Paul Churchland

For over three decades, Paul Churchland has been a provocative and controversial philosopher of mind and philosopher of science. He is most famous as an advocate of "eliminative materialism," whereby he suggests that our commonsense understanding of our own minds is radically defective and that the science of brain demonstrates this (just as an understanding of physics reveals that our commonsense understanding of a flat and motionless earth is similarly false). This collection offers an introduction to Churchland's work, as well as a critique of some of his most famous philosophical positions. Including contributions by both established and promising young philosophers, it is intended to complement the growing literature on Churchland, focusing on his contributions in isolation from those of his wife and philosophical partner, Patricia Churchland, as well as on his contributions to philosophy as distinguished from those to Cognitive Science.

Brian L. Keeley is an Associate Professor of Philosophy at Pitzer College in Claremont, California. His research has been supported by the National Science Foundation, the National Institute for Mental Health, the McDonnell Project for Philosophy and the Neurosciences, and the American Council of Learned Societies. He has published in the *Journal of Philosophy*, *Philosophical Psychology*, *Philosophy of Science*, *Biology and Philosophy*, and *Brain and Mind*.

Contemporary Philosophy in Focus

Contemporary Philosophy in Focus offers a series of introductory volumes to many of the dominant philosophical thinkers of the current age. Each volume consists of newly commissioned essays that cover major contributions of a preeminent philosopher in a systematic and accessible manner. Comparable in scope and rationale to the highly successful series **Cambridge Companions to Philosophy**, the volumes do not presuppose that readers are already intimately familiar with the details of each philosopher's work. They thus combine exposition and critical analysis in a manner that will appeal to students of philosophy and to professionals as well as to students across the humanities and social sciences.

FORTHCOMING VOLUMES:

Ronald Dworkin edited by Arthur Ripstein
Jerry Fodor edited by Tim Crane
Saul Kripke edited by Alan Berger
David Lewis edited by Theodore Sider and Dean Zimmermann
Bernard Williams edited by Alan Thomas

PUBLISHED VOLUMES:

Stanley Cavell edited by Richard Eldridge
Donald Davidson edited by Kirk Ludwig
Daniel Dennett edited by Andrew Brook and Don Ross
Thomas Kuhn edited by Thomas Nickles
Alasdair MacIntyre edited by Mark Murphy
Hilary Putnam edited by Yemina Ben-Menahem
Richard Rorty edited by Charles Guignon and David Hiley
John Searle edited by Barry Smith
Charles Taylor edited by Ruth Abbey

Paul Churchland

Edited by

BRIAN L. KEELEY
Pitzer College

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 2RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521830119

First published in print format

ISBN-13 978-0-511-18301-0 eBook (MyiLibrary)
ISBN-10 0-511-18301-1 eBook (MyiLibrary)

ISBN-13 978-0-521-83011-9 hardback
ISBN-10 0-521-83011-7 hardback

ISBN-13 978-0-521-53715-5 paperback
ISBN-10 0-521-53715-0 paperback

Contents

Preface *page* ix
BRIAN L. KEELEY

Acknowledgments xiii

List of Contributors xv

1 Introduction: Becoming Paul M. Churchland (1942–) 1
BRIAN L. KEELEY

2 Arguing For Eliminativism 32
JOSÉ LUIS BERMÚDEZ

3 The Introspectibility of Brain States as Such 66
PETE MANDIK

4 Empiricism and State Space Semantics 88
JESSE J. PRINZ

5 Churchland on Connectionism 113
AARRE LAAKSO AND GARRISON W. COTTRELL

6 Reduction as Cognitive Strategy 154
C. A. HOOKER

7 The Unexpected Realist 175
WILLIAM H. KRIEGER AND BRIAN L. KEELEY

8 Two Steps Closer on Consciousness 193
DANIEL C. DENNETT

Index 211

Preface

Philosophy is, among other conceptions no doubt, a human quest for comprehension, particularly *self*-comprehension. Who am I? How should I understand the world and myself? It is in this context that the philosophical importance of Paul M. Churchland (PMC) is most evident. For three decades and counting, PMC has encouraged us to conceive of ourselves from the "Neurocomputational Perspective" – not only as a minded creature, but also as minded due to our remarkable nervous system. Our brains, ourselves. This represents a unique and interesting way to approach this hoary philosophical enquiry.

However, his lasting intellectual contribution as we enter a new millennium is not so much some particular way of seeing ourselves, but rather his unwavering belief that we are capable of perceiving the world and ourselves in ways very different from the norm. PMC has made a career as a sort of Patron Saint of Radical Re-conceptualization. Again and again he argues that we do not *have* to see ourselves in ordinary and well-worn terms. Copernicus had us throw out our commonsense framework of a flat, motionless Earth, wandering planets, and a sphere of fixed stars and showed us how to see the night sky with new eyes. PMC urges us to consider the possibility that many more such conceptual revolutions await us, if only we would give them a fair hearing.

The invocation of Copernicus is fitting. PMC is a philosopher of mind whose intuitions and ideas are primarily informed by science and the philosophy of science. As he put it in the preface to his 1989 *A neurocomputational perspective: The nature of mind and the structure of science*, "The single most important development in the philosophy of mind during the past forty years has been the emerging influence of philosophy of science. . . . Since then it has hardly been possible to do any systematic work in the philosophy of mind, or even to understand the debates, without drawing heavily on themes, commitments, or antecedent expertise drawn from the philosophy of science" (xi). Whereas for many, philosophy of psychology (or philosophy of cognitive science) is primarily a branch of philosophy of mind, PMC

sees it as a branch of philosophy of science; that is, as the exploration into the unique philosophical problems raised in the context of the scientific study of the mind/brain.

In the pages of this collection of papers, a number of Paul Churchland's contemporaries explore and assess his contributions to a variety of discussions within philosophy. The various authors will discuss his views both with an eye toward explicating his sometimes counterintuitive (and therefore often provocative) positions and another toward critiquing his ideas. The result should be a deeper appreciation of his work and his contribution to the present academic milieu.

In addition to a number of articles over the years, there have been a small number of book length works and collections on the philosophy of Paul Churchland (jointly with that of his wife, Patricia). Notable among these has been McCauley's 1996 collection, *The Churchlands and their critics* (McCauley 1996), which brings together a number of philosophers and scientists to comment critically on various aspects of their philosophy along with an informative response by the Churchlands. A very accessible, short-but-book-length exploration is Bill Hirstein's recent *On the Churchlands* (Hirstein 2004). While both of these are recommended to the reader interested in learning more about Churchland's philosophy, the present volume attempts to be different from, while at the same time being complementary to, this existing literature.

As with Hirstein's volume, the present collection attempts to be accessible to the nonexpert on the neurocomputational perspective. But unlike it, we do so from the multiple perspectives of the contributors and cover a wider array of topics. Where Hirstein's volume has the virtue of a single author's unified narrative, the present volume has the virtue of a variety of perspectives on the philosopher at hand.

The McCauley volume is also a collection of papers by various authors, but the goal there is explicitly critical; whereas in the present volume, the critical element is strongly leavened with exegetical ingredients. All the authors here spend a good amount of space spelling out Churchland's position before taking issue with it. Also, the explicit target here is to understand the work of Paul Churchland *as a philosopher*. Because Churchland works in the highly interdisciplinary field of Cognitive Science and spends much of his time engaging neuroscientists of various stripes, it is often useful to consider his contributions to the world as a cognitive scientist. While a laudable endeavor, that is not the approach taken here. Here we are attempting to come to grips with Churchland's contribution to the philosophical realm, although this should not be taken as devaluing his contributions elsewhere.

Finally, other secondary literature dealing with the work of Paul Churchland – including the two volumes discussed previously – often consider his work as of a piece with that of his wife, Patricia Churchland. That is not the approach here. Instead, we have set our sights on the work of Paul, although his wife's work is discussed as is necessary to understand Paul's philosophical insights. While their work is clearly interdependent at a very deep level – often Paul's work is the *yin* to Pat's *yang* – each is a clear and cogent thinker in his and her own right. To avoid having it seem that Pat acts as the mere handmaiden to Paul's work (or vice versa), we primarily deal with Paul's work here.[1]

Brian L. Keeley, Pitzer College

Note

1. Although see Note 1 of Chapter one for more on the difficulties of separating the discussion of either philosopher from that of the other.

Works Cited

Churchland, P. M. (1989). *A neurocomputational perspective: The nature of mind and the structure of science*. Cambridge, MA, The MIT Press (A Bradford Book).

Hirstein, W. (2004). *On the Churchlands*. Toronto, Thomson Wadsworth.

McCauley, R. N., Ed. (1996). *The Churchlands and their critics. Philosophers and their critics*. Cambridge, MA, Blackwell Publishers Ltd.

Acknowledgments

Approximately a decade ago, I was sitting in Pat and Paul Churchland's hot tub – yes, the rumors are true: West Coast philosophy does occur under such conditions. I asked Paul to reveal to me the key to a successful career in philosophy. "Get other people to write about *you*," is my memory of his response. Although this advice might seem as useful as "Buy low; sell high," to a graduate student spending his days writing about this or that philosophical figure, it did convey an important message about how one needs to think about one's future scholarship. That out-of-the-classroom lesson explains in part why I took on the project of editing this book. It offers me the chance to pay back in a very appropriate way the debt for this and many other lessons Paul has taught me over the years.

Much of what I learned about Paul's work came not from him, but through my contact with Pat Churchland. She was one of the two chairs of my Ph.D. dissertation committee; and, as a member of her Experimental Philosophy Lab, and in countless classrooms, office hours, talk receptions, and so on, I have learned from Pat not only how to be a scholar and philosopher, but quite a lot about how her and Paul's views have developed over a long, fruitful career. I would not have had the confidence to undertake a volume like this if it were not for her influence.

I owe a big debt of gratitude to the contributors to this volume who hung in there, despite the seemingly slow process.

Bill Bechtel was, as always, an early and indefatigable supporter of my own work in general, and this volume in particular. Carrie Figdor read over portions of my contributions and offered valuable feedback.

I should acknowledge the financial support the *McDonnell Project in Philosophy and the Neurosciences*, as directed by Kathleen Akins, while I was working on this collection. The group of scholars she gathered together for that project resulted in the initial contributors to this volume.

Some of the early work of my Chapter was carried out while I was in residence as a Fellow of the *Center for Philosophy of Science* at the University of Pittsburgh. The members of the Center, along with the faculty, staff,

and students of the History and Philosophy of Science Department there, deserve my thanks for both a pleasant as well as edifying four months in Fall 2003. I should thank Sandy Mitchell (not incidentally, the *other* of the two chairs of my Ph.D. dissertation) in particular.

Finally, my thanks goes to my friends and colleagues at Pitzer College for their continuing support of faculty scholarship, specifically in the form of several awards from the Research & Awards Committee and the granting of my sabbatical leave in Fall 2003.

Contributors

JOSÉ LUIS BERMÚDEZ is Professor of Philosophy and Director of the Philosophy-Neuroscience-Psychology program at Washington University in St. Louis. He is the author of *The Paradox of Self-Consciousness*, *Thinking without Words*, and *Philosophy of Psychology: A Contemporary Introduction*.

GARRISON W. COTTRELL is Professor of Computer Science and Engineering at the University of California, San Diego. His main research interest is in building working models of cognitive processes using neural networks. His most recent work has been on understanding face and object processing. His work has been published in *Journal of Neuroscience*, *Nature*, *Philosophical Psychology*, *Psychological Science*, and the *Journal of Cognitive Neuroscience*, among others.

DANIEL C. DENNETT is University Professor and Director of the Center for Cognitive Studies at Tufts University. His most recent awards are the Barwise Prize, presented by the American Philosophical Association's Committee on Philosophy and Computers, the Bertrand Russell Society Award for 2004, and Humanist of the Year, 2004, from the American Humanist Association. He is the author of many books, including most recently, *Freedom Evolves* and *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness*.

C. A. HOOKER holds a Chair of Philosophy at the University of Newcastle, Australia. He is a Fellow of the Australian Academy of Humanities. He has published eighteen books and more than one hundred research papers, including *Reason, Regulation and Realism: Toward a Naturalistic*, *Regulatory Systems Theory of Reason*, and *A Realistic Theory of Science*.

BRIAN L. KEELEY is Associate Professor of Philosophy at Pitzer College in Claremont, CA. He is a member of the McDonnell Project in Philosophy and the Neurosciences and has recently been awarded a Charles A. Ryskamp Fellowship from the American Council of Learned Societies. He

has published a number of papers, including two in *Journal of Philosophy*: "Making Sense of the Senses: Individuating Modalities in Humans and Other Animals" and "Of Conspiracy Theories."

WILLIAM H. KRIEGER is a Lecturer in Philosophy at the California State Polytechnic University, Pomona. He is the author of a forthcoming book on philosophical and archaeological explanation: *Can There Be a Philosophy of Archaeology? Processual Archaeology and the Philosophy of Science*. He is also a field director at Tell el-Farah, South Archaeological Excavations.

AARRE LAAKSO is a Postdoctoral Fellow in Psychology at Indiana University, Bloomington. His research concerns links between psychology and philosophy, such as cognitive architectures and the nature of psychological explanation, spatial representation and reference, and language acquisition and nativism. His work has appeared in *Philosophical Psychology*, *Psycoloquy*, *Metapsychology*, and *Behavioral and Brain Sciences*.

PETE MANDIK is Associate Professor of Philosophy and Coordinator of the Cognitive Science Laboratory at William Paterson University. He is a member of the McDonnell Project in Philosophy and the Neurosciences. His work has appeared in *Cognition and the Brain: The Philosophy and Neuroscience Movement* and he is an editor of *Philosophy and the Neurosciences: A Reader*.

JESSE J. PRINZ is Associate Professor of Philosophy at the University of North Carolina at Chapel Hill. He has research interests in the philosophy of cognitive science, philosophy of language, and moral psychology. His books include *Furnishing the Mind: Concepts and Their Perceptual Basis*, and *Gut Reactions: A Perceptual Theory of Emotion*.

1 Introduction: Becoming Paul M. Churchland (1942–)

BRIAN L. KEELEY

The goal of this chapter is two-fold. First, I will present an overview of the philosophical vision of Paul M. Churchland (PMC). This will help situate the more detailed, and necessarily narrower, discussions of the other authors in this volume. Second, the more substantive goal here is to show that Paul Churchland's views have not developed in a vacuum. While he has clearly developed his own unique view of the philosophical terrain, he is not without his influences – influences that he in no way attempts to hide. His work is a unique blend of ideas encountered as a nascent philosopher. The philosophers I will be discussing are not always so well known to today's students of philosophy, so there is value in considering how these views of the preceding generation are being passed on within the work of one of today's more influential philosophers of mind and science.

I will begin by sketching Paul Churchland's personal biography. After getting the basic facts on the table, I will turn to the three philosophers whose influence on PMC are my foci: Russell Hanson, Wilfrid Sellars, and Paul Feyerabend. Each of these thinkers made philosophical contributions that are reflected in the work of PMC. Next, I will show how all three of these thinkers contributed to the philosophical position most closely associated with Churchland, namely "Eliminative Materialism." My comments critical of Churchland's version of eliminative materialism are meant to set the stage for the rest of this volume's contributions, as this philosophical framework is at the core of PMC's view of science, the mind, and the science of the mind.

PERSONAL HISTORICAL OVERVIEW

PMC was born a Canadian and earned a B.A. from the University of British Columbia, and in 1969, he was awarded a Ph.D. in Philosophy from the University of Pittsburgh. There, he wrote a dissertation under the direction of Wilfrid Sellars. He spent the first 15 years of his career at the University

of Manitoba, taking advantage of its relative isolation to further develop his own approach to the ideas to which he was exposed during his graduate education. In addition to a number of important early papers on eliminative materialism and the status of commonsense reasoning, he published his first two books. The first is his still-insightful monograph, *Scientific Realism and the Plasticity of Mind* (1979). Here, he lays out his views on the nature of scientific process and how it is based in the cognitive capacities of adult, human scientists.

His second book, *Matter and Consciousness* (1984, revised and updated 1988; translated into five languages), has become one of the most popular textbooks in the philosophy of mind. (Rumor has it that this book is the all-time bestseller for the Bradford Books imprint of the MIT Press; quite an impressive achievement given the competition from the likes of Jerry Fodor, Dan Dennett, Stephen Stich, and Fred Dretske, to name only a few.) *Matter and Consciousness* provides an introduction to the Churchland worldview; how the problems of the philosophy of mind are to be approached from a perspective developed out of the neural sciences. The book is an important step in PMC's development because it contains the first sustained discussions of contemporary neuroscience and how these theories and discoveries provide grist for the traditional philosophical mill.

Several of PMC's early papers were co-authored with his perennial partner in crime: his wife, Patricia Smith Churchland. Starting early in their respective careers, these two have worked closely together; a more-than-three-decades-long collaboration so close that it is often difficult to determine who is ultimately responsible for this or that idea.[1]

In 1984, the Churchlands moved to the institution with which they would become most closely associated: the University of California, San Diego (UCSD).[2] There, he fell in with the then-burgeoning Connectionist (a.k.a. Parallel Distributed Processing (PDP)) movement in cognitive science. According to the proposals of this group, the mind is best understood as a computational system formed of networks of simple processing units. The units are modeled on neurons (in that they sum inputs analogously to the behavior of dendrites and either "fire" or not in a process akin to a paradigmatic neuron's either producing an action potential down its axon or not). While other models of the mind made use of language-like units (say, formal symbols in a "language of thought" (Fodor 1975)), the PDP approach was intended to present a "sub-symbolic" alternative to such theories of mind in that the fundamental units are vectors of activation across networks of neuron-like entities (cf., Smolensky 1988; Clark 1989). The two-volume bible of this approach came out of the San Diego–based

PDP Research Group two years later (McClelland and Rumelhart 1986; Rumelhart and McClelland 1986).

From this point forward, the science of connectionism and what came to be known more generally as "computational neuroscience" became the main source of scientific theories and ideas used by Churchland to present his new theory of mind. His next two major works explore how to apply the insights resulting from thinking of the mind as a neural net to a variety of problems within philosophy: *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science* (1989) and *The Engine of Reason, The Seat of the Soul: A Philosophical Journey into the Brain* (1995, translated into six languages). A collection of papers by Paul and Pat, separately and together, has also been published (Churchland and Churchland 1998).

As of the writing of this chapter, Paul is still as productive as ever and continues his career as Professor of Philosophy at UCSD.

INFLUENCES

The question of influences on a thinker is necessarily irresolvable in any final way. The influence of some – Socrates, Plato, Hume, Kant – are so wide ranging that there is little value in trying to pick out their specific contributions to any given philosopher. Anyone with a reasonably strong background in philosophy can see their influences on most who followed them. Two clear influences on PMC whose ubiquity, even in a very short span of time, is wide ranging are W. V. O. Quine and Thomas Kuhn. Quine's promotion of naturalized epistemology opened the way for the highly naturalized approach that PMC has undertaken.[3] Kuhn's post-positivist exploration of the dynamics of theory change within science places a strong emphasis on the psychological processes of individual human scientists. This foreshadows PMC's own concerns with the scientist as learning machine and the human learner as a kind of scientist. That said, it seems as though it is practically impossible for philosophers to avoid reading Quine and Kuhn these days, so spotting these influences is less than earth shattering.

In what follows, I will concentrate on three philosophers – Russell Hanson, Wilfrid Sellars, and Paul Feyerabend – all of whose work is clearly reflected in the mature philosophy of Paul Churchland. Furthermore, their work is sometimes overlooked by recent generations of philosophers,[4] such that, while reading Churchland, it may be unclear what is his unique contribution and what he takes from those upon whose shoulders he stands. While he is clearly influenced by these thinkers, it is not fair to say that he is

merely parroting them. With each influence, he accepts some aspects of the proffered theory and weaves those ideas into a tapestry of his own making. He clearly rejects some elements as misguided or otherwise wrongheaded. It is instructive to undertake an investigation into such a personal history of ideas because it reveals decisions on the part of Churchland as to what component ideas to embrace and which to leave by the wayside.

HANSON

Norwood Russell Hanson (1924–67) is not so well known today, in part because he did his most important philosophical work in the years after the disillusionment with Logical Positivism but before the rise of some of the more popular post-positivist approaches to philosophy of science, such as found in the work of Lakatos and Kuhn. Therefore, his oeuvre gets short shrift. This is a shame because Hanson's work is an important stepping-stone from the positivist dreams of Carnap, Ayer, and others to the contemporary work of philosophers such as PMC.

One belief that Hanson and PMC share is that philosophy of science is best done with a solid understanding of the practice of science. Large chunks of Hanson's work in philosophy of science involve detailed discussion of the minutia of science and its practice. In the introduction to his landmark *Patterns of Discovery*,[5] Hanson writes,

> The approach and method of this essay is unusual. I have chosen not to isolate general philosophical issues – the nature of observation, the status of facts, the logic of causality, and the character of physical theory – and use the conclusions of such inquiries as lenses through which to view particle theory [in physics]. Rather the reverse: the inadequacy of philosophical discussions of these subjects has inclined me to give a different priority. Particle theory will be the lens through which these perennial philosophical problems will be viewed. (1958: 2)

As a result of this novel approach, a significant portion of Hanson's book contains a fairly detailed discussion of then-current particle microphysics.[6] Decades later, it would be PMC's books that would be filled with the details of science. The reason for this is not mere "scientism" on the part of Hanson and Churchland (despite what some critics might believe (Sorell 1991)). Instead, their reason is that it is in the practice of science – particularly of new and unsettled disciplines – that one finds the most interesting philosophical

problems and often the material for their solution. What Hanson wrote of particle physics in 1958 would be equally true of the neural and cognitive sciences of the 1980s: "In a growing research discipline, inquiry is directed not to rearranging old facts and explanations into more elegant formal patterns, but rather to the discovery of new patterns of explanation. Hence the philosophical flavour of such ideas differs from that presented by science masters, lecturers, and many philosophers of science" (1958: 2). Like Kuhn, Hanson stressed the importance of studying how science is actually conducted (and not how it is mythologized after the fact). It is in the practice of actual science that one finds explanatory genesis. For Hanson, the chosen source was particle physics; for Churchland, it is computational neuroscience.

So, what image of science did Hanson get from this detailed look at physics and how did it differ from that of his allegedly misinformed predecessors? First, Hanson argued that one of the central tenets of Logical Positivism – the distinction between the context of discovery and the context of justification – was a nonstarter. According to the dogma Hanson sought to challenge, there are two different aspects to the formation of new theories. The first aspect, the *context of discovery*, is the often-mysterious process of the creation of new hypotheses. How does a scientist generate a new hypothesis? The second, the *context of justification*, is the more structured and logical process of determining whether a given hypothesis is correct. Given a hypothesis, how does a scientist figure out whether it is correct?

The classic illustrative example of this distinction is Friedrich Kekulé's famous description (years after the event) of how he came to discover the chemical structure of benzene (Kekulé 1890/1996). As he describes it, the idea that the benzene molecule had a ring structure came to him as he was dozing next to a fire during an evening break from trying to work out a solution to this structural problem. Having arrived at this proposal, "... I spent the rest of the night working out the consequences of the hypothesis" (34). Thus, while the creative process through which the hypothesis was generated seems relatively mysterious (it just came to him while he napped), that process is distinct from the more rigorous (and fully conscious) process of *working out* the logical consequences of the idea in order that it may be tested.

The work done by this distinction in the positivist story is the demarcation of a division of labor within the study of the scientific method. The context of discovery, with its apparently irrational intuitive leaps and the

like, is the purview of psychologists. The logic of the context of justification is not so unconstrained and willy-nilly, and this is where philosophy of science must necessarily dig in and set the rules. The creative aspect of discovery is, in essence, rule-breaking whereas the justification process is essentially rule-driven. Philosophy of science, according to the positivists, has the goal of determining what those rules should be.

While such a division of labor offers a neat and clean picture of the scientific process and a clear role for philosophical inquiry, Hanson argued that it is simply not an accurate portrayal of the scientific process. The only way one might come to believe it *is* the correct picture would be by concentrating too much on such cleaned up "text book" examples as Kekulé's. Instead, when one looks at how science is actually done, it is revealed that the discovery of explanatory patterns is not only tractable and interesting, it is perhaps *the* most interesting part of the scientific method: "The issue is not theory-using, but theory-finding; my concern is not with the testing of hypotheses, but with their discovery. Let us examine not how observation, facts and data are built up into general systems of physical explanation, but how these systems are built into our observations, and our appreciation of facts and data" (1958: 3).

The idea that theories are "built into our observations" brings us to Hanson's most lasting contribution to philosophy of science: the thesis that scientific observation is inescapably "theory-laden" (to use the term he introduces into the philosophical lexicon in Hanson (1958: 19–24); see also Hanson (1971: 4–8). Positivist dogma held that an essential component of the logic of justification is the claim that the process of observation is independent of our theorizing about the world. After working out the empirical consequences of a particular hypothesis, we evaluate it by observing the world and determining whether its predictions obtain. On the positivist view, in order to be an arbiter of theory evaluation, observation must, in principle, be independent of theory. Again, the merely psychological (the physiology of perception) is distinct from the philosophical (the interpretation of observations as evidence either for or against a particular theory).

Hanson again rejects this simplifying distinction, arguing that observation cannot be so cleanly separated from theory: "The color-blind chemist needs help from someone with normal vision to complete his titration work – whether this someone be another chemist, or his six-year-old son, does not matter. But, now, are there any observations that the latter, the child, could *not* make?" (1971: 4). Hanson's answer is "yes."

After citing a passage from Duhem (1914: 218) that foreshadows the claim he wants to propose, Hanson asks what is presupposed by an act of genuine scientific observation. The ability to sense is one thing.

> *Knowledge* is also presupposed; scientific observation is thus a 'theory-laden' activity. . . . Brainless, photosensitive computers – infants and squirrels too – do not make scientific observations, however remarkable their signal-reception and storage may be. This can be no surprise to any reader of this book. That the motion of Mars is retrograde, that a fluid's flow is laminar, that a plane's wing-skin friction increases rapidly with descent, that there is a calcium deficiency in Connecticut soil, that the North American water table has dropped – these all concern observations which by far exceed the order of sophistication possible through raw sense experience. Nor are these cases of simply requiring physicobiological 'extensions' to the senses we already have; for telescopes, microscopes, heat sensors, etc., are not sufficient to determine that Mars' motion is retrograde, that blood poisoning is settling in, that volcanic activity is immanent. Being able to make sense of the sensors requires knowledge and theory – not simply more sense signals. (Understanding the significance of the signal flags fluttering from the bridge of the *Queen Elizabeth* does not usually require still *more* flags to be flown!) (1971, 5).

This inseparable intermixing of theory and observation is central to Hanson's thought. Along with the importance of engaging actual scientific practice, the theory-ladenness of observation becomes a foundation stone in PMC's philosophy as well. We will turn to where PMC parts company with Hanson later, following a discussion of his affinities with the two other philosophers considered here.

SELLARS

Wilfrid Sellars (1912–1989), son of philosopher Roy Woodward Sellars (1880–1973), taught at the University of Minnesota and Yale, before finally settling at the University of Pittsburgh, where he supervised a doctoral thesis by Paul Churchland.[7]

According to Sellars (1960/1963),

> The aim of philosophy, abstractly formulated, is to understand how things in the broadest possible sense of the term hang together in the broadest possible sense of the term. Under "things in the broadest possible sense" I

> include such radically different items as not only "cabbages and kings", but numbers and duties, possibilities and finger snaps, aesthetic experience and death. To achieve success in philosophy would be, to use a contemporary turn of phrase, to "know one's way around" with respect to all these things, not in that unreflective way in which the centipede of the story knew its way around before it faced the question, "how do I walk?", but in that reflective way which means that no intellectual holds are barred. (1)

This is to say that Sellars sees the academic discipline of Philosophy as not so much asking the "Big Questions" as asking the "*Broad* Questions." It is that which stitches together all of our various understandings of the world – those provided by the natural and social sciences, those of the humanities, as well as those of ordinary humans just grappling with their multifarious worlds – into a coherent, unified conception of the world. By "unified," we should not think of anything akin to a classical reductionist picture in which every legitimate form of explanation should eventually be translated into some single language (cf., deVries and Triplett 2000: 114–16). Instead, there will likely be many different understandings, with philosophy providing the intellectual resources for understanding how they, as he says, "hang together."[8]

During his long career, Sellars made a number of contributions to philosophy, quite a few of which had an impact on the work of his apprentice. The first I will note is a key distinction Sellars draws in the ways that we humans understand ourselves, referred to earlier. Sellars distinguishes two "images" or very general philosophical frameworks for understanding human activity. The first is the *manifest image* – the embodiment of our commonsense understanding of human behavior, including our own personal behavior. Sellars (1960/1963) characterizes "... the manifest image of man-in-the-world as the framework in terms of which man encountered himself – which is, of course, when he came to be man" (6). This image is not pre-theoretical in the sense of being unreflective. Rather this is the image of oneself achieved upon taking oneself as an object of understanding; what humans got when they first realized that they, too, were something that required understanding, in addition to all the other confusing aspects of the world, including other animals, the weather, the night sky, etc.[9] Furthermore, it is a framework in which the basic ontological category is that of "persons." In the manifest image, everything understood is understood in terms of being a kind of person. As deVries and Triplett (2000) put it, "It is our refined commonsense conception of what the world and ourselves are and how they interact" (190).

The manifest image is contrasted with what Sellars calls the *scientific image*. This is the image of our self and the world provided by the explicit theorizing of post-Enlightenment science. There is a strong "what-you-see-is-what-you-get" element in Sellars' conception of the manifest image. He cites Mill's inductive method as central to the method of the manifest image; such explanation is generated by noting the correlations of observed events in the world (1960/1963: 7). In contrast, what demarcates the method of the scientific image is its method of hypothesis and the postulation of the unobserved and the unobservable in the service of explanation. The fundamental ontology of the manifest image (persons) is directly observable to everyone; indeed if all one had was the manifest image, persons are all one would ever see. By contrast, the fundamental ontology of the scientific image, say that provided by contemporary physics, is one of unobservable atomic elements, atomic forces, and the like.[10]

What is the relationship between these two images? They are often taken to be opposed to one another. As one striking example, one line of thought derives from taking the scientific perspective on humans themselves and seeing them not as persons in the sense of the manifest image but rather as a collection of abstract, scientific entities (cell assemblies, molecules, expressed DNA, quarks, what have you): "Even persons, it is said (mistakenly, I believe), are being 'depersonalized' by the advance of the scientific point of view" (Sellars 1960/1963: 10). This is "mistaken" because he takes the goal of philosophy to be explanation in the broadest sense; he sees both images as essential to a full understanding of humans, the world, and the place of humans in the world. He likens the relationship between the two to be that of the different component images of a stereoscopic diagram. Properly viewed through a pair of stereoscopic lenses, the two images combine to provide an image with dimensions lacking in either component image on its own.[11]

Sellars' notion of these two different images of ourselves and the world around us show up in PMC's career-long concern with what have come to be known as "folk theories." Folk theories are what they sound like: the commonsense theories possessed by the average person. In particular, PMC is concerned with *folk psychology*, our commonsense theory of animal (most important, human) thought and behavior.[12] While PMC accepts Sellars' distinction between the two images, how he treats the relationship between these two images represents perhaps his largest break from his dissertation advisor, but that will addressed in the following section.

Another contribution Sellars made to contemporary philosophy – the contribution he is likely best known for today – is his attack on

foundationalist epistemology, such as one finds, for example, in the work of C. I. Lewis (1929, 1945). Like Hanson, Sellars disagreed with the positivist tenet that there was some store of human-independent data upon which we can build our scientific knowledge by using these data to arbitrate between hypotheses. However, where Hanson attacks the notion that such data can exist independently of our theories, Sellars takes a slightly different tack. Sellars takes issue with the very notion of this fund of data, what he calls the "Myth of the Given." His *Empiricism and the Philosophy of Mind* is a long argument intended to expose this myth and undermine its foundation (Sellars 1956/1997). As Richard Rorty puts it in his introduction to the recent republication of this essay, this work, "... helped destroy the empiricist form of foundationalism by attacking the distinction between what is 'given to the mind' and what is 'added by the mind.' Sellars' attack on the Myth of the Given was a decisive move in turning analytic philosophy away from the foundationalist motives of the logical empiricists. It raised doubts about the very idea of 'epistemology,' about the reality of the problems which philosophers had discussed under that heading" (Rorty 1997: 5).

Along with Hanson's related arguments for theory-laden observation, PMC takes Sellars' Myth of the Given arguments on board in his own work.

FEYERABEND

Paul K. Feyerabend (1924–94) was a sometimes self-deprecating[13] epistemologist and philosopher of science. The slogan, "Anything goes," summed up his approach to philosophy (and probably explains some of his popular cachet in the radical 1960s and early 1970s). He was passionate in his defense of explanatory pluralism and tried to keep alive the iconoclastic spirit of early Enlightenment science against the growing hegemony of industrialized and institutionalized science. He saw that the tables had turned; whereas once science had to eke out a precarious existence in the shadow of culture-dominating seventeenth century ecclesiastical powers, in the late twentieth century he saw the need to write papers with titles such as "How to defend society against science" (Feyerabend 1975). Following World War II, science was quickly becoming one of the dominant cultural institutions of the world. Having served in Hitler's army as a young man, Feyerabend was deeply suspicious of any tyrannical force in society, no matter how benevolent its stated intentions.

Feyerabend sees science – properly understood – as a fundamentally democratic process, rather than as a necessarily truth-seeking one. In fact, he

noted these two sentiments are in potential conflict as the search for a single "truth" can often be used to shut down the alternative points of view that genuinely democratic processes need. When given a choice between truth and freedom, Feyerabend invariably picked freedom, even if that meant embracing the freedom to be wrong. Relying too heavily on the goal of seeking a single truth is dangerous to scientific progress, as John Preston has recently summarized Feyerabend's view:

> . . . as long as we use only one empirically adequate theory, we will be unable to imagine alternative accounts of reality. If we also accept the positivist view that our theories are summaries of experience, those theories will be void of empirical content and untestable, and hence there will be a diminution in the critical, argumentative function of our language. Just as purely transcendent metaphysical theories are unfalsifiable, so too what began as an all-embracing scientific theory offering certainty will, under these circumstances, have become an irrefutable dogma, a *myth*. (Preston 2000: 143, emphasis in original)

It is important not to see Feyerabend as an opponent or enemy of science. Rather, he is a critic who wishes to save science from itself. Science is too important not to do correctly.

Feyerabend is a scientific realist and a metaphysical materialist. Feyerabend attempts to defend materialism by defining its real opponent as the commonsense idiom that all of us share and bring with us to philosophical and scientific debates concerning the nature of the mind. After generally railing against typical philosophical objections arising from the tendency to criticize a new approach as impossible before it is even given a chance to develop, Feyerabend turns to the two most common claims against a materialist theory of mind: (1) that such a theory would be meaningless, and even if this is not the case, (2) that materialism is simply false.

Against the charge of meaninglessness, Feyerabend points out that what this claim really means is that a materialist theory of mind is in serious conflict with our commonsense idiom (in today's terms, our "folk psychology" (FP), as discussed in connection with Sellars). But what exactly is it that makes the commonsense idiom the bedrock of meaning? It cannot be its wide and common usage alone. Feyerabend (1963a) asks rhetorically: "Is it really believed that a vigorous propaganda campaign which makes everybody speak the materialist language will turn materialism into a correct doctrine?" (50). Elsewhere, he writes,

> The objection [that a new descriptive language must be related to a previous one] assumes that the terms of a general point of view and a correspondence

> language can obtain meaning only by being related to the terms of some other point of view that is familiar and known by all. Now if that is indeed the case, then how did the latter point of view and the latter language ever obtain its familiarity? And if it could obtain its familiarity without help "from outside," as it obviously did, then there is no reason to assume that a different point of view cannot do equally well. (1963b: 173)

So, the common idiom cannot be the foundation of meaning by virtue of its commonality. It also cannot be true by virtue of its practical success. J. L. Austin argued for this feature of commonsense theory:

> Our common stock of words embodies all the distinctions men have found worth drawing, and the connexions they have found worth marking, in the lifetime of many generations: these surely are likely to be more numerous, more sound, since they have stood up to the long test of survival of the fittest, and more subtle . . . than any that you or I are likely to think up . . . (as quoted in Feyerabend (1963a: 50–1))

According to Feyerabend (1963a), this argument fails on three counts: First, "such idioms are adapted not to facts, but to beliefs" (51). That is, such an idiom succeeds or fails by virtue of its ability to be defended by cultural institutions, how it jibes with the hopes and fears of the community, and so on. The truth of these beliefs need not be questioned in this process, and it is just these beliefs that materialism calls into question. Second, the commonsense idiom is never tested as scientific hypotheses and theories are. It is never tested in the systematic and controlled way that modern science has developed to test the mettle of hypotheses. Third, even if the idiom was "tested" in the correct sense (that is, assuming that the second point is overstated), then one *still* cannot throw out the materialist story for being less "successful" than the common idiom. Although he explains this in more detail elsewhere (1962), the general notion is that one cannot compare the materialist theory with "the facts" because the facts are formulated in terms of the common idiom, hence prejudicing the evidence in its favor. Certain "facts" are only empirically accessible from within certain language games, to use Wittgensteinian terminology. Feyerabend (1963a) concludes this stage of the debate by claiming that ". . . if you want to find out whether there are pains, thoughts, and feelings in the sense indicated by the common usage of these words, then you must become (among other things) a materialist" (53).

In other words, what we find in Feyerabend is a coming together of Hanson's embrace of scientific practice as a source of philosophical

understanding with Sellars' identification of a distinct commonsense view of the world (the manifest image) and his contention that we cannot accept that the way things appear to us are necessarily as they are (the myth of the given). Feyerabend observes that this commonsense view (which is typically dualist or at least in some way non-materialist) may well be in conflict with the materialist view of the mind being presented to us by current science. The conclusion that Feyerabend draws from this confluence of ideas is that we must be open to the possibility that the commonsense view of the world may be *radically false*. If the commonsense view is not epistemically given and it represents a theory-like view of the world, then it is only logical to conclude that it might be false, no matter how wrenching that possibility might be to our tightly held intuitions.

This view, which Feyerabend dubs "eliminative materialism," is laid out by him in a number of often quite short papers (1962, 1963a, 1963b). PMC has spent much of his career carrying the Feyerabend mantle forward. Below, we need to take a look at PMC's eliminative materialism in more detail. However, before turning to that, we will look at the aspects of Hanson, Sellars, and Feyerabend that PMC fails to take on board, because they are instructive concerning PMC's unique take on the ideas Hanson, Sellars and Feyerabend held.

WHAT CHURCHLAND REJECTS

Churchland makes use of many of the post-positivist insights of our three central thinkers. He, like Sellars, is aware of the power of our commonsense intuitions and, with Feyerabend, he is downright suspicious of them. With Hanson, he sees the important role that embryonic science can have in providing interesting material for a philosopher to work with. He agrees with all three of these philosophers that to rely on a clear separation of what we think, theoretically, is the case from what we see requires the confidence of a fool.

However, he doesn't accept everything that comes from these thinkers who paved the way for a post-positivist philosophical world. Instead of noting small, picky points here and there, I will instead concentrate on what I see as perhaps the biggest difference between PMC and all three of these influences. Just as PMC is a product of the zeitgeist of his generation of philosophers, so too were Hanson, Sellars, and Feyerabend. One apparently inescapable influence in their time was the work of Ludwig Wittgenstein. Wittgenstein's influence is explicit and notable in the work of all three

of our precursors to PMC. However, PMC's own work barely mentions Wittgenstein, resulting in the impression that what we have in Churchland's mature philosophy is the philosophy of Hanson, Sellars, Feyerabend, and others stripped of most of its Wittgensteinian elements (aside from the occasional reference to family-resemblance similarity metrics in the relationship between concepts).

What is this influence of Wittgenstein that fails to appear in the work of Churchland? First and foremost, Wittgenstein is deeply concerned with *language*. Human, natural language is where philosophy begins and ends for Wittgenstein, and that concern is reflected in the theories of our three influencing philosophers. For example, when Hanson argues that our observation is laden with theories, what form does that influencing theory take? With the positivists, Hanson accepts the idea that theories are collections of statements in a language. Further, notice Feyerabend's focus on the role of the common *idiom* on our philosophical views in my earlier discussion of his views. In Sellars, we find him basically agreeing with the central role of language; with Wittgenstein, Sellars is of the view that the mind as we understand it is a product of language and not the other way around.[14] Churchland rejects this idea of the priority of language and the strong emphasis it places on ordinary language. Indeed, it ought to be the first lesson of a thorough-going eliminative materialism: while it seems natural to think of language as the beginning of philosophy – how else could we pose questions, after all? – we ought not take that centrality for granted.

Incidentally, Churchland's rejection of ordinary language philosophy goes some way toward explaining his grounds for rejecting 1960s-style mind-brain identity theory (a lá, Place 1956; Smart 1962). One might think that PMC would embrace this early attempt to eliminate folk-psychologically based mental talk and replace it with a language developed out of neuroscience. (So, instead of speaking of "pain," we ought to speak of "C-fiber activity of such-and-such type," according to the common example of the time.) While PMC clearly honors these forebears of neurophilosophy, he cannot embrace their program for a simple reason: They believe that we will be able simply to take our commonsense idiom and straightforwardly reduce it to brain-talk. In other words, like Wittgenstein and (as I am proposing here) Hanson, Sellars, and Feyerabend, the identity theorists believe that our ordinary ways of speaking of the mind represent a cogent and coherent theory of the mind, such that we ought to seek its translation in neural terms. PMC rejects this, to his mind, rosy view of the status of our folk psychology. Instead, he argues that it will need to be radically changed before it can be reduced to a mature neuroscience. For example,

in the example given, we should expect that the folk psychological category "pain" might simply not hold up to rigorous examination. However, we are getting a bit ahead of ourselves. We'll return to Churchland's eliminative materialism.

A second point of difference in Churchland is illustrated by his different attitude toward common sense and ordinary language, which runs hand-in-hand with the point just made. To illustrate this, consider what Sellars sees as the proper relationship between his two frameworks for understanding the world: the manifest and the scientific images. Recall that, for Sellars, considered properly, we should think of these images as equally important and complementary. They are like the two images of a stereoscopic pair; each is different, but when each image is presented separately and simultaneously to the eyes, our brain fuses the two images into a marvelous, more full perception of the world. On this view, both the manifest and the scientific images individually contribute elements to an understanding of the world that its counterpart cannot. The scientific and manifest images are equally important, different, and necessary to a full understanding of the world, including ourselves. While this view is most explicit in Sellars, I believe it also makes sense of both Feyerabend and Hanson; indeed, it gives a primacy of role to ordinary language and commonsense philosophy to which Wittgenstein always held.

Paul Churchland's view, I suggest, is subtly different on this point, although the small difference is everything in this case. For PMC, the proper relationship between the two views is not that of a balanced combination of a pair of images in a stereogram, because this view does not take seriously Hanson's observation that theories – understood as the way in which one's brain is wired up to perceive the world, not as collections of sentences – invariably "infect" our view of the world. Independent manifest and scientific images cannot exist to be combined stereographically, because each invariably infects the other from the outset. As a result, it is better (according to my reading of PMC) to think of *converting* the manifest image into a scientific image of the world. Churchland's charge is that we must make the scientific image our manifest image.

This conversion of the scientific image into our manifest image (or perhaps "subversion" is a better term to use here) is what I take as the point of one of PMC's most famous (and oft-repeated by him) examples. He asks the reader to consider the night sky when multiple planets are visible:

> . . . [T]here is a simple theory with which almost everybody is sufficiently familiar, but which has yet to put into *observational* gear by all but the most devoted observers of the heavens. I have in mind here Copernicus' theory

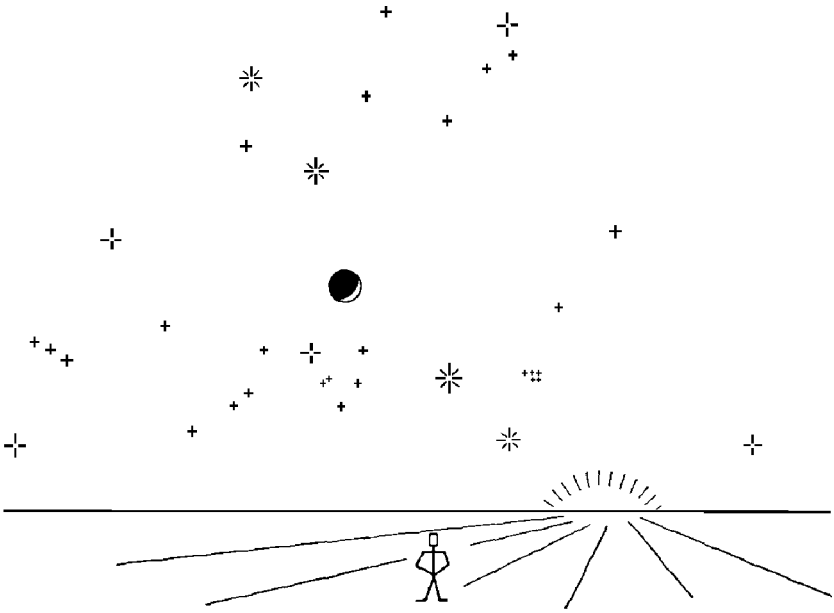


Figure 1.1: The night sky, as depicted by Paul Churchland (1979: 31).

> of the arrangement and motions of the solar system. Our minds, perhaps, have been freed from the tyranny of a flat immobile Earth, but our *eyes* remain in bondage. Most of us could pen quite successfully the relevant system of coplanar circles and indicate the proper directions of revolution and rotation, but when actually confronted with the night sky most of us have only the vaguest idea of how to relate what we have drawn to what we can see. And yet the structure of our system and the behaviour of its elements can readily be made visually transparent, and the magnitude of the "gestalt shift" involved is rather striking. (Churchland 1979: 31–2, emphasis in original)

He goes on to ask us to experience the Copernican gestalt shift for ourselves, providing us with some helpful diagrams to aid us in fully grasping the scientific image of the night sky. What I have reproduced as Figure 1.1 is his depiction of an observer on an appropriate evening just after sunset. It depicts the horizon, the moon, and four of the brightest objects in the sky (in order, from the horizon upward) – Mercury, Venus, Jupiter, and Saturn – all lying roughly along a straight line. Although this is merely curious to most observers, in fact, this linear arrangement is due to the fact that all these bodies lay along the plane of the ecliptic.

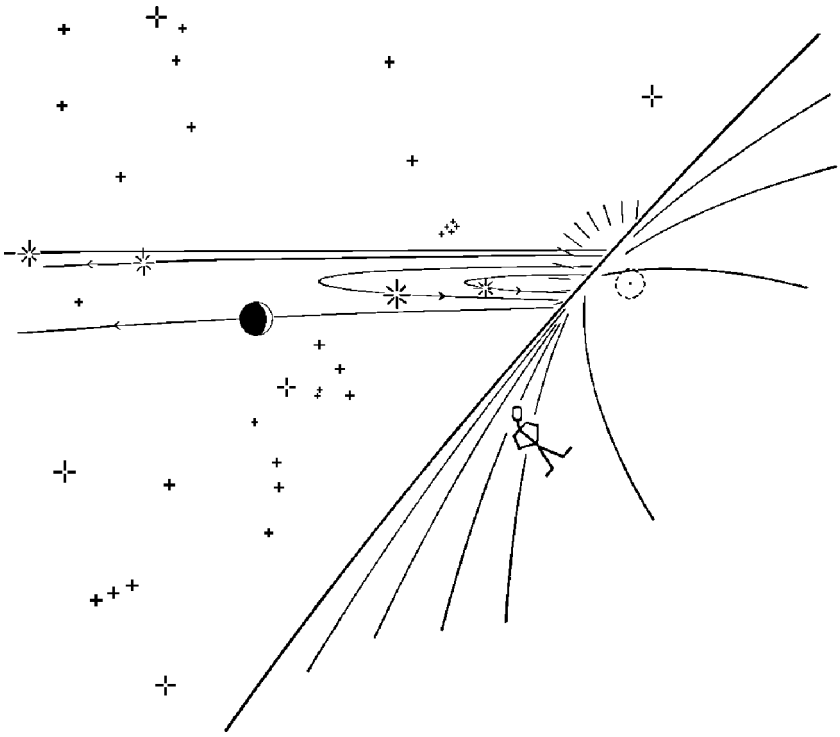


Figure 1.2: The night sky with the benefit of Copernican theory. From (Churchland 1979: 33).

Churchland continues:

> In order to see the situation "as it really is" (as per [Figure 1.2]), what the observer in fig. 1.1 must do is *tilt his head* to the right so that the relevant line (the ecliptic, in fact) becomes a horizontal line in his visual field. This will help him fix his bearings within the frame of reference or coordinate system whose horizontal plane is the plane of the ecliptic, whose origin or centre point is at the Sun, and in which all the stars are reassuringly motionless. If this can be achieved – it requires a non-trivial effort – then the observer need only exploit his familiarity with Copernican astronomy to perceive his situation as it is represented [in Figure 1.2]. . . . [As a result of seeing the world in this new way,] [h]is brow need no longer furrow at the changing appearance of one entire hemisphere of his visual environment: the shifting configurations of the solar family are now visually recognizable by him for what they are. He is at home in his solar system for the first time. (32–4, emphasis in original)

Notice that what PMC is presenting us with is not a way to fuse together two different understandings of the night sky, as Sellars recommends, but rather a way of using the *scientific* image of the world to encounter our *everyday* world in new ways. He seeks to transform our scientific image into a new, thoroughly modern and up-to-date commonsense vision of the universe:

> If our perceptual judgements must be laden with theory in any case, then why not have them be laden with the best theory available? Why not exchange the Neolithic legacy now in use for the conception of reality embodied in modern-era science? Intriguingly, it appears that this novel conceptual economy could be run directly on the largely unappreciated resources of our own sensory system as constituted here and now. . . . Should we ever succeed in making the shift, we shall be properly at home in our physical *universe* for the very first time. (Churchland 1979: 35, emphasis in original)

However, perhaps at the end of the day PMC would argue that I'm making a mountain of an interpretational molehill here. There are certainly moments in Sellars where he, too, notes the close connections between the manifest and scientific images. PMC could easily point to places where he and Sellars seem rather close in spirit. For example, at one point, Sellars observes that, "The truth of the matter . . . is that science is continuous with common sense, and the ways in which the scientist seeks to explain empirical phenomena are refinements of the ways in which plain men, however crudely and schematically, have attempted to understand their environment and their fellow men since the dawn of intelligence" (1956/1997: 97). All explanation and understanding are of a piece and everything, ultimately, is up for grabs.

PAUL CHURCHLAND'S ELIMINATIVE MATERIALISM

Having discussed Paul Churchland's influences, in this final part of this Chapter, I want to turn to the major philosophical element of who he became. More than anything, PMC is known as a present-day proponent of the position Feyerabend initially named: eliminative materialism. Because it is central to Churchland's views on all topics, it will be useful to review his position, and my few critical comments following this discussion should help bring what ideas PMC presents into clearer focus.

In sketch, PMC's eliminative materialism looks like this:

P1) Folk psychology exists

P2) Folk psychology is a theory

P3) Theories are falsifiable (fallible).

/∴ C1) Folk psychology is falsifiable, i.e., *eliminable*.

P4) There is good reason to believe that folk psychology is indeed an incorrect scientific psychological description (of us, of animals).

/∴ C2) Folk psychology should be considered to be false, i.e., *eliminated*.

The first part of the argument (P1–3, C1) follows directly from the work of the three philosophical influences I discussed earlier. Hanson gives us the idea that our perception of the world is influenced by the theories (commonsense and otherwise) we hold. Feyerabend contributes a distrust of common sense (in the form of holding it up as something that could be wrong). Sellars goes on to give this commonsense view a name: it is the manifest image and he more fully explores its nature. The Churchlandian argument for eliminative materialism puts these parts together: This manifest image, with which perception is laden, is just as fallible and likely to be wrong as any other theory that may infect its user's view of the world.

However, PMC goes beyond these original arguments for the *possible* elimination of folk psychology to present a variety of reasons for believing that the imagined elimination should *in fact* be carried out. He has three main arguments for P4, the linchpin premise. These arguments are taken from his landmark 1981 *Journal of Philosophy* paper, "Eliminative materialism and the propositional attitudes" (Churchland 1981/1989). First, Churchland argues that folk psychology cannot account for a plethora of psychological phenomena: sleep, the dynamics of mental illness, perceptual illusions and hallucinations, intelligence differences between individuals, the dynamics of learning, and so on (6–7). He notes that folk psychology sheds "negligible light" on how the brain constructs an elaborate three-dimensional visual world from photons falling on the retina or how humans can pull up memories from a vast store in only a matter of moments (7). So, as theories go, folk psychology leaves much to be desired.

Second, unlike a relatively new theoretical framework, such as neuroscience, folk psychology cannot claim it is still at an early stage of development. In other words, it cannot explain away the lacuna in its explanatory coverage by offering a promissory note. Folk psychology is not an area of ongoing development. Indeed, Churchland claims that "... both the content and the success of [folk psychology] have not advanced sensibly in two or three thousand years. The FP of the Greeks is essentially the FP

we use today" (8). In this light, Churchland accuses folk psychology of being a Lakatosian "degenerative research programme" (see Lakatos and Musgrave 1970); that is, it has shown little, if any, change over the extent of recorded history. Therefore, one should not hope that it will soon address the failings noted in the first complaint. PMC is blunt: "The history [of folk psychology] is one of retreat, infertility, and decadence" (7).[15]

Finally, folk psychology is inconsistent with surrounding explanatory frameworks, such as those provided by contemporary biology, neuroscience, and psychology. And, as explanatory frameworks must both be internally consistent as well as consistent with adjoining frameworks, this is a very telling flaw in folk psychology. Unless folk psychology is to take on the status of modern theology, it simply will not do to envision an isolated framework that applies only to the "normal," waking behavior of human beings, while at the same time being highly inconsistent with the way we understand other, related phenomena.

These three reasons, according to PMC, taken in combination seem to warrant the exploration of other frameworks; frameworks that, one would hope, capture all that folk psychology presently captures, and more. In other words, the arguments discussed earlier are largely *negative* in nature; they explain how we shouldn't go about understanding ourselves and the world around us, despite how appealing – indeed, indispensable – the commonsense approach may seem. PMC is not content to leave us with only that negative contribution to philosophy. Much of his work explores a more positive contribution: the development of a successor explanatory framework with which to replace folk psychology. That positive framework is developed out of the neural sciences and involves conceiving of the mind/brain as a "neural network." As PMC puts it,

> Eliminative materialism is the thesis that our commonsense conception of psychological phenomena constitutes a radically false theory, a theory so fundamentally defective that both the principles and the ontology of that theory will eventually be displaced, rather than smoothly reduced, by completed neuroscience. (Churchland 1989: 1)

CRITICAL COMMENTS ON PMC'S ARGUMENTS FOR ELIMINATIVE MATERIALISM

My main complaint with PMC's presentation of eliminative materialism is that it fails to capture the situation adequately. In particular, it fails to make what I think is a crucial distinction: the distinction between *folk psychology*

per se (FP_x) and *belief-desire* folk psychology ($FP_{b\text{-}d}$). Folk psychology is the way in which the folk conceive themselves as psychological beings, however that may be. Belief-desire folk psychology is a *particular way* that the folk, especially contemporary, Western folk, tend to conceive themselves as psychological beings.

Making this distinction helps head off one potentially disastrous misinterpretation of the EM program. On hearing it presented as it normally is (that is, without clearly distinguishing between FP_x and $FP_{b\text{-}d}$), eliminative materialism sounds suspiciously like reductionist arguments calling for the elimination of psychology. For example, one may think here of the mind-brain identity theorists mentioned earlier. These reductionist eliminative arguments bear many superficial similarities to EM, but they call for the elimination of an entire level of explanation; usually arguing that some lower level, typically neuroscience or physics, is all that is needed for full and complete scientific explanation. This is clearly *not* what PMC has in mind. Incidentally, this is why he is in favor of functionalism (properly construed, of course). Computational neuroscience is offered not merely as a low-level replacement for psychology, but instead as a multi-level theory capable of bridging the immense gap between mind and brain. On this interpretation, EM is actually a much more moderate claim than is often feared.[16]

PMC himself does not steer us away from potential misinterpretations when he argues, as mentioned, that FP is a Lakatosian degenerative research program. His argument there seems to go too far. Someone arguing, as PMC does, that "x is a *bad* f" must be careful not to overdo it, or somebody will point out that it makes more sense to conclude that "x just isn't an f." Is $FP_{b\text{-}d}$ a really bad example of a theory, or is it the case that it just isn't a theory at all? Some, such as Wilkes (1984), draw just such a conclusion from PMC's arguments.

I see several reasons for backing away from Churchland's strong arguments for the degenerative nature of $FP_{b\text{-}d}$. First, arguments for replacing a particular conceptual scheme are not precise, so one can back off from the strong stance PMC has taken and still conclude that $FP_{b\text{-}d}$ should be eliminated, on the basis of the other arguments he has presented.[17] Second, saying that $FP_{b\text{-}d}$ is an unchanging monolith seems inconsistent with the fact that such a theory must have come into existence at some point in time. Presumably, PMC wouldn't argue that $FP_{b\text{-}d}$ was born in its full glory, Athena-like, out of the head of *Hominid* man or woman. It must have undergone some kind of development to come to its present state of fulfillment, and it seems unduly procrustean to claim that all that development was completed prior to the invention of writing.[18]

Third, arguing for the historically changing nature of FP allows one to account for all sorts of historical facts about psychology that on PMC's account are left unexplained. Historians of psychology are fond of pointing out the many different ways the folk have psychologically understood themselves and their conspecifics: demon possession and witches, differences in conceptions of the mind before and after Freudian and Jungian conceptions of the unconscious, humor psychologies of the Enlightenment, the ancient Greek focus on inherited character, and so on.

Finally, the picture of FP as an evolving enterprise is appealing because it allows us to say something positive about the *future* of this endeavor. Part of the problem with the Churchland program is one of terminology, for the most charitable interpretation of his program construes it as one of *revision* and not *elimination*.[19] If one claims, as I do here, that FP has evolved through history, then one can propose that what PMC's neurocomputational perspective offers us is the next step in that evolution. I am suggesting that PMC ought to be arguing that the FP of the future will/ought to be a computational, neuroscientifically informed folk psychology (FP_{CNS}). The required change will be on par with the change required when humanity shifted from a Ptolemaic to a Copernican worldview, but this does not seem so untenable when placed against the background of less-dramatic changes. Folk psychology was revised in light of the theoretical paucity of humor psychology several centuries ago, and the time has come, the eliminativist argues, to bring about the next great change. The point I wish to stress here is that what needs to be *eliminated* is the reference to beliefs and desires (and related concepts). This is done by *revising* folk psychology in light of contemporary science. Feyerabend (1963a) himself seemed to have a conception of the history of FP as evolving. When arguing against those who criticized the then embryonic brain sciences, he counters: "It took a considerable time for ordinary English to reach its present stage of complexity and sophistication. The materialist philosopher must be given at least as much time" (54).

THIS VOLUME

If successful, the preceding discussion should give the reader a synoptic overview of Paul Churchland's philosophical worldview. It is intended to set the stage for the other contributions in this volume.

Eliminative materialism and our ability to conceive of ourselves in ways divorced from those of folk psychology are the topics of Chapters 3 and 4.

In "Arguing for eliminativism," **José Luis Bermúdez** is generally happy with the ultimate goals of eliminative materialism; he too finds folk psychological explanations to be inadequate in the light of what we have come to learn from contemporary cognitive science. However, he is also unimpressed with the kinds of arguments PMC gives for EM (some of which I rehearsed here). So, Bermúdez takes on some more recent criticisms of eliminativism, such as the work of Paul Boghossian (1990) as a springboard for exploring good and not-so-good ways of defending eliminativism. In his programmatic paper, Bermúdez tells us what he believes PMC *ought* to be saying to arrive at the conclusions Bermúdez and he share.

In his contribution, "The introspectibility of brain states as such," **Pete Mandik** tackles a related topic. He asks whether we can really take PMC seriously when Churchland says that we see the world through neurally informed eyes. If that were the case, then when we turn our minds inward and introspect, we in fact perceive our own brain in action. However, according to Mandik, PMC goes further, defending what Mandik terms the *Introspection Thesis*: "A person with sufficient neuroscientific education can introspect his or her brain states *as* brain states." This is just the mind-brain version of the "tilt your head to the right and see the planets as lying on the ecliptic" example discussed earlier. Although the introspection thesis is a clear consequence of PMC's eliminative materialism, as Mandik notes, it is far from obvious what it would even mean to introspect one's brain states as such, much less whether such an odd-seeming claim is true. Mandik ends up concluding that Churchland's thesis, surprisingly enough, is more plausible than opposing, yet supposedly more obvious, theses about introspection one finds in the philosophy of mind literature these days.

Turning from eliminative materialism proper, the next two contributions explore PMC's successor theory and its scientific source – the science of *connectionism* – in more detail. In his "Empiricism and state space semantics," **Jesse J. Prinz** enters into the fray primarily between Churchland (on one side) and Jerry Fodor and Earnest LePore (on the other) over how to understand the nature of mental content. PMC derives his account of mental content – he has dubbed it "state space semantics" – from computational neuroscience and connectionism. According to PMC, connectionism provides us with an account of how brains acquire mental content through learning. The resulting account explains content as patterns or vectors of activation ("prototypes") across appropriate populations of neurons (in us) or neuron-like units (in the case of connectionist systems). Fodor, LePore, and others have been at pains to show that such an account just cannot be made to work for the same reasons that David Hume's superficially similar

associationist accounts of mental content fail. Instead of defending Hume, Churchland tends to stress how his own account is different. Prinz instead argues that Churchland should lose his fear of Hume and embrace the kind of empiricist account of mental content – properly understood, of course – that Prinz himself has been proffering of late (Prinz 2002). Prinz argues that by doing this, PMC would find a way of making the world safe for state space semantics.

Aarre Laakso and Garrison W. Cottrell are also concerned with PMC's use of connectionism in the development of state space semantics. However, they are less sanguine than Prinz about the theory of mental content PMC has developed, primarily because they fear that Churchland's treatment of connectionism has been, at its worst moments, incorrect, and more often, incomplete. In their chapter, "Churchland on connectionism," they take issue with Churchland's presentation of connectionist findings, suggesting that he has a far too rosy view of the ease with which connectionist models can be translated into a theory of the mind.[20] The source of this criticism is important to note explicitly. *Prima facie*, Laakso and Cottrell know whereof they speak. Cottrell is a highly respected computer scientist and connectionist; indeed, Cottrell's research has figured prominently in the connectionist literature PMC has drawn on for several decades. Laakso is a philosopher who spent much of his graduate school tenure working in Cottrell's lab learning the intricacies of neural modeling. More to the point, Laakso and Cottrell's work on measuring similarity in representational capacities across networks with differing architectures became the primary source in PMC's most recent salvo in the Churchland versus Fodor/LePore debates.[21] Laakso and Cottrell's contribution then provides us with a valuable opportunity to fact check PMC's use of connectionism.[22]

The next two contributions look at PMC as a philosopher of science. The chapter contributed by **Clifford A. Hooker**, "Reduction as a cognitive strategy," critically explores the relationship PMC draws between theoretical reduction (or elimination[23]) in science and the cognitive strategies of individual human thinkers. Hooker is ideally suited to undertake this exploration: PMC has drawn on Hooker's work on reductionism in developing his own view (particularly, Hooker 1981a, 1981b, 1981c, as well as Hooker 1975), so like the Laakso and Cottrell contribution, this is another opportunity for a kind of constructive dialogue between Churchland and his interlocutors on key elements of PMC's philosophy.[24]

Philosophy of science is also the subject of "The unexpected realist" by **William H. Krieger and Brian L. Keeley**. In this contribution, PMC's view on scientific realism is considered. As the title of his first

book – *Scientific realism and the plasticity of mind* – suggests, he has been a proponent of the view that the entities invoked in scientific theories are *metaphysically real* in some robust sense of that concept. The non-visible theoretical posits of science, PMC maintains, are not merely instrumentally useful measuring devices nor are they merely pragmatically parsimonious posits that help make our scientific understanding of the world in line with common sense. Quite the contrary, PMC holds that electrons and quarks are just as real as shoes, ships, and sealing wax. Churchland is termed an *unexpected* realist because, given the kind of pedigree I ascribe to him in the discussion earlier in this chapter, one might expect him to be a scientific antirealist of some stripe. After all, if perception is invariably infected by theory, and theory is ever evolving, then it is hard to see what grounds we have for claiming that the particular posits of today's theory are any more real than the supposedly false theorizing that came before. Indeed, embracing this lack of a place to stand, metaphysically speaking, is part of Churchland's hero, Feyerabend's, motivation for claiming that "Anything goes" when it comes to science. But, despite this, PMC finds a principled reason to be a scientific realist after all.

The final chapter of this collection brings us back to ultimate point of Paul Churchland's philosophy: to explain the nature of mind. All this discussion of eliminative materialism, folk psychology, connectionism, reductionism, and scientific realism is certainly interesting in its own right, but at the end of the day, these are all mere stepping stones to what motivates Paul Churchland in the first place: to attempt to develop an understanding of what the "mind" is and how it is related to the body, more specifically, the brain. At its most devilish, this is the problem of the nature of consciousness. And, on this topic, PMC has probably had no more worthy an opponent than **Daniel C. Dennett**. To my eyes, I have always thought that Dennett and Churchland probably agree on far more than they differ. But one consequence of agreeing with somebody on 90% of a philosophical position is that you can *really* dig in and disagree vociferously on the remaining 10%. As Dennett traces in his contribution, "Two steps closer on consciousness," these two able philosophers have been dialectically engaged over the problem of consciousness for some 25 years, and I am pleased to include here the latest step in the progression.[25] Dennett's goal here is to show PMC (and the rest of us) that the 10% difference to which I just alluded is exaggerated. Rhetoric aside, Dennett argues that on the topic of the mind as (what Dennett refers to as) a "virtual machine," he and PMC are in significant agreement. Much of what PMC takes to be signs of disagreement are, according to Dennett, based on misunderstanding on Churchland's part.

The disagreement between these two on the relationship between memes and consciousness is more substantive, according to Dennett. Here, Dennett argues that PMC's position would be stronger and more self-consistent if Churchland would only embrace Dennett's eliminativism with respect to our folk theories of conscious phenomena.

Taken together, the contributions in this collection represent a broad overview of the often-radical philosophy of Paul M. Churchland. I suspect that this is far from the last word both on (as well as from) this dynamic and provocative philosopher. Although the title of this chapter is "Becoming Paul M. Churchland," the man and the philosopher will no doubt continue to develop, and the way in which we perceive him will likewise develop as others create new theories of who he is and what he represents to philosophy.

Notes

1. In a volume dedicated to one-half of such a partnership, it should be noted explicitly that Pat has had a strong influence on Paul's philosophical development; perhaps the strongest influence of anybody (and vice versa, of course). As Paul himself put it in 1995: "After twenty-five years of affection and collaboration, I often feel we have become the left and right hemispheres of a single brain" (1995: xii). In terms of bringing neuroscience to the philosophical masses, Patricia's *Neurophilosophy* (1986a) has had perhaps the largest impact of any single work, including *Matter & Consciousness*. She clearly deserves her own volume in this series.
2. The date of the move is important for explaining why the Churchlands fail to be featured in Dan Dennett's "Logical geography" of computational approaches in the philosophy of mind (Dennett 1986). That piece was originally written in 1984. However, one suspects that if it had been written even a few years later, the Churchlands would have been noted as high mucky mucks of "West Pole" (sub-category: "San Diego School," a lá Don Norman and David Rumelhart) Computationalism.
3. Quine clearly appreciated PMC's philosophical contribution. In a "blurb" that appears on the back cover of the paperback edition of Churchland's 1995 *The engine of reason, the seat of the soul*, Quine describes PMC's book as "an outstanding philosophical achievement, integrating artificial intelligence, brain neurology, cognitive psychology, ethnology, epistemology, scientific method, and even ethics and aesthetics, into an interlocking whole."
4. Of these three, I suspect that the inclusion of Sellars among the "relatively not read" is the most likely to invoke rebuke. In my defense, I note that in his introduction of a recent re-issue of Sellars' *Empiricism and the Philosophy of Mind*, Richard Rorty includes this work (along with Quine's "Two Dogmas" and Wittgenstein's *Philosophical Investigations*) as the three seminal works in the mid-twentieth-century shift to post-positivist philosophy. However, he observes, "Of these three, Sellars' long, complicated, and very rich essay is the least known and discussed," by "historians of recent Anglo-American philosophy" (Rorty 1997: 1–2).

5. Hilary Putnam (1959) said of this book, "In my opinion, this is the most exciting book on the philosophy of science to appear in the last ten years. It is exciting for various reasons, but the most important single reason is that at last we have a philosopher of science who is in fact writing about science and not about the papier-mâché constructions that frequently replace science in the writings of philosophers and logicians of science" (1666). It should probably be noted that Putnam (1963) was less impressed with Hanson's next book, *The concept of the positron* (1963), although this was primarily because he saw Hanson as trying to defend the Copenhagen interpretation of quantum mechanics, a task that Putnam finds to be a fool's errand.
6. Which may have appealed to the young Paul Churchland more than other philosophers – what with his Bachelor of Arts (with Honors) in Philosophy, Physics, and Mathematics.
7. Among other things, no doubt!
8. Philosophy is, for Sellars, a deeply *important* endeavor. Perhaps it is not entirely tongue-in-cheek here to wonder if Sellars believes, with Benjamin Franklin, that if we don't hang together, we shall hang separately, intellectually speaking.
9. Although it is often useful for exegetical reasons to discuss the manifest image as that which may have been held by our evolutionary predecessors, this emphasis is misleading. It is still at play today in all of us as we negotiate our worlds.
10. Sellars (1960/1963) observes that "... the basic objects of current theoretical physics are notoriously imperceptible and unimaginable" (9).
11. I cannot help but observe that 35 years later Sellars' Ph.D. student, for different reasons, would publish a book, each copy of which included not only stereoscopic images but also a pair of stereoscopic glasses for viewing them (Churchland 1995). The production of stereoscopic photographs has been a longstanding hobby for Paul Churchland.
12. Although I have not discussed it here, Sellars is also concerned with commonsense psychology. In what is likely the most famous argument in *Empiricism and the Philosophy of Mind* – the "Myth of Jones" thought experiment of Section XII – Sellars presents a fictionalized account of how primitive humans may have developed a theory of the psychology of their brethren.
13. In his autobiography (Feyerabend 1995), he sometimes suggests that the reason that he was hired at prestigious institutions such as the University of California, Berkeley, was because of his "big mouth," not because of his philosophical acumen or the quality of his ideas.
14. I see this issue of the relative ontological priority of language vs. mind as the watershed that separates some students of Sellars (e.g., PMC) from other contemporary Sellarsians (e.g., Robert Brandom).
15. It is this claim upon which I will put pressure later.
16. On my reading of him, Sorell's (1991) critique of Churchland makes exactly the misinterpretation of position that I describe here.
17. Albeit with much less certainty. If one looks at the structure of the argument of Churchland (1981/1989), it becomes clear that a lot of the weight of the argument rides on the degenerative nature of FP.

18. The source of this concern is actually PMC himself; he raised it in a seminar I took from him at UCSD in 1992. However, to my knowledge, he has never addressed it.
19. As Pat Churchland (1986b) explains it, they once considered adopting the name "revisionary materialism" for their position. However, given the inertia already behind Feyerabend's chosen moniker, "eliminative materialism," they decided to keep it.
20. Or of the *brain*, for that matter. Laakso and Cottrell argue that, early optimism in the field aside, the biological plausibility of neural nets is not as obvious as it could be.
21. The salvo: "Conceptual similarity across sensory and neural diversity: The Fodor/Lepore challenge answered" (Churchland 1998). The ammunition for the salvo can be found in (Laakso and Cottrell 1998, 2000).
22. Prinz, on the one hand, and Laakso and Cottrell, on the other, also constitute a pincer attack on Churchland with respect to his development of a philosophical position out of computational neuroscience: In essence, Prinz is criticizing him for not being radical enough and not really taking the results of connectionism to their proper conclusion. Laakso and Cottrell, to the contrary, criticize PMC for reading far too much into connectionism than the science will bear.
23. To what extent PMC is more properly thought of as a "reductionist" and/or an "eliminativist" is something Hooker explores at length. This is more than a mere terminological concern, as Hooker helps place PMC's work within the larger tradition of other philosophers of science who address the question of the relationship between levels of scientific explanation, such as Batterman (2001) and Bickle (1998, 2003).
24. What's more, Hooker and Churchland have previously worked together, co-editing a volume of papers on the work of Bas van Fraassen (Churchland and Hooker 1985).
25. Churchland's most recent contribution to this debate – to which Dennett's contribution here can be seen as a direct response – is found in PMC's contribution, "Catching consciousness in a recurrent net," to be found in the volume in this series dedicated to exploring the work of Dennett, edited by Andrew Brook and Don Ross (Churchland 2002).

Works Cited

Batterman, R. W. (2001). *The devil in the details: Asymptotic reasoning in explanation, reduction and emergence*. Oxford, Oxford University Press.

Bickle, J. (1998). *Psychoneural reduction: The new wave*. Cambridge, MA, MIT Press (A Bradford Book).

_____. (2003). *Philosophy and neuroscience: A ruthlessly reductive account*. Dordrecht, The Netherlands, Kluwer Academic Publishers.

Boghossian, P. (1990). "The status of content." *The Philosophical Review* **99**: 157–84.

Churchland, P. M. (1979). *Scientific realism and the plasticity of mind*. New York, Cambridge University Press.

_____. (1981/1989). "Eliminative materialism and the propositional attitudes." *The Journal of Philosophy* **78(2)**: 67–90. Reprinted as Chapter 1 of Paul M. Churchland, *A neurocomputational perspective: The nature of mind and structure of science*. Cambridge, MA, The MIT Press (A Bradford Book), 1989.

_____. (1989). *A neurocomputational perspective: The nature of mind and the structure of science*. Cambridge, MA, The MIT Press (A Bradford Book).

_____. (1995). *The engine of reason, the seat of the soul: A philosophical journey into the brain*. Cambridge, MA, The MIT Press (A Bradford Book).

_____. (1998). "Conceptual similarity across sensory and neural diversity: The FodorLepore challenge answered." *The Journal of Philosophy* **95**: 5–32.

_____. (2002). "Catching consciousness in a recurrent net." *Daniel Dennett*. A. Brook and D. Ross (Eds.) New York, Cambridge University Press, 64–81.

Churchland, P. M., and P. S. Churchland (1998). *On the contrary: Critical essays, 1987–1997*. Cambridge, MA, The MIT Press (A Bradford Book).

Churchland, P. M., and C. A. Hooker, Eds. (1985). *Images of science: Essays on realism and empiricism, with a reply from Bas C. Van Fraassen*. Chicago, The University of Chicago Press.

Churchland, P. S. (1986a). *Neurophilosophy: Toward a unified science of the mindbrain*. Cambridge, MA, The MIT Press (A Bradford Book).

_____. (1986b). "Replies to comments." *Inquiry* **29**: 241–72.

Clark, A. (1989). *Microcognition: Philosophy, cognitive science, and parallel distributed processing*. Cambridge, MA, The MIT Press (A Bradford Book).

Dennett, D. C. (1986). "The logical geography of computational approaches: A view from the east pole." *The representation of knowledge and belief*. M. Brand and M. Harnish (Eds.) Tucson, University of Arizona Press: 59–79. Reprinted as Daniel C. Dennett, *Brainchildren: Essays on designing minds*. Cambridge, MA, The MIT Press (A Bradford Book), 1998, 215–34.

deVries, W. A., and T. Triplett (2000). *Knowledge, mind, and the given: Reading Wilfrid Sellars's "empiricism and the philosophy of the mind."* Indianapolis, Hackett Publishing Company, Inc.

Duhem, P. (1914). *La théorie physique*. Paris, M. Rivière.

Feyerabend, P. K. (1962). "Explanation, reduction and empiricism." *Minnesota studies in the philosophy of science: Scientific explanation, space and time*. H. Feigl and G. Maxwell (Eds.) Minneapolis. Volume **3**: 28–97.

_____. (1963a). "Materialism and the mind-body problem." *Review of Metaphysics* **17**: 49–66.

_____. (1963b). "Mental events and the brain." *The Journal of Philosophy* **LX**: 11.

_____. (1975). "How to defend society against science." *Radical Philosophy* **11**: 3–8. Reprinted in *Introductory readings in the philosophy of science, revised edition*. E.D., Klemke, Robert Hollinger, & David A. Kline. (Eds.). Buffalo, NY, Prometheus Books, 1998, 54–65.

_____. (1995). *Killing time: The autobiography of Paul Feyerabend*. Chicago, University of Chicago Press.

Fodor, J. A. (1975). *The language of thought*. Cambridge, MA, Harvard University Press.

Hanson, N. R. (1958). *Patterns of discovery: An inquiry into the conceptual foundations of science*. Cambridge, UK, Cambridge University Press.

_____. (1963). *The concept of the positron: A philosophical analysis*. New York, Cambridge University Press.

_____. (1971). *Observation and explanation: A guide to philosophy of science*. New York, Harper Torchbooks.

Hooker, C. A. (1975). "The philosophical ramifications of the information-processing approach to the brain-mind." *Philosophy and Phenomenological Research* **36**: 1–15.

_____. (1981a). "Towards a general theory of reductionism, part i: Historical framework." *Dialogue* **20**: 38–59.

_____. (1981b). "Towards a general theory of reductionism, part ii: Identity and reduction." *Dialogue* **20**: 201–36.

_____. (1981c). "Towards a general theory of reductionism, part iii: Cross-categorial reduction." *Dialogue* **20**: 496–529.

Kekulé, F. A. (1890/1996). Origin of benzene and structural theory. *Introduction to the philosophy of science*. A. Zucker (Ed.) Upper Saddle River, NJ, Prentice Hall, 34–5.

Laakso, A., and G. W. Cottrell (1998). "How can I know what you think?: Assessing representational similarity in neural systems." *Proceedings of the twentieth annual cognitive science conference, Madison, Wi.* Mahwah, NJ, Lawrence Erlbaum.

_____. (2000). "Content and cluster analysis: Assessing representational similarity in neural systems." *Philosophical Psychology* **13**: 47–76.

Lakatos, I., and A. Musgrave, Eds. (1970). *Criticism and the growth of knowledge*. Cambridge, UK, Cambridge University Press.

Lewis, C. I. (1929). *Mind and the world-order*. New York, Charles Scribner's Sons.

_____. (1945). *An analysis of knowledge and valuation*. La Salle, IL, The Open Court Publishing Company.

McClelland, J. L., and D. E. Rumelhart (1986). *Parallel distributed processing: Explorations in the microstructures of cognition*. Cambridge, MA, The MIT Press (A Bradford Book).

Place, U. T. (1956). "Is consciousness a brain process?" *The British Journal of Psychology* **47**: 44–50. Reprinted in *Identifying the Mind: Selected Papers of U.T. Place*. U. T. Place, Elizabeth R. Valentine, George Graham. Oxford, Oxford University Press, 2003, 45–52.

Preston, J. (2000). Feyerabend. *A companion to the philosophy of science*. W. H. Newton-Smith (Ed.) Malden, MA, Blackwell Publishers, 143–8.

Prinz, J. J. (2002). *Furnishing the mind: Concepts and their perceptual basis*. Cambridge, MA, The MIT Press.

Putnam, H. (1959). "Review of Norwood Russell Hanson, *Patterns of discovery. An inquiry into the conceptual foundations of science*." *Science* **129**(3364): 1666–7.

_____. (1963). "Review of Norwood Russell Hanson, *The concept of the positron: A philosophical analysis*." *Science* **139**(3556): 745.

Rorty, R. (1997). "Introduction." *Empiricism and the philosophy of mind*. W. Sellars (Ed.) Cambridge, MA, Harvard University Press, 1–12.

Rumelhart, D. E., and J. L. McClelland (1986). *Parallel distributed processing: Explorations in the microstructures of cognition*. Cambridge, MA, The MIT Press (A Bradford Book).

Sellars, W. (1956/1997). *Empiricism and the philosophy of mind*. Cambridge, MA, Harvard University Press. *Minnesota Studies in the Philosophy of Science*, *vol. 1: The Foundations of Science and the Concepts of Psychology and Psychoanalysis*, Herbert Feigl and Michael Scriven (Eds.) Minneapolis, University of Minnesota Press, 1956.

_____. (1960/1963). "Philosophy and the scientific image of man." *Science*, *perception and reality*. Atascadero, CA, Ridgeview Publishing Company, 1–40.

Smart, J. J. C. (1962). Sensations and brain processes. *The philosophy of mind*. V. C. Chappell (Ed.) Englewood Cliffs, NJ, Prentice-Hall. (This is a slightly revised version of a paper first published in *The Philosophical Review*, **68**, 1959, 141–56.)

Smolensky, P. (1988). "On the proper treatment of connectionism." *Behavior & Brain Sciences* **11**(1): 1–23.

Sorell, T. (1991). *Scientism: Philosophy and the infatuation with science*. London, Routledge.

Wilkes, K. (1984). "Pragmatics in science and theory in common sense." *Inquiry* **27**: 339–61.

2 Arguing for Eliminativism*

JOSÉ LUIS BERMÚDEZ

I am sure I am not alone in reporting that the greater exposure I have to experimental work in scientific psychology and neuroscience the less value there seems to be in our commonsense psychological framework of belief, desire, and the other propositional attitudes. Commonsense psychological concepts hardly feature at all in cognitive science and cognitive neuroscience. Theorists in these areas either eschew psychological vocabulary altogether or appeal to shadowy neologisms such as "cognize" or "encode." It is often difficult to see where the points of contact are between the serious scientific study of the mind and the apparent platitudes that philosophers tend to take as their starting point. And even when one can see where the points of contact are, scientific psychology and commonsense psychology are often in tension with each other. Many areas of scientific psychology place serious pressure on our image of ourselves as rational, consistent agents with stable character traits. Yet this image of ourselves is at the heart of commonsense psychology. Commonsense psychology tells one story about the "springs of action" – about how and why we behave the way we do – while the story (or rather, stories) told by scientific psychology and cognitive neuroscience seem completely different and in many ways incompatible with a commonsense understanding of human behavior.

In the face of all this some philosophers, most prominently of course Paul Churchland, have argued for a wholesale rejection of our commonsense ways of thinking about the mind. Very few philosophers have found Churchland's eliminative materialism to be a palatable option. Whatever tensions there might be between commonsense psychological thinking and the scientific investigation of cognition and behavior, it is natural to ask whether we can really abandon the conceptual framework of commonsense psychology. What would happen if we tried to do without our concepts of belief and desire and the ways of thinking about how and why people behave that those concepts bring with them? This paper considers how best an eliminativist might argue for the radical falsity of commonsense psychology. I will be arguing that Paul Churchland's "official" arguments

for eliminative materialism (in, e.g., Churchland 1981) are unsatisfactory, although much of the paper will be developing themes that are clearly present in Churchland's writings. My aims are, first, to refocus the debate on what I consider more interesting and fruitful arguments for eliminativism and, second, to explore how far those arguments might take us. This introductory section explores some general issues about the form that a plausible argument for eliminative materialism should take.

In clarifying the issues in this area it will be helpful to begin with an influential attack on eliminative materialism. Paul Boghossian's "The status of content" argues forcefully that eliminative materialism cannot be coherently formulated. As he brings out, eliminative materialism is best viewed as a species of irrealism. Irrealism with respect to a particular area of discourse is the thesis that there are no real objects or properties corresponding to the kind terms in that discourse. Just as irrealism about evaluative discourse about art or morality holds that there is nothing in the world corresponding to the evaluative predicates used in aesthetics and ethics, eliminative materialism holds that there is nothing in the world corresponding to the psychological kind terms used in commonsense psychology.

As Boghossian notes, irrealist theories in general, and hence eliminative materialism in particular, come in two flavors. Irrealism can be developed either as a form of error theory, or as a form of nonfactualism. Error-theoretic forms of irrealism hold that, although meaningful and substantive claims are made by the particular discourse in question, those claims are all false. So, for example, an error theory of moral discourse would hold that evaluative statements are no less truth-apt than any other form of statements. It is just that none of them are true. In contrast, nonfactualist versions of irrealism hold that surface grammatical form is misleading and deny that the discourse in question is truth-apt. The nonfactualist denies that the apparently assertoric form of the discourse in question should be taken at face value. An example of a nonfactualist approach to moral discourse would be the thesis that moral discourse should be understood as expressing particular attitudes to people and/or situations, rather than as making claims about the (moral) properties of those people and/or situations.

Plainly, eliminative materialism about the propositional attitudes is an error-theoretic variety of irrealism. Commonsense psychology makes certain claims about the world and, according to the eliminative materialist, those claims are uniformly false. These claims occur primarily in the context of explaining and predicting behavior. They invoke both particular psychological states and generalizations defined over classes of such states.

Yet, if eliminative materialism is correct, there are no such states and no true generalizations defined over them.

Arguing convincingly for eliminative materialism requires identifying and diagnosing the errors of which commonsense psychology is being convicted. We need to know (a) what the relevant claims are, and (b) why they are supposed to be false. In the remainder of this section and the next section I will focus predominantly on the first question. The rest of the paper will emphasize the second question. As will soon become apparent, however, it is very difficult to keep these two questions distinct.

Boghossian explicitly argues that eliminativism has to be construed as an error theory about all forms of content. He objects to all attempts to limit eliminativism to a thesis about the psychological, and in particular to the idea that we can continue to think about language and linguistic behavior in semantic terms while rejecting the conceptual framework of commonsense psychology. In fact, his anti-eliminativist argument depends on this very broad reading of the scope of eliminative materialism, for what he effectively offers is an argument to the effect that global eliminativism about content is incoherent. Since his argument against global eliminativism is fairly convincing, our first task must be to determine whether eliminative materialism can be formulated without a commitment to global eliminativism.

Prima facie, there is no direct entailment from eliminativism about the psychological to eliminativism about the linguistic. It is true that there are models of linguistic meaning and the requirements of communication that effectively bring linguistic behavior and linguistic understanding within the scope of propositional attitude psychology. If one thinks that the meaning of a sentence is given (at least in part) by the communicative intentions of its utterer, and that communication is achieved when the listener manages to work backward from the heard sentence to the communicative intention that it embodies, then it is obvious that the domain of the linguistic stands or falls with the domain of the psychological. But this model of linguistic meaning is far from compulsory. Accounts of meaning in terms of use, for example, offer a natural retreat for the eliminative materialist.

Boghossian's argument in fact hinges on what he takes to be the most powerful arguments in support of eliminative materialism. These are all, he thinks, arguments directed at the notion of content in the abstract, rather than at the content of propositional attitudes. He writes:

> The real difficulty with the suggestion that one may sustain differential attitudes towards mental and linguistic content stems from the fact that

> the *best* arguments for claim that nothing mental possesses content would count as *equally* good arguments for the claim that nothing linguistic does. For these arguments have nothing much to do with the items being *mental* and everything to do with their being *contentful*: they are considerations, of a wholly general character, against the existence of items individuated by content. If successful, then, they should tend to undermine the idea of linguistic content just as much as they threaten its mental counterpart. (Boghossian 1990: 171)

The arguments Boghossian discusses are arguments from (1) the indeterminacy of content, (2) the holistic nature of content, (3) the irreducibility of content, and (4) the "queerness" of content.

It is quite plausible, I think, that these four arguments are each just as applicable to linguistic content as they are to propositional attitude content. But there is no need to conclude with Boghossian that eliminative materialism entails global eliminativism. We can instead take it as a recommendation to look elsewhere for a plausible argument for eliminative materialism. The eliminative materialist needs an argument that can plausibly be confined to the psychological domain and that will not leave her open to Boghossian's charge of incoherence.

The eliminative materialist's first task must be to identify a class of putative errors committed by commonsense psychology that can plausibly be confined within the realm of the psychological. The putative error must be more localized than simply a commitment to content-bearing states in general. It must be something deriving either from the particular type of content-bearing states invoked in commonsense psychology or from the use to which they are put (or, of course, both). This paper explores both possibilities.

An eliminativist can argue that there is a fundamental error in interpreting the scope of commonsense psychology. The claim here is that we actually use the machinery of commonsense psychology far less frequently than we think we do. Whereas opponents of eliminativism hold that commonsense psychology is an indispensable tool for navigating the social world, the eliminativist can point to many occasions on which we manage to understand what is going on and engage in coordinated social behavior without bringing into play the conceptual framework of commonsense psychology.

Of course, this line of argument can only be part of the overall eliminativist strategy. Showing that commonsense psychology has a far more peripheral role to play in social understanding and social coordination than it is standardly held to have does not give us eliminativism. The eliminativist

has to go beyond this to tackle directly the core tenet of commonsense psychology. This core tenet is that propositional attitudes are the "springs of action." We act the way we do in virtue of our beliefs, desires, hopes, fears, and so on. These states are distinguished by having a particular type of content, namely, propositional content of the type that might be captured by a "that – " clause. It would be powerful support for the eliminativist thesis if it could be shown that invoking psychological states with propositional content is simply not the right way to think about the springs of action. The eliminativist can try to show that, although we need to appeal to representational states in explaining and predicting behavior, these representational states do not have propositional content and are fundamentally different from states that do have propositional content. The project here would be to establish a fundamental mismatch between two models of representation – the model of representation implicated in just about all ways of thinking about commonsense psychology, on the one hand, and the family of models of representation that seem to provide the best general picture of how the brain can be representational, on the other.

This paper is programmatic. Its aim is to explore how eliminativism might best be formulated and defended. Section 1 clarifies the nature of commonsense psychology, and hence by extension the eliminativist thesis. This allows us to see why Paul Churchland's standard arguments for eliminativism are unlikely to provide the desired conclusion. I argue that the most promising strategy for the eliminative materialist has two components. The first part of the strategy is to put pressure on what I have elsewhere called the broad construal of the scope of commonsense psychology (Bermúdez 2003, 2005). In most general terms "commonsense psychology" simply picks out the complex of skills and capacities that collectively make possible social understanding and social coordination. It is a substantial thesis that these skills and abilities all invoke the conceptual framework of commonsense psychology. The first step for the eliminative materialist is to block this move by arguing that commonsense psychology has a far more peripheral role to play in social understanding and social coordination than it is standardly held to have. This line of argument is discussed in Section 2.

But narrowing the scope of commonsense psychology does not give us eliminativism. The second part of the eliminativist strategy tries to establish a fundamental mismatch between two models of representation – the model of representation implicated in just about all ways of thinking about commonsense psychology, on the one hand, and the family of models of representation that seem to provide the best general picture of how the

brain can be representational, on the other. Sections 3 and 4 discuss this second component of the eliminativist strategy.

1. COMMONSENSE PSYCHOLOGY AND ELIMINATIVE MATERIALISM

In its most general sense the term "commonsense psychology" picks out the complex of social abilities and skills possessed by all normal, encultured, non-autistic and non-brain–damaged human beings. These are the skills and abilities that allow us to navigate the social world. Taken in this very general sense, commonsense psychology is an *explanandum* rather than an *explanans*. We would expect it to be the sort of thing of which a theoretical account is given, rather than something that can itself do theoretical and explanatory work.

The expression "commonsense psychology" is used more determinately to characterize what is in effect a particular conceptual framework deemed to govern our social understanding and social skills, where this conceptual framework can be thought of as an account of what underlies the general abilities and skills just identified. Here is a useful characterization of this second way of thinking about commonsense psychology from the introduction to a recent collection of essays:

> It has become a standard assumption in philosophy and psychology that normal adult human beings have a rich conceptual repertoire which they deploy to explain, predict and describe the actions of one another and, perhaps, members of closely related species also. As is usual, we shall speak of this rich, conceptual repertoire as 'folk psychology' and of its deployment as 'folk psychological practice'. The conceptual repertoire constituting folk psychology includes, predominantly, the concepts of belief and desire and their kin – intention, hope, fear, and the rest – the so-called propositional attitudes. (Davies & Stone 1995: 2)

In very general terms, then, our skills in social understanding and social coordination are underpinned by the conceptual framework of propositional attitude psychology. We can make sense of other people and coordinate our behavior with theirs in virtue of our ability to apply the concepts of belief, desire, and so forth.

This general characterization leaves unanswered questions about how the conceptual framework of propositional attitude psychology is applied in practice. This brings us to a third way of thinking about commonsense psychology. This is where we find the much-discussed distinction between

theory theorists and simulation theorists. A number of influential theorists accept the view that social understanding and social coordination rest upon an implicitly known, and essentially theory-like, body of generalizations connecting propositional attitude states with overt behavior and with each other. Paul Churchland is of course one of these, as are David Lewis, Frank Jackson, and Jerry Fodor. On this view (the so-called *theory theory*), social understanding involves subsuming observed behavior and what is known of a person's mental states under these generalizations in order to understand why they are behaving in a certain way and how they will behave in the future.

In recent years the theory theory has been challenged within both philosophy and psychology by theorists promoting the simulationist approach to commonsense psychology.[1] Simulationists hold that we explain and predict the behavior of other agents by projecting ourselves into the situation of the person whose behavior is to be explained/predicted and then using our own mind as a model of theirs. Suppose that we have a reasonable sense of the beliefs and desires that it would be appropriate to attribute to someone else in a particular situation, so that we understand both how they view the situation and what they want to achieve in it. And suppose that we want to find out how they will behave. Instead of using generalizations about how mental states typically feed into behavior to predict how that person will behave, the simulationist thinks that we use our own decision-making processes to run a simulation of what would happen if we ourselves had those beliefs and desires. We do this by running our decision-making processes *off-line*, so that instead of generating an action directly they generate a description of an action or an intention to act in a certain way. We then use this description to predict the behavior of the person in question.

We can, therefore, distinguish three different ways of construing commonsense psychology, as in the following table.

1	The complex of skills and abilities that underlie our capacities for social understanding and social coordination.
2	A particular conceptual framework for explaining social understanding and social coordination in which the propositional attitudes are central.
3	A particular account of how the conceptual framework in (2) is applied in the service of explanation/prediction.

It is plain that these different ways of construing commonsense psychology yield different ways of construing eliminative materialism. There is

no prospect of an eliminative materialism defined in terms of the first way of construing commonsense psychology, since this would involve denying the existence of successful social interaction and coordination. However, eliminative materialism looks very different depending on whether it is understood against the background of the second or third construals of commonsense psychology.

The most wide-ranging form of eliminativism is directed at the simple idea that we can only make sense of other people's behavior, and coordinate our own behavior with theirs, through the interpretive framework of propositional attitude psychology. This form of eliminativism takes issue with the very idea that our canonical way of getting purchase on the explanation, prediction, and coordination of behavior is through the machinery of the propositional attitudes. It is completely neutral on the issues that divide simulationists and theory theorists. Any arguments that can show that social coordination and social understanding do not rest upon the machinery of propositional attitude psychology will be effective against both theory theorists and simulationists.

One striking feature of Churchland's version of eliminative materialism is that, although the position is clearly targeted against commonsense psychology in its second construal (its intended target is the legitimacy of any use of content-bearing propositional attitudes in the enterprise of explaining, predicting, and coordinating behavior), the arguments standardly put forward make the most sense in the context of the third construal. The enterprise of showing that commonsense psychology does not count as a legitimate and productive scientific theory clearly presupposes the truth of some version of the theory theory. Of the three canonical arguments for eliminativism, two clearly fall into this category. Consider the following:

> The folk psychology of the Greeks is essentially the folk psychology we use today, and we are negligibly better at explaining human behavior in its terms than was Sophocles. This is a very long period of stagnation and infertility for any theory to display, especially when faced with such an enormous backlog of anomalies and mysteries in its own explanatory domain. (Churchland 1981: 124)

The idea that commonsense psychology is a theory is both explicit in the argument and essential to it. The same holds for the argument from the irreducibility of commonsense psychology:

> If we approach *Homo sapiens* from the perspective of natural history and the physical sciences, we can tell a coherent story of his constitution, development and behavioral capacities which encompasses particle physics,

> atomic and molecular theory, organic chemistry, evolutionary theory, biology, physiology and materialistic neuroscience . . . But folk psychology is no part of this growing synthesis. (idem.)

Only if commonsense psychology is a theory can it even be a candidate for reduction.

It might seem that an easy way to respond to the concerns just raised would be to argue on independent grounds for the inadequacy of the simulationist alternative to the theory theory. This in fact is what Paul Churchland has done. His 1989 paper "Folk psychology and the explanation of human behavior" contains a sustained argument against what he calls the 'anti-theoretical view of our self understanding.' Surely, if simulationism is not a serious contender, then arguments against the theoretical construal of commonsense psychology will be sufficient to establish the bankruptcy of commonsense psychology *tout court*.

This would be a mistake. There is a very significant difference between a successful *local* argument against a particular way of applying the conceptual framework of propositional attitude psychology and a successful *global* argument against propositional attitude psychology as such. What needs to be shown for the global conclusion to emerge is that no possible version of the theory theory could resolve the problem that the three local arguments identify. And nothing like this follows from the three arguments we have briefly considered. If they are telling at all (and of course this has been fiercely contested) they are telling only against our current commonsense psychological theories. To take just one example, if our commonsense psychology were to change then it would no longer be the same commonsense psychology enjoyed by the ancient Greeks, and the first argument would be blunted. But nothing has been said to rule out the possibility of a significantly developed commonsense psychology that continues to operate within the parameters of propositional attitude psychology. Nor is it clear how the sort of considerations that Churchland brings into play could possibly show this. What would be needed would be a principled argument to show that certain very general features of propositional attitudes (features general enough to apply to any possible version of propositional attitude psychology, and hence to any possible descendant of our current commonsense psychology) rule them out of the project of explaining and predicting human behavior. The type of arguments that Churchland brings into play in support of the standard version of eliminative materialism just seem to be the wrong type of arguments.

On the other hand, however, the resources for such an argument can be found elsewhere in Churchland's writings. There is an illuminating passage

in "Eliminative materialism and the propositional attitudes." Churchland is (implicitly) criticizing the idea that we can think of what a person believes in terms of the sentences to which that person would assent, so that we can use those sentences to characterize the *content* of what they believe. This is, of course, simply a special case of the general principle that propositional attitudes have contents that can be specified by means of "that –" clauses, where the complement of a "that –" clause is a declarative sentence.

> A declarative sentence to which a speaker would give confident assent is merely a one-dimensional *projection* – through the compound lense of Wernicke's and Broca's areas onto the idiosyncratic surface of the speaker's language – of a four- or five-dimensional "solid" that is an element in his true kinematical state . . . Being projections of that inner reality, such sentences do carry significant information regarding it and are thus fit to function as elements in a communication system. On the other hand, being *sub*-dimensional projections, they reflect but a narrow part of the reality projected. They are therefore *un*fit to represent the deeper reality in all its kinematically, dynamically, and even normatively relevant respects. (Churchland 1981: 129)

The motivation for eliminativism suggested here is not that it is an impoverished theory, nor that it cannot be reduced to neuroscience, nor that it is limited in its explanatory scope. Rather, we should be eliminativists because commonsense psychology rests upon an untenable model of representation.

Before going on to consider that line of argument in Sections 3 and 4, however, we should note that the distinction between different ways of thinking about commonsense psychology provides a way of clarifying the first part of the eliminativist strategy discussed earlier. It is effectively a matter of definition that we navigate the social world in virtue of commonsense psychology in the first of the three identified senses. But it by no means follows that our skills and abilities in social understanding and social coordination are underwritten by commonsense psychology in the second sense. The eliminativist can argue that many of our social skills and abilities have nothing to do with propositional attitude psychology. This line of argument is explored in the next section.

2. NARROWING THE SCOPE OF COMMONSENSE PSYCHOLOGY[2]

No eliminativist denies that we are capable of accommodating ourselves to other people's behavior and of engaging in coordinated social action. The eliminativist's question is how these forms of social accommodation

and coordination are achieved. To what extent, if any, are the propositional attitudes involved? One obvious way to broach this general question is to ask whether success in social exchanges and social interactions rests upon attributions of propositional attitudes. Do we navigate the social world by attributing beliefs, desires, hopes, and fears to the people whom we encounter in order to explain and predict their behavior? Or do we employ more basic mechanisms that allow us to interact successfully with others without using the machinery of propositional attitude psychology?

It will be helpful to break this question down into two further questions.

(1) Does successful social behavior always require explaining and/or predicting the behavior of other participants?

(2) In those cases where social behavior does depend on explaining and predicting the behavior of others, do such explanations and predictions have to involve propositional attitude psychology?

In this section I will briefly survey some reasons for answering these two questions in the negative – by giving examples of (a) social interactions that seem to proceed without any form of explanation or prediction, and (b) processes of explanation and prediction that seem to proceed without involving attributions of propositional attitudes.

Let me begin with the first question. Are there any social interactions that can be modeled without assuming that the parties involved are engaged in explaining or predicting each other's behavior? Surprisingly, game theory has some interesting implications in this area. Game theorists have long been interested in a particular class of strategic interaction where the dominant strategy for each player leads inevitably to an outcome in which each player is worse off than he could otherwise have been. A dominant strategy is one that is more advantageous than the other possible strategies, irrespective of what the other players do (it *dominates* those other strategies). The classic example of this kind of strategic interaction is the so-called prisoners' dilemma, which has the following pay-off table.

		Player B	
		Betray	Keep Silent
Player A	Betray	5, 5	0, 10
	Keep Silent	10, 0	2, 2

Each entry represents the outcome of a different combination of strategies on the part of prisoner A and B. The bottom left-hand entry represents the outcome if prisoner A keeps silent at the same time as being betrayed by prisoner B. The outcomes are given in terms of the number of years in prison that will ensue for prisoners A and B respectively. So, the outcome in the bottom left-hand box is 10 years for prisoner A and none for prisoner B. Imagine looking at the pay-off table from Prisoner A's point of view. You might reason as follows.

> Prisoner B can do one of two things – betray me or keep silent. Suppose he betrays me. Then I have a choice between five years in prison if I also betray him – or ten years if I keep silent. So, my best strategy if he betrays me is to betray him. But what if he keeps silent? Then I've got a choice between two years if I keep silent as well – or going free if I betray him. So, my best strategy if he keeps silent is to betray him. Whatever he does, therefore, I'm better off betraying him.

Unfortunately, prisoner B is no less rational than prisoner A and things look exactly the same from her point of view. In each case the *dominant* strategy is to defect. So, the two prisoners will end up betraying each other and spending five years each in prison, even though they both would have been better off keeping silent and spending two years each in prison.

The psychology of the prisoners' dilemma is not very complicated. There is no question of either player having to predict what the other player is doing. If betrayal is the dominant strategy then one should follow it whatever one thinks the other person is going to do, which just means that there is no need to think about what they are going to do. One can, in effect, read off what to do from the pay-off table.

Things get more complicated when we come to social interactions that have the same logic as the prisoner's dilemma but are iterated. When it is not known how many plays there will be and/or the rationality of the other participant is not known, scope opens up for cooperative play. This is where we rejoin the question of the domain of commonsense psychology. Suppose that we find ourselves, as we frequently do, in social situations that have the structure of an indefinitely repeated prisoner's dilemma. How do we negotiate the situation? One answer might be that I make a complex set of predictions about what the other player (or players) will do, based on my assessment of their preference orderings and their beliefs about the probability of each of us betraying as opposed to cooperating, and then factor in my own beliefs about how what will happen in the future

depends on whether or not I cooperate – and so on. This, of course, would be an application of the general explanatory framework of commonsense psychology (on the simplification that utilities and probability assignments are regimentations of desires and beliefs).[3]

But even if we can make sense of the idea that strategic interaction involves these kinds of complicated multi-layered predictions involving expectations about the expectations that other people are expected to have, one might wonder whether there is a simpler way of determining how to behave in that sort of situation. In fact game theorists have directed considerable attention to the idea that social interactions taking the form of indefinitely repeated prisoner's dilemmas might best be modeled through simple heuristic strategies in which, to put it crudely, one bases one's plays not on how one expects others to behave but rather on how they have behaved in the past. The best known of these heuristic strategies is TIT-FOR-TAT, which is composed of the following two rules:

1. Always cooperate in the first round
2. In any subsequent round do what your opponent did in the previous round

The TIT-FOR-TAT strategy is very simple to apply, and does not involve any complicated folk psychological attributions or explanations/predictions. All it requires is an understanding of the two basic options available to each player, and an ability to recognize which of those strategies has been applied by other players in a given case. The very simplicity of the strategy explains why theorists have found it such a potentially powerful tool in explaining such phenomena as the evolutionary emergence of altruistic behavior (see Axelrod 1984 for an accessible introduction and Maynard Smith 1982 and Skryms 1996 for more detailed discussion).[4]

Strategies such as TIT-FOR-TAT plainly do not involve any exploitation of the categories of folk psychology. In fact, they do not involve any processes of explanation or prediction at all. To apply TIT-FOR-TAT, or some descendant thereof, I need only decide whether the behavior of another player should best be characterized as a cooperation or a defection – and indeed determine which previous behaviors are relevant to the ongoing situation. This will often be achievable without going into the details of why that player behaved as they did. Of course, sometimes it will be necessary to explore issues of motivation before an action can be characterized as a defection or a cooperation – and sometimes it will be very important to do

this, given that identifying an action as a defection is no light matter. But much of the time one might well get by perfectly well without going at all deeply into why another agent behaved as they did.

The TIT-FOR-TAT heuristic is an interesting example of how one might navigate the social world without explaining or predicting the behavior of others. But suppose we are dealing with a social situation in which some form of explanation and/or prediction of the behavior of other participants is required. Is this a situation that we can only navigate by using the conceptual framework of commonsense psychology? Not necessarily. Consider everyday social interactions, such as buying food in a shop or ordering in a restaurant. These are coordination problems that can only be successfully negotiated because one has certain beliefs about why people are doing what they are doing and about how they will continue to behave. But there is no need for these beliefs to be second-order beliefs about the psychological states of other participants in the interaction. In such routine situations all that is required is to identify the person approaching the table as a waiter, or the person standing behind the counter as a butcher. Simply identifying social roles provides enough leverage on the situation to allow one to predict the behavior of other participants and to understand why they are behaving as they are. There is no need to think about what the waiter might desire or the butcher believe any more than they need to think about what I believe or desire. The social interaction takes care of itself once the social roles have been identified (and I've decided what I want to eat).

One lesson to be drawn from highly stereotypical social interactions such as these is that explanation and prediction *need not* require the attribution of propositional attitudes. Identifying someone as a waiter is a matter of understanding him as a person who typically behaves in certain ways within a network of social practices that typically unfold in certain ways. This is a case where our understanding of individuals and their behavior is parasitic on our understanding of the social practices in which their behavior takes place. Nor, of course, is this understanding of social practices a matter of mastery of a primitive theory. We learn through experience that certain social cues are correlated with certain behavior patterns on the part of others and certain expectations from those same individuals as to how we ourselves should behave. Sometimes we have these correlations pointed out to us explicitly; more often we pick them up by monitoring the reactions of others when we fail to conform properly to the "script" for the situation.

These programmatic remarks provide some support for eliminative materialism. Commonsense psychology, understood as a conceptual framework for making sense of behavior in terms of propositional attitude psychology, is based on the overarching assumption that people act in virtue of their beliefs and desires, and hence that we can only make sense of other people's behavior by interpreting it within the web of propositional attitude psychology. We have been considering examples of social interactions that (if my interpretation is correct) it would be a fundamental mistake to view in these terms. To the extent, then, that these social interactions are typical and widespread, rather than isolated outliers, we have a clear sense in which commonsense psychology could turn out to be false in a significant number of cases.

3. NEURAL REPRESENTATION AND ELIMINATIVE MATERIALISM

Let us turn now to the argument for eliminativism from neural representation. This is an argument that can be found in Churchland's writings, although it is not as foregrounded as the "standard" arguments briefly considered earlier. Cognition is, Churchland thinks, a form of information-processing and the representations over which that processing takes place are distributed in something like the way that representations are distributed over a large number of units and weights in an artificial neural network. Here is a very general presentation of the approach:

> The basic idea is that the brain represents the world by means of very high-dimensional *activation vectors*, that is, by a pattern of activation levels across a very large population of neurons. And the brain performs computations on those representations by effecting various complex *vector-to-vector transformations* from one neural population to another. This happens when an activation vector from one neural population is projected through a large matrix of synaptic connections to produce a new activation vector across a second population of nonlinear neurons. Mathematically, the process is a process of multiplying a vector by a matrix and pushing the result through a nonlinear filter. (Churchland 1992, reprinted in Churchland & Churchland 1998: 41)

Churchland is betting, in effect, that a complete account of neural computation will be defined over complex patterns of activation across large populations of neurons. This means that neural representations will have a

huge number of degrees of freedom. There will be as many dimensions of variation in a neural representation as there are neurons whose activation values can vary independently. In mathematical terms we need to consider neural representations as *n*-place vectors where *n* is the number of neurons.

Philosophical discussion of the ramifications of the distributed nature of neural representations have tended to focus on the issue of whether mental representations are structured. The parameters of the debate should be familiar. Proponents of the language of thought hypothesis argue that propositional attitudes cannot be causally efficacious unless the vehicles of those propositional attitudes are isomorphic to the structure of their contents (so that the vehicle of the belief that Paris is in France has distinguishable components corresponding to the separable and recombinable components of the content of the belief). It is noted that there seems little prospect of capturing such structure if representations are distributed in anything like the way they are in artificial neural networks. Then the conclusion is drawn *either* that artificial neural networks cannot be a good guide to the structure of mental representation *or* that propositional attitude psychology is seriously cast in doubt. There is a sense, of course, in which this can be taken as an argument for eliminativism from the (putatively) distributed nature of neural representations. But the argument is, to put it mildly, far from compelling. I would imagine that opponents of eliminativism have far more confidence in the validity of propositional attitude psychology than they have in the principle that causally efficacious propositional attitudes require isomorphically structured vehicles. It is hard to imagine this line of argument having much suasive force.

Yet there seems to be a much more direct line of argument from the distributed nature of neural representation to eliminativism. If neural representations are distributed in a way that yields as many degrees of freedom as there are neurons whose activation values can vary independently of each other, then we need to think of each individual neural representation as a location in a multidimensional state space where each dimension is given by the range of possible activation values for each unit. The state space of a neural representation defined over *n* independently varying neurons contains *n* dimensions, and assigning to each unit a particular activation value uniquely determines a single point within that *n*-dimensional space (a point that could equally be represented by a vector comprising an ordered sequence of those activation values). Paul Churchland is betting that this does indeed turn out to be the case. Here is the possibility he envisages.

> Suppose that research into the structure and activity of the brain, both fine-grained and global, finally does yield a new kinematics and correlative dynamics for what is now thought of as cognitive activity. The theory is uniform for all terrestrial brains, not just human brains, and it makes suitable conceptual contact with both evolutionary biology and non-equilibrium thermodynamics. It ascribes to us, at any given time, a set or configuration of complex states, which are specified within the theory as figurative 'solids' within a four- or five-dimensional phase-space. The laws of the theory govern the interaction, motion and transformation of these 'solid' states within that space, and also their relations to whatever sensory and motor transducers the system possesses. (Churchland 1981: 129)

How will this conceptual framework of solids within multidimensional phase space relate to the familiar framework of the propositional attitudes?

The argument for eliminativism from the distributed nature of neural representation is simply that there is a fundamental mismatch between the vocabulary of the propositional attitudes and the complex multidimensional representational states that our explanations of behavior are trying to capture. We act the way we do because of how our brains represent the world. Yet the complexity of those neural representations far outstrips the linguistic resources with which we are trying to capture them. As Churchland notes in the passage quoted in the first section, there seems to be a fundamental difference between one-dimensional linguistic characterizations of mental states and what he describes as the underlying multidimensional neural reality

It is built into our thinking about propositional attitudes that the content of a propositional attitude can be specified in a "that –" clause where what follows the "that" is a declarative sentence. The contents of propositional attitudes are exhausted by the meanings of the sentences that express them. If I believe that Paris is in France then there can be no more to the content of what I believe than can be expressed through the sentence "Paris is in France." We can see this general notion of propositional attitude content at play in the familiar debate about the relation between the content of belief and the content of perception. Theorists who argue that there is a fundamental distinction between the content of belief and the content of perception tend to highlight features of the content of perception that outstrip the concepts and words that we can deploy in specifying how a given perception represents the world. So, for example, the acuity of our color perception allows us to make discriminations that far outstrip our color vocabulary. The same holds for shape perception. We can perceptually represent the world as containing shapes for which we have no concepts or

words. This is frequently taken to indicate that the content of perception is fundamentally different from the content of belief – if the content of perception were content of the same type as in beliefs then it would be fully expressible without remainder in a "that –" clause.

We do not have anything like a plausible model of neural representation, but if Churchland's bet is well founded and neural representations are distributed in a way that gives them a very high number of degrees of freedom then it seems clear that there will be an even greater lack of fit between our propositional attitude vocabulary and the multidimensional representations in virtue of which we behave the way we do than there is between our propositional attitude vocabulary and our perceptual representations. Yet why should this provide an argument for eliminativism about the propositional attitudes? How do we get from the premise that propositional attitude vocabulary does not get things fully right to the conclusion that it gets things fundamentally wrong?

What is at issue here, of course, is the precise nature of the lack of fit between our propositional attitude talk and the underlying neural representations. Let us look again at the lack of fit between perceptions and epistemic reports of those perceptions (reports that specify what is sometimes called the conceptual content of perceptions). If I look out of the window, see my car in the drive and then say that I see that my car is in the drive, it is certainly true that I have failed fully to characterize my perceptual state. There are indefinitely many ways in which things could visually appear to me that it would be correct to describe as my seeing that my car is in the drive. But I have not, of course, said anything false. Not telling the whole story is not the same as telling a story that is wholly wrong – or even, for that matter, partly wrong. The eliminativist is arguing for something much stronger than the claim that commonsense psychology stands to the underlying neural reality in the relation that perceptual reports employing "that –" clauses stand to the underlying perceptual reality. Clearly, then, if we are to derive an argument for eliminativism from the multidimensional nature of neural representation there must be more going on than in the perceptual case.

4. ELIMINATIVE MATERIALISM AND SUBSYMBOLIC REPRESENTATION

How might such an argument be motivated? Suppose we identify a locus of representational content at a very low level in neural network models. That is to say, suppose that we identify particular units, or small groups of

units, as carrying out particular representational functions and tasks. This would be to make something like the "natural assumption" that Churchland sketches out in the following passage.

> If we are to assign specific semantic or representational *contents* to collective units of this kind, a natural first assumption is that any unit must in some way inherit its overall content from the individual and more basic representational significance of each of its many constituting elements, namely, the activation level of each of its many neurons. After all, it is these individual neurons that are the recipients of information from the environment: either directly, through their interaction with ambient light, heat and various mechanical and chemical impingements; or indirectly, through their many synaptic connections with neurons earlier in the processing hierarchy. (Churchland 1998, in Churchland & Churchland 1998: 83)

We can think of these units or groups of units as the representational primitives of the network – the place we need to start if we are to build up to an account of the representational character of the network as a whole.[5] It is natural to think that these representational primitives will be representing what are often called microfeatures of the environment. That is to say, they code features that are much more finely grained than those encoded within the vocabulary that we employ to specify the content of propositional attitudes. These microfeatures are, to use the familiar jargon, subsymbolic.

As far as eliminativism is concerned, the crucial question is the relation between this subsymbolic level of representation in terms of microfeatures and the symbolic level of representation in terms of objects and properties. The eliminativist needs to argue that there is a fundamental mismatch between the two different types of representation, so that what is going on at the subsymbolic level comes apart from what is going on at the symbolic level. Strikingly, discussions of the relation between subsymbolic and symbolic levels have tended to try to assimilate the two levels, rather than to drive a wedge between them. Theorists such as Smolensky, for example, have tried to show how approximations to our symbolic-level concepts can be constructed from complexes of microfeatures (see, for example, Smolensky 1988). Smolensky's well-known example of the distributed representation of *coffee cup* has some heuristic value, but is not very helpful because the microfeatures are all located squarely at the symbolic level. Smolensky's microfeatures include *hot liquid*, *burnt odor*,

finger-sized handle, and so on. Clearly there will be no difficulties integrating these microfeatures into a symbolic level account of what is going on in the network. For present purposes the following characterization from Fodor and Pylyshyn's well-known critique of neural network models is more helpful. Fodor and Pylyshyn note that the relation between subsymbolic and symbolic representations is standardly taken to be analogous to the relation between a defined expression and its feature analysis, and go on to say:

> Since micro-features are frequently assumed to be derived automatically (i.e., via learning procedures) from the statistical properties of samples of stimuli we can think of them as expressing the sorts of properties that are revealed by multivariate analysis of sets of stimuli (e.g. by multi-dimensional scaling of similarity judgments). In particular, they need not correspond to English words; they can be finer-grained than, or otherwise atypical of, the terms for which a non-specialist needs to have a word. (Fodor & Pylyshyn 1998 in Macdonald & Macdonald 1995)

Even though clusters of subsymbolic microfeatures approximate to the relevant symbolic representations, rather than mapping precisely onto them, Fodor and Pylyshyn are assuming a basic continuity between the subsymbolic and the symbolic level stories. And this is not at all surprising, given that (as they note) the subsymbolic characterization of what is going on in a network is often arrived at by working backward from the symbolic characterization. This might be by the feature analysis of concepts (as in Smolensky's coffee example) or by more sophisticated methods of analyzing the relevant activation space, such as the similarity measures proposed by Laakso and Cottrell (1998) or the dendogram analysis of NETtalk offered by Rosenberg and Sejnowski (1987).

If there is such a mapping, even if only an approximate mapping, from the subsymbolic to the symbolic level then there seems no prospect of arguing for an interesting version of eliminativism from the mismatch between the multidimensionality of neural representation and the unidimensionality of commonsense psychology. It is striking, therefore, that Churchland himself places considerable emphasis on the existence of such a mapping. In his 1998 paper "Conceptual similarity across sensory and neural diversity: The Fodor-Lepore challenge answered" he uses the Laakso-Cottrell similarity measure to argue that the semantic properties of a point in a network's activation space should be understood in terms of "stable and objective *macrofeatures* of the environment" (Churchland 1998: 85). If this is how we

are to understand the semantic properties of neural networks then there seems little prospect of arguing that the vocabulary and conceptual framework of propositional attitude psychology is fundamentally and in principle unsuitable for characterizing how brains represent the world. Eliminativism seems to be in tension with other aspects of Churchland's theoretical framework.

5. EVIDENCE FOR MICROFEATURAL REPRESENTATION

The eliminativist needs to show that we can explain behavior in terms of a level of neural representation that is incommensurable with the conceptual framework of propositional attitude psychology. This requires showing that the representations serving as the "springs of action" are representations of features of the environment that do not mesh with the features of the environment that are represented in propositional attitude psychology. But is there any evidence that this is true?

Let me begin with some more general comments. The explanatory power of folk psychology depends on beliefs, desires, and other propositional attitudes being the "springs of action." This way of thinking about the springs of action brings with it a particular interpretation of the architecture of cognition – specifically, a sharp distinction between "central" cognitive processes that involve propositional attitudes and "modular" cognitive processes that are not defined over propositional attitudes but instead provide inputs to the propositional attitude system. These modular processes have certain characteristics (such as informational encapsulation, domain-specificity, speed, and so on) that make it natural to classify them as subpersonal, in opposition to the personal-level propositional attitude system, which has none of these characteristics.

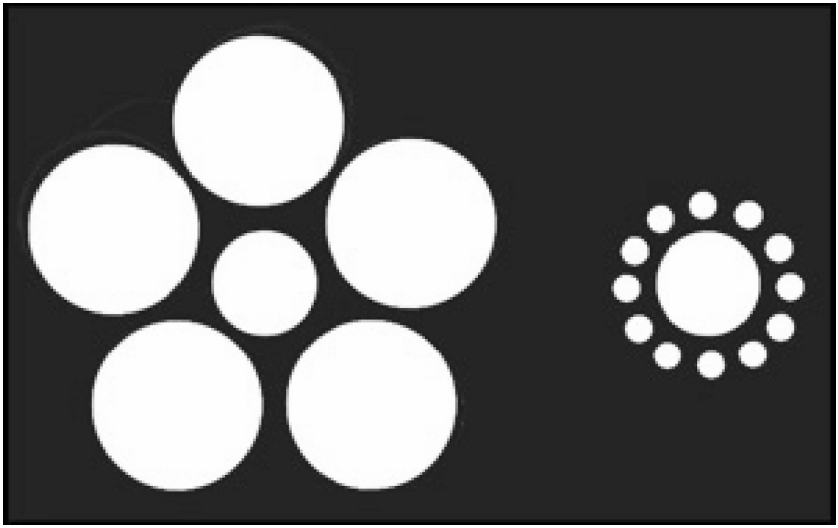
There is a number of ways of putting pressure on this way of thinking about the architecture of cognition. We looked at ways of making sense of the behavior of others that do not involve the attribution of propositional attitudes and hence that do not involve the explanatory framework of commonsense psychology. Much of our understanding of other people rests upon a range of relatively simple mechanisms and heuristics that allow us to identify patterns in other people's behavior and to respond appropriately to the patterns detected. The simplest such patterns are a function of mood and emotional state, while the more complex ones involve social roles and routine social interactions.

Of course, our ways of explaining behavior are not invariably a good guide to how that behavior came about, and so we need to look further. In the remainder of this section I point toward some experimental findings and research programs that might serve the eliminativist project by offering examples of this form of microfeature representation.

Two Visual Pathways

Psychologists and neuroscientists agree that there are (at least) two different ways of thinking about how our perceptions of the environment feed into action. There is some controversy about precisely how we are to understand both the function and the neuroanatomy of these two different pathways, but a considerable consensus that some sort of distinction needs to be made between "vision for action" and "vision for identification."[6] Neuropsychological dissociations are an important source of evidence. Researchers have reported a double dissociation between the capacity to act on objects and the capacity to name them. Patients with optic ataxia are able to identify objects but are severely impaired in tasks that involve reaching objects or working out their orientation, while patients with various types of agnosia have the reverse impairment – they can act on objects but are often completely unable to identify them. Neuroanatomical evidence points toward a distinction between two different visual pathways leading from the visual cortex – the dorsal pathway projecting to the posterior parietal cortex and the ventral pathway leading to the inferotemporal cortex. The functional distinction between the dorsal and ventral pathways was originally construed in terms of the distinction between "where" and "what," with the dorsal stream primarily involved in computing distance and location and the ventral stream specialized for the type of color and form processing that feeds into object identification (Mishkin and Ungerleider 1982). More recent investigation has suggested that the dorsal pathway is also involved in computing the "how" of action (Milner & Goodale 1995).

The two visual systems hypothesis offers an interesting example of how behavior can be explained in terms of the representation of microfeatures. One of the striking experimental data that has emerged from investigation of the differences between vision for action and vision for identification is that the two different systems can come into conflict. We see this, for example, in work that has been done on visual illusions, where the illusions have a much greater effect on perceptual reports than on action performance. The Ebbinghaus size contrast illusion is a case in point.



As the diagram indicates, a circle surrounded by other circles appears smaller if the surrounding circles are enlarged. When (normal) subjects are presented with two circles of the same size, one of which is surrounded by small circles and the other surrounded by large circles, they reliably judge the one surrounded by small circles to be larger than the one surrounded by large circles. Yet the illusion does not carry over to action. When subjects are asked to reach out as if they were going to pick up the circles their grip aperture is constant for the two circles. Similar effects have been observed with Muller-Lyer and Ponzo illusions. The dissociations between behavior and report in these visual illusions suggest that we respond to properties such as *graspability* that are a function of the size of the object and yet that are clearly distinct from the object's *perceived size*. These microfeatures are difficult to assimilate within the conceptual framework of commonsense psychology. It is a key tenet of commonsense psychology, for example, that we act on objects in virtue of how they appear to us, so that it is because an object looks a certain size to us that we make the appropriate hand movements for grasping it. Yet the experimental evidence suggests that things cannot be as simple as this. If subjects acted on how objects appear to them (more strictly: on how they report objects appearing to them) then they would act differently in the two cases. Instead it looks as if subjects are sensitive to properties of objects that are correlated with actual size but are independent of perceived size in a way that operates outside the realm of conscious awareness.

The Dimensional-Action System

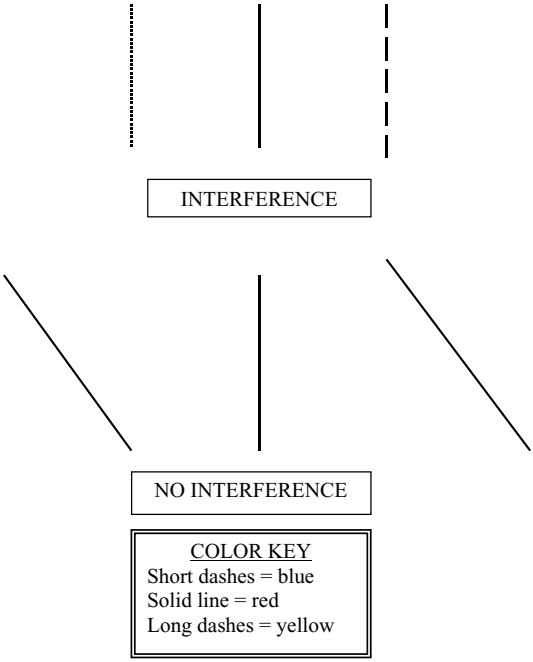
The eliminativist can place pressure on the hegemony of propositional attitude psychology by identifying links between perception and action that are based on the representation of microfeatures and that shortcut anything that might plausibly be described as "central processing." Intriguing recent results in perceptual psychology provide further examples of such microfeature-based perception-action links.

Traditional information processing accounts make a sharp distinction between perceptual processing and postperceptual processing and see all motor processing and response selection as falling clearly on the postperceptual side of the divide. Whereas perceptual processing is widely held to involve the separate processing of microfeatures (with shape, form, color, and so on all being processed in neurally distinct areas), postperceptual processing is thought to take place downstream of the "binding" of those microfeatures to form representations of objects. Yet some intriguing recent experimental evidence has led theorists to postulate highly specialized perception-action links that are explicitly tied to the perception of microfeatures (Cohen and Feintuch 2002). As with the dissociations discussed in the two visual systems case, the experimental evidence seems to show that we can act on isolated features of objects in complete independence of other features of those objects. It has been know for a long time that there are regions of the mammalian visual system perceptually sensitive simply to color or to shape, but it has always been thought that we can only act on the world by somehow combining these separately processed features into representations of objects – into representations that operate at the symbolic level of commonsense psychology, rather than at the subsymbolic level of microfeatures. Yet this assumption appears to be called into question by research into the dimensional action system.

One representative set of experiments was carried out with the so-called *flanker task interference paradigm*. Subjects are instructed to make differential responses to types of object presented at the center of a display while ignoring peripheral distractors flanking the target object. In the experiments reported in Cohen and Shoup (1997) the targets and responses are as follow. The first response is to be made to the appearance either of a red vertical line or of a blue right diagonal line, while the second is to be made to the appearance either of a green vertical line or of a blue left diagonal line. So, if the target object has a vertical orientation the appropriate response can only be made on the basis of color, while if it is blue the appropriate

response can only be made on the basis of orientation. The distractors are lines of varying colors and orientations.

Strikingly, interference effects are only observed when the relevant responses for target and distractor are on the same dimension.



So, for example, there is an interference effect if a red vertical line is flanked by differently colored vertical lines, but not if it is flanked by red diagonal lines. Similarly, there is interference if a blue left diagonal is flanked by other diagonal lines, but not if it is flanked by differently colored diagonal lines. The conclusion drawn by the experimenters is that there are distinct processing channels linking the detection of individual microfeatures (a particular orientation, or a particular color) with particular responses. These processing channels operate without any "binding" of the relevant micro-features.

In this case it is not the perception of the microfeatures that is surprising, or the fact that we are able to act on perceived color and perceived shape. What is surprising, and difficult to assimilate within the conceptual framework of propositional attitude psychology, is that there seem to be perception-action links triggered by representations of microfeatures that

are independent of representations of objects, or indeed of representations of other microfeatures.

The Ecological Approach to Perception and Action

J. J. Gibson's ecological approach to perception and action is best known for the claim that the line between perception and cognition is far less sharply defined than it is standardly taken to be. There are ways of perceiving the world that have direct implications for action. Frequently what we perceive are the possibilities that the environment "affords" for action, so that we can act on how we perceive the world to be, without having to form or exploit beliefs and other propositional attitudes. An affordance is a resource or support that the environment offers a particular creature (the possibility of shelter, for example, or the availability of food). These affordances are objective features of the environment that are supposed to be directly perceived in the patterns of light in the optic flow. Gibsonian psychologists describe organisms as "resonating to" or "directly sensitive to" the affordances in the distal environment.

This basic claim about affordances is not particularly radical, and in itself provides little or no support for the eliminativist project. If we describe affordances in more neutral terms as the instrumental properties of objects in the environment then it is hard to see why there should be any incommensurability between perceptual sensitivity to those instrumental properties and the conceptual framework of commonsense psychology.

Scope for incommensurability does appear, however, when we consider the general account of the workings of perception that underlies Gibsonian claims about the direct perception of affordances. One of Gibson's major contributions to the study of vision is the proposal to reconstrue the visual field as a constantly moving and constantly reconfiguring set of illuminated surfaces and concomitant solid visual angles, rather than in terms of empty space containing bounded objects (figures on a ground). We do not, he thinks, ever see empty space surrounding discrete objects. What we see is a complex and gapless structure of surfaces. Some of these surfaces are surfaces of objects, while others are not (the various surfaces in the sky, for example). To each surface there corresponds a solid visual angle with its base at the face of the visible surface and its apex at the point of observation. We can, for simplicity's sake, think of these solid angles as cones, although of course their shape will vary with the visible outline of the surface in question. As the observer moves through the environment the solid angles change, as one surface moves in front of another (relative to the perceiver)

or as the observer approaches or moves away from the surface. This is what Gibson terms optic flow.

The ecological analysis of visual perception gives us an example of how representation in terms of microfeatures might work in practice. There is a fundamental mismatch between a characterization of the distal environment in terms of objects and properties (of the sort that might feature in specifications of propositional attitudes) and a characterization of the distal environment in terms of optic flow. Gibson's perspective on perception rests upon the perception of microfeatures that resist assimilation to macrofeatures. These microfeatures include, for example:

- texture gradients (the decrease with distance of discriminable fineness of detail)
- the focus of expansion (the aiming point of locomotion that is also the vanishing point of optic flow)
- visual solid angles

Gibsonian accounts of perception attempt to show how behavior is controlled by perceptual sensitivity toward these microfeatures, which are properties of the global optic array, rather than of individual objects. Within its sphere of applicability, Gibsonian psychology certainly seems to provide an account of the "springs of action" that support eliminativism by offering explanations of behavior that resist assimilation to the concepts and categories of commonsense psychology.

Dynamic Touch and Rotational Inertia

The eliminativist needs to identify behaviors that can be shown to involve responding to microfeatures of the environment that resist assimilation within the conceptual framework of commonsense psychology. A good example of this has come from perceptual psychologists working within a broadly Gibsonian tradition and exploring the phenomenon of *dynamic touch*. It is well known that people can make accurate assessments of the spatial properties of objects by manipulating those objects. So, for example, people are remarkably accurate at detecting the length of objects by grasping those objects at a single point and moving the object around (without running their fingers over the whole object). One can get a feel for the phenomenon by picking up a pen, closing one's eyes and rotating the pen with one's finger, or taking a slightly larger object, such as a metal rod, and rotating it around one's wrist.

Dynamic touch is a puzzling phenomenon because the haptic system does not have access to any direct perceptual information about the length of the pen or the ruler. The physiological underpinnings of dynamic touch are mechanoreceptors yielding information about the stretching, contraction, and twisting of muscles and tendons. Very little work is done by receptors on the surface of the skin, even at the point at which contact is made with the object. Clearly there is some mechanical property (or properties) of objects that is reliably correlated with changes in the mechanoreceptors. The obvious candidates are weight and rotational force (torque), but neither of these can do the job. Perceived length is independent of both weight and torque, as can easily be appreciated by manipulating a pen and a pencil of the same length but different weights, and by manipulating both of them with very different twisting forces. The key mechanical property must remain invariant through changes in torque and weight.

It turns out (Turvey 1996, Carello and Turvey 2004) that the relevant physical invariant is what is known as the *inertia ellipsoid*. The inertia ellipsoid is, roughly speaking, a way of characterizing an object that measures the object's resistance to being rotated. It is derived from the object's principal moments of inertia, where a principal moment of inertia quantifies an object's resistance to rotation around one of its axes of symmetry. An object's moment of inertia will vary according to the distribution of its mass, with higher concentrations of mass away from the object's center of gravity yielding a higher moment of inertia. Suppose, for example, that we hang a weight from a metal rod. The further away the attachment point is from the rod's center of gravity the greater the moment of inertia – and the more force will be required to rotate it. Once we know an object's axes of symmetry and its principal moments of inertia we can characterize its overall resistance to being rotated in terms of an ellipse whose center is the intersection of the three axes of symmetry and whose surface is obtained from the reciprocal of the square roots of the principal moments of inertia.

It is a robust finding that an object's rotational inertia, as given by the inertia ellipsoid, is the invariant underlying perceived length. Nor is length the only quantity that can be detected by perceived touch. People can make reliable estimates of an object's weight from wielding the object. Amazeen and Turvey (1996) have established that perceived heaviness is also a function of rotational inertia, both when perceived heaviness accurately tracks an object's weight and when (as in the size-weight illusion) it leads to misleading estimates.[7]

This sensitivity to rotational inertia is a further example of the type of microfeatural sensitivity suggested by the Gibsonian approach to

perception and action. We act on the world in virtue of our perceptual attunement to properties of objects and of the optic array that are fundamentally alien to the conceptual framework of commonsense psychology. The inertial ellipsoid is a mathematical object that stands to our commonsense thinking about objects and their dynamic and kinematic properties in something like the relation that the rules of transformational grammar stand to our everyday use of English. Subtle experimental work is required to identify rotational inertia as the relevant parameter in our haptic sensitivity to the spatial properties of objects.

The Influence of Situation in Social Psychology

The "springs of action" have been investigated by social psychologists as well as cognitive psychologists and neuroscientists. The research has been two-pronged, investigating both why people behave the way they do and how we interpret that behavior. Two features of this research are particularly salient in the present context. The first has to do with the genesis of behavior. There is an overwhelming body of evidence highlighting the importance of the situation in determining behavior. Situational changes that might seem at first sight to be insignificant have been shown to have a serious effect on behavior. The second feature is that subjects systematically underestimate the significance of the situation, making what has come to be known as the "fundamental attribution error" of overestimating the significance of character traits and personality in explaining and predicting behavior.

In one famous set of experiments (Darley and Batson 1973) groups of students at a Princeton theological seminary were sent from one building to another as part of an experiment putatively on religious education. Their task in the second building was to give a talk, with one group giving a talk on the Good Samaritan and another on jobs in seminaries. On the way over they passed an experimenter slumped in a doorway and moaning. Overall 40% of the subjects stopped to offer some sort of assistance. What is striking, though, is the drastic difference between subjects who were told that they were running late (only 10% offered assistance) and subjects who were told that they had time to spare (where the figure was 63%). This discrepancy did not seem to be correlated with any other differences between the participants.

Other experiments have found what seem *prima facie* to be even more trivial situational factors having a large impact on behavior. Mathews and Canon (1975) explored the influence of ambient noise, showing that

subjects are five times less likely to help an apparently injured man who has dropped some books when there is a power mower running nearby than when ambient noise is at normal levels. Isen and Levin (1972) found an even more striking effect, discovering that people who had just found a dime were 22 times more likely to help a woman who has dropped some papers than people who had not found a dime. Experiments such as these have been carried out many times and the powerful influence of situational factors has proved a very robust finding.

This experimental tradition poses numerous interesting philosophical problems, particularly with respect to the role that character plays in ethical theory (Doris 2002). For present purposes what is interesting is the perspective that situationist social psychology casts on the genesis of behavior. It looks very much as if features of situations that do not in any sense count as commonsense psychological reasons for action can play a large role in determining how people behave. At least in the experimental paradigms it seems fundamentally inappropriate to seek explanations in terms of propositional attitudes that act as reasons for action. An important element in the springs of action seems to be relatively low-level features of the situation – what might, in fact, be termed situational microfeatures.

6. PROSPECTS FOR ELIMINATIVISM?

The constraints on a successful argument for eliminativism should by now be clear. Eliminative materialism is an error theory and hence the most plausible way of arguing for it is to identify precisely the error that commonsense psychology is supposed to be committing. The error must come in the very idea that the springs of action can be explained through the concepts and categories of propositional attitude psychology, rather than in a particular way of thinking about how those concepts and categories are to be applied or in a very general feature such as a commitment to the role of representational content in psychological explanation. The eliminativist's aims are not best served by focusing, as Paul Churchland sometimes does, on the idea that commonsense psychology is a proto-scientific theory to be judged by the standards appropriate to scientific theories. Nor is the eliminativist well served by mounting a global attack on the very idea of content-bearing states. The eliminativist needs an argument that will engage those who take a fundamentally different approach to what commonsense psychology is and how it is to be applied. And it is important that the eliminativist thesis be formulated in a way that avoids the charges of

incoherence that might plausibly be leveled against global attacks on the status of content.

The most plausible way of meeting these various constraints, I have argued, is for the eliminativist to identify two very different types of error in commonsense psychology and how we apply it. The first putative error is an error about the scope of commonsense psychology. The eliminativist can argue that we rely far less on commonsense psychology than is generally assumed by pointing to far more primitive mechanisms that control social coordination and governing our understanding of ourselves and others. The case here seems plausible. It seems very likely that the significance of commonsense psychology is significantly overstated by philosophers of mind.

But the real motivation for eliminativism has to come from a more direct attack on the role that propositional attitudes are supposed to play in the genesis of behavior. The eliminativist needs to argue that the representations that feed into action are fundamentally different from those invoked by propositional attitude psychology. The "springs of action" are representations of features that are much more finely grained than those encoded within the vocabulary that we employ to specify the content of propositional attitudes. The eliminativist's most promising strategy is to argue that, whatever we might think about why we behave the way we do, careful experimental work will show that we are in fact acting in virtue of representations of properties and microfeatures that fall completely outside the ambit of propositional attitude psychology. We have considered a number of examples of how this eliminativist strategy might be developed, ranging from the implications of the two visual systems hypothesis to research in social psychology into the role that situational factors play in controlling action.

Of course, even if the examples I have proposed on the eliminativist's behalf are accepted, the eliminativist case is a long way from being made. With the exception of the experiments on situational factors in social psychology, the examples given are all of relatively low-level motor tasks. The real question is how well the examples scale up. Can the eliminativist give comparable accounts of what is going on in the far more complex cases that tend to be cited when the thesis of the ineliminability of commonsense psychology is proposed? It is very difficult to say. Simpler tasks have been much more comprehensively studied than complex ones. No doubt many will think that beliefs, desires, and other propositional attitudes have an irreducible role to play in explaining complex actions and social coordination. On the other hand, the case for ineliminability needs to be made. One

clear benefit of exploring the dialectic of eliminativism in the way we have been doing is that it makes clear the level at which the debate about the status of commonsense psychology will have to be made. It is simply not good enough to proclaim the ineliminability of the conceptual framework of propositional attitude psychology, or to make intuitive appeals to the possibility of formulating generalizations that capture patterns in behavior that cannot be captured without bringing in commonsense psychological concepts. The debate about the status and future of commonsense psychology must take place on the basis of a detailed study of the genesis and execution of action. One thing that we learn from looking at simple motor behavior is that our intuitions about what is going on when we make simple discriminations or wield objects are systematically misleading. Could the same be going on in more complex behaviors? If, so then it may well turn out, as the eliminativist claims, that the scope of commonsense psychology will shrink to an extensionless point, as the representations that drive behavior are discovered to have little in common with the belief-desire pairs that are the mainstay of propositional attitude psychology. That would be an outcome very much in the spirit of Paul Churchland's eliminative materialism, even if the means by which it is reached are rather different from Churchland's "official" arguments for eliminativism.

Notes

* I am grateful to Brian Keeley for very helpful comments on an earlier draft of this paper, and to audiences at the University of Cincinatti and at the 2004 meeting of the Austrian Ludwig Wittgenstein Society at Kirchberg-am-Wechsel. Some of the material in Section 2 is taken from Chapter 7 of Bermúdez, 2005.

1. Gordon (1986) and Heal (1986) are key statements of the simulationist position. The principal readings in the debate between theory theorists and simulation theorists are collected in Davies and Stone 1995a and 1995b. Further essays will be found in Carruthers and Smith 1996. This collection includes interesting material from developmental psychologists and students of primate cognition. Currie and Ravenscroft 2002 develops a theory of imagination in the context of a simulationist approach to social understanding.
2. This section incorporates material from Bermúdez 2003 and forthcoming.
3. For more on this simplifying assumption see Lewis 1994 and Pettit 1991.
4. TIT-FOR-TAT has only a limited applicability to practical decision-making. In a situation in which two players are each playing TIT-FOR-TAT, a single defection will rule out the possibility of any further cooperation. This is clearly undesirable, particularly given the possibility in any moderately complicated social interaction that what appears to be a defection is not really a defection (suppose, for example, that my colleague misses the examination meeting because her car broke down).

So any plausible version of the TIT-FOR-TAT strategy will have to build in some mechanisms for following apparent defections with cooperation, in order both to identify where external factors have influenced the situation and to allow players the possibility of building bridges back towards cooperation even after genuine defection. One possibility would be TIT-FOR-TWO-TATS, which effectively instructs one to cooperate except in the face of two consecutive defections.

5. In fact, for reasons brought out in Churchland and Sejnowski (1993), the representational primitives are far more likely to be distributed across groups of units than to individual units. See their discussion of local coding versus vector coding on pp. 163ff.
6. See the essays in section I of Prinz and Hommel (2002) for up-to-date surveys of current thinking in this area. The tutorial by Rossetti and Pisella is particularly helpful.
7. The size-weight illusion is the illusion that larger objects of the same weight are perceived as heavier.

References

Amazeen, E. L., & Turvey, M. T. (1996). "Weight perception and the haptic size-weight illusion are functions of the inertia tensor." *Journal of Experimental Psychology: Human Perception and Performance*, **22**: 213–32.

Axelrod, R. (1984). *The Evolution of Cooperation*. Harmondsworth, Penguin.

Bermúdez, J. L. (2005). "The domain of folk psychology." In A. O'Hear (Ed.), *Mind and Persons*. Cambridge, Cambridge University Press.

Bermúdez, J. L. (2005). *Philosophy of Psychology: A Contemporary Introduction*. London, Routledge.

Boghossian, P. (1990). "The status of content." *The Philosophical Review* **99**: 157–84.

Carello, C., & Turvey, M. T. (2004). "Physics and psychology of the muscle sense." *Current Directions in Psychological Science* **13**: 25–8.

Churchland, P. M. (1981). "Eliminative materialism and the propositional attitudes." *Journal of Philosophy* **78**: 67–90.

Churchland, P. M. (1989). "Folk psychology and the explanation of human behavior." *Philosophical Perspectives* **3**: 225–41.

Churchland, P. M. (1992). "Activation vectors vs. propositional attitudes: How the brain represents reality." *Philosophy and Phenomenological Research* **52**: 419–524.

Churchland, P. M. (1998). "Conceptual similarity across sensory and neural diversity: The Fodor/Lepore challenge answered." *Journal of Philosophy* **65**: 5–32.

Churchland, P. M., & Churchland P. S. (1998). *On the Contrary: Critical Essays, 1987–1997*. Cambridge, MA, MIT Press.

Churchland, P. S., & Sejnowski, T. J. (1992). *The Computational Brain*. Cambridge, MA, MIT Press/Bradford Books.

Cohen, A., & Feintuch, U. (2002). "The dimensional-action system: A distinct visual system." In W. Prinz and B. Hommel (Eds.), *Common Mechanisms in*

Perception and Action: Attention and Performance XIX. Oxford, Oxford University Press.

Cohen, A., & Shoup, R. (1997). "Perceptual dimensional constraints on response selection processes." *Cognitive Psychology* **32**: 128–81.

Darley, J. M., & Batson, D. (1973). "From Jerusalem to Jericho: A study of situational and dispositional variables in helping behavior." *Journal of Personality and Social Psychology* **27**: 100–8.

Davies, M. K., Stone, T. (eds.) (1995). *Folk Psychology: The Theory of Mind Debate*, Oxford, Blackwell.

Doris, J. (2002). *Lack of Character*. New York, Cambridge University Press.

Fodor, J., & Pylyshyn, Z. (1988). "Connectionism and cognitive architecture: A critical analysis." *Cognition* **28**: 3–71.

Isen, A. M., & Levin, P. A. (1972). "Effect of feeling good on helping: Cookies and kindness." *Journal of Personality and Social Psychology* **21**: 384–8.

Laakso, Aarre, & Cottrell, Garrison W. (1998). How can I know what You think?: Assessing representational similarity in neural systems. *Proceedings of the Twentieth Annual Cognitive Science Conference, Madison, WI.* Mahwah, NJ, Lawrence Erlbaum.

Lewis, D. (1994). "Reduction of mind." In S. Guttenplan (Ed.), *Companion to the Philosophy of Mind*. Oxford, Blackwell.

Matthews, K. E., & Canon, L. K. (1975). Environmental noise level as a determinant of helping behavior. *Journal of Personality and Social Psychology* **32**: 571–7.

Maynard Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge, Cambridge University Press.

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

P. Pettit (1991). "Decision theory as folk psychology." In M. Bacharach & S. Hurley (eds.), *Foundations of Decision Theory*. Oxford, Blackwell.

Sejnowski, T. J., & Rosenberg, C. (1987). "Parallel networks that learn to pronounce English text." *Complex Systems* **1**:145–68.

Skyrms, B. (1996). *The Evolution of the Social Contract*, Cambridge, Cambridge University Press.

Smolensky, P. (1988). "On the proper treatment of connectionism." *Behavioral and Brain Sciences* **11**(1): 1–23.

Turvey, M. T. (1996). Dynamic touch. *American Psychologist* **51**: 1134–51.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale & R. J. W. Mansfield (Eds). *Analysis of Visual Behavior*. Cambridge, MA, MIT Press.

3 The Introspectibility of Brain States as Such

PETE MANDIK

Paul Churchland has defended various bold theses throughout his career. Of particular interest to the current chapter is what I shall call Churchland's *Introspection Thesis*.

> A person with sufficient neuroscientific education can introspect his or her brain states *as* brain states.[1]

Is the Introspection Thesis true? It certainly isn't obvious. Introspection is the faculty by which each of us has access to his or her own mental states. Even if we were to suppose that mental states are identical to brain states, it doesn't follow immediately from this supposition that we can introspect our mental states *as* brain states. This point is analogous to the following. It doesn't follow immediately from the mere fact that some distant object is identical to a horse that we can perceive it *as* a horse. Further, it isn't obvious that any amount of education would suffice to make some distant speck on the horizon *seem* like a horse. It may very well be the case that no matter how well we *know* that some distant speck is a horse; as long as we are sufficiently distant from it we will only be able to *see it as* a speck. Analogously then, it may very well be the case that no matter how well we *know* that our mental states are brain states, we will only be able to *introspect them as* irreducibly mental.

Not only is the introspection thesis not obviously true, it is not obvious what it would be like for it to be true. We can easily imagine seeing a horse as a horse. Can we similarly imagine introspecting brain states as brain states? I think, indeed, we can. Though I think the case will be even clearer once we review Churchland's arguments, it will be useful, at this early stage, to get a sketch of what it is we are supposed to imagine. It helps to begin by noting the distinction between something that can be perceived and something that can be figured out based on what is perceived. If I stick my finger into a hot cup of coffee, I can perceive the heat of the coffee. If, without touching the coffee, I see the steam rising from it, it is the steam that I see and based on what I perceive, I figure out the approximate temperature

of the coffee. As will be elaborated below, for Churchland the crucial distinction here depends not on the degree to which theoretical knowledge is involved but on how *automatic* its application in perception is. If, on having a certain sensation, I come to *automatically* apply the concept of heat to the cause of the sensation, then, for Churchland, that counts as perceiving heat. If I learn, then, to apply the concept of heat automatically (that is, without going through some intermediary inference) to coffee on the sight of steam rising from it, then what I've learned to do is see the heat of the coffee. I may perceive the coffee as being hot even in situations in which I see it without feeling it. Churchland's defense of the Introspection Thesis depends on an analogous view of introspection. Suppose that, in addition to being able to apply the concept of heat to an external object as an automatic reaction to some sensation, I learn the concept that applies to the neural basis of that sensation. Or, to pick a different kind of sensation, suppose the neural basis of motion perception involves activity in area V5 of cerebral cortex and that I learn to apply the concept of activity in V5 as an automatic response to a sensation of motion. Under such conditions, then, I would be introspecting my brain states *as* brainstates and in this case I would be introspecting the sensation of motion as a pattern of activity in area V5 of cerebral cortex.

Why care whether the Introspection Thesis is true? Churchland cares about the Introspection Thesis because it provides him a defense of his favored brand of materialism against attacks by anti-materialists who would base their claims on introspection. Churchland is concerned to show, then, by showing how introspection itself can reveal that mental states are brain states, that introspection does not provide an unassailable refuge for the antimaterialist. Further interest in the Introspection Thesis is that it has deep implications for current work on consciousness and it impacts debates not only between materialists and antimaterialists but also debates among materialists. In particular, it arguably undermines *representationalism*, an approach to consciousness that has many materialist adherents.[2] Although this will be unpacked further later, to a first approximation representationalism is the view that there is nothing more to qualia – the phenomenal characteristics of conscious experience – than representational content. The tension between representationalism and the Introspection Thesis becomes apparent once we consider a thesis oft associated with representationalism: the *Transparency Thesis*.

> When a person introspects his or her own conscious mental states he or she only has access to the properties those states represent objects as having.

To spell out the Transparency Thesis in terms of an example, it is the claim that when I have a conscious experience of a blue square, my introspective access to my experience only puts me in touch with features represented as instantiated in the environment – blueness and squareness. I have no introspective access, then, to features of the experience itself. In direct opposition to Churchland's Introspection Thesis, then, I can have no introspective access to any of the stuff going on in my brain when I have a conscious experience of a blue square on the wall. The metaphor of transparency is appropriate here insofar as when I examine my experiences I inevitably "look through" them to an external world of objects and properties that the experiences represent. The Transparency Thesis is oft appealed to as a premise in arguments for representationalism.[3] Further, representationalism is arguably true only if the Transparency Thesis is true. As Kind (2003) puts the point, if we are able to introspect aspects of experiences other than their representational contents, then properties other than representational contents of experience figure into the phenomenal character of experience.[4]

My goal in this chapter is to adjudicate between the competing theses of Transparency and Introspection, arguing ultimately that Transparency is the weaker of the two. I discuss further what the prospects are for representationalism without the Transparency Thesis. The organization of the rest of this chapter is as follows. First I spell out Churchland's arguments for the Introspection Thesis. Next I spell out the Transparency Thesis and the related notions of qualia, consciousness, and representationalism. Finally I discuss the degree to which the tension between Transparency and Introspection is a problem for Churchland and what resources his larger body of work makes available to resolve it.

CHURCHLANDISH INTROSPECTION

Churchland's argument for the Introspection Thesis depends on a particular view of perception and an analogy between perception and introspection. The view of perception at play here is that "perception consists in the conceptual exploitation of the natural information contained in our sensations or sensory states." (Churchland 1979:7). Analogously then, introspection is the conceptual exploitation of natural information that our sensations or sensory states contain about themselves. Fleshing out Churchland's views of perception and introspection requires us to flesh out what Churchland thinks the conceptual exploitation of natural information is. Crucial here

is a distinction Churchland draws between two kinds intentionality that sensations can have, that is, two ways in which a sensation can be a sensation *of* φ. A sensation can have "objective intentionality" as well as "subjective intentionality" and Churchland adopts the typographical convention of subscripts to distinguish "sensation of$_o$ φ" from "sensation of$_s$ φ". Spelling out the distinction semiformally, Churchland provides:

> *Objective* intentionality:
> A given (kind of) sensation is a sensation of$_o$ φ with respect to a being x if and only if
>
> under normal conditions, sensations of that kind occur in x only if something in x's perceptual environment is indeed φ.
>
> *Subjective* intentionality:
> A given (kind of) sensation is a sensation of$_s$ φ with respect to a being x if and only if
>
> under normal conditions, x's characteristic non-inferential response to any sensation of that kind is some judgment to the effect that something or other is φ. (ibid, p. 14)

The objective intentionality of sensations is the information that sensations actually carry about the environment regardless of whether we exploit that information. The objective intentionality of sensations determines what it is that we are *capable* of perceiving. What we actually *do* perceive depends on subjective intentionality. That is, what we actually do perceive depends on what concepts we bring to bear in the judgments that our sensations noninferentially elicit. So, for example, whether I am capable of seeing the tiny insect on the far side of the room depends on whether I have states of my visual system that reliably co-vary with the presence of that object, and if my eyesight is insufficiently acute, I will lack such states. Whether I actually do perceive that object depends on more than just good eyesight. It depends on whether I actually do employ my conceptual resources to interpret my visual sensations as indicating the presence of an insect. Thus, enriching our conceptual repertoire allows us to better exploit, in perception, the information already contained in sensation (ibid: 16). For example, with sufficient education, we can move beyond the coarse-grained common-sense temperature concepts in virtue of which we feel things as 'hot,' 'warm,' and 'cold' and instead exploit scientific concepts in order to feel "that the mean kinetic energy of the atmospheric molecules in this room is roughly 6.2×10^{-21} kg m^2/s^2" (ibid: 26). Multiplying examples, Churchland offers

that with sufficient conceptual augmentation we can hear "the occurrence and properties of compression wave trains in the atmosphere – most obviously of both their wavelength (from 15 m to 15 mm) and their frequency (from 20 to 20,000 cycles per second)" (ibid: 26) and we can see "the dominant wavelength (and/or frequency) of incoming electromagnetic radiation in the range $0.38–0.72 \times 10^{-16}$ m, and of the reflective, absorptive, and radiative properties of the molecular aggregates from which it comes" (ibid: 27). Our sensory states already carry this information and it is thus there waiting to be picked up by a suitably theoretically informed set of concepts.

Human perceivers are importantly analogous to measuring instruments, according to Churchland. Both have states that serve as reliable indicators of certain aspects of the environment. Further, in both cases reliable indication relies on interpretation functions that map distinct states onto distinct propositions (ibid: 38). In the case of measuring instruments, the interpretation function is determined when we calibrate the measuring instrument to map, for instance, the needle positions on an ammeter "onto distinct propositions such as 'there is a 5 ampere (A) current flowing in the circuit'" (ibid: 38). In the case of the conceptual exploitation of sensory information, while Churchland acknowledges that we do not explicitly and consciously use an interpretation function to formulate our perceptual judgments, he nonetheless points out that,

> *insofar as our conceptual responses to our sensations do display determinate and identifiable patterns* . . . we *embody* or *model* a set of interpretation functions . . . implanted in childhood as we learned to think and talk about the world . . . [and that] are just as properly subjects for evaluation, criticism, and possible replacement as are interpretation functions in any other context." (ibid: 39; emphasis in original)

With the above view of perception in hand, Churchland goes on to spell out what introspection would amount to. Focusing on introspective judgments about sensory states "e.g. 'I have a visual sensation of an orange circle'" (ibid: 40), Churchland describes introspection as involving "a temporary disengagement from the interpretation functions that normally govern our conceptual responses, and the engagement instead of an interpretation function that maps (what we now conceive as) sensations, etc., onto judgments *about* sensations, etc." (ibid: 40; emphasis in original). One consequence of this view of introspection, important both to Churchland and for points I'll raise subsequently, is that introspective judgments are no more likely to be incorrigible or infallible than perceptual judgments more

generally. Churchland illustrates by continuing the analogy to measuring instruments:

> [C]onsider an ammeter with a graduated dial marked '5 A', '10 A', and so on. Suppose it [is] constructed so that at the flick of a switch it flips another dial into place behind the needle, a dial marked '0.01 gauss', '0.02 gauss', and so on. This second dial is so calibrated that the needle positions on the dial now *overtly* reflect the simultaneous strength of the variable magnetic field inside the instrument, the very field whose action moves the spring-loaded needle. Our ammeter is now operating in "introspective mode". (ibid: 40)

A measuring instrument not only has states that carry information about its immediate environment, its states carry information about themselves and a calibration of the instrument can just as easily latch on to the one kind of information as the other. To use an example perhaps more accessible than those Churchland provides, the height of the column of mercury in a thermometer not only carries information about the temperature of the surrounding medium, but also information about how high the mercury is. We could put a mercury thermometer in "introspective mode," then, by changing the marks on it from measurements of degrees in Celsius to measurements of height in millimeters. And again, there is no guarantee of accuracy, for the calibration scheme may very well say that the current height is 3 mm when in reality it is 3.5 mm. However, when the device is correctly calibrated, what indicates that the height is 3 mm is when the 3 mm mark is even with the top of the mercury column that is, in fact, 3 mm in height.

The Churchlandish introspection of brain states involves exploiting the information that a state of the nervous system carries about itself. Churchland offers possible examples of what this neurophysiologically informed introspection would be like. His remarks on these possibilities are worth quoting at length for they simultaneously serve to bolster the plausibility of the Introspection Thesis and cast doubt on the Transparency Thesis.

> The considerable variety of states currently apprehended in a lump under 'pain', for example, can be more discriminately recognized as sundry modes of stimulation in our A-delta fibres and/or C-fibres (peripherally), or in our thalamus and/or reticular formation (centrally). What are commonly grasped as "after images" can be more penetratingly grasped as differentially fatigued areas in the retina's photochemical grid, and the chemical behaviour of such areas over time – specifically, their resynthesis of rhodospin (black/white) and the iodopsins (sundry colours) – is readily followed

> by suitably informed introspection. The familiar "phosphenes" can be recognized as spontaneous electrical activity in the visual nervous system. Sensations of acceleration, and of falling, are better grasped as deformations and relaxations of one's vestibular maculae, the tiny jello-like linear accelerometers in the vestibular system. Rotational "dizziness" is more perspicuously introspected as a residual circulation of the inertial fluid in the semicircular canals of the inner ear, and the increase and decrease of that relative motion is readily monitored. The familiar "pins and needles" at a given site is more usefully apprehended as oxygen deprivation of the nerve endings there located. (ibid: 118–119)

Before moving on to consider how Churchland's Introspection Thesis bears on discussions of the alleged transparency of conscious experience, it will be useful to summarize the key points from the above discussion and relate them to examples that are perhaps a bit easier to get an intuitive grasp of than Churchland's examples of, say, perceiving red light as electromagnetic radiation in the range $0.38–0.72 \times 10^{-16}$ m. The crucial aspects of Churchland's account of perception are those that allow for the reconstruction of the distinction between what is perceived without inference and what is inferred but not perceived. Let us consider the following situation to illustrate this distinction. Two friends, George and John, are lunching in a well-lighted location when, as part of some publicity stunt, a man in a realistic gorilla suit runs through the area. Suppose that both gorilla suit and gorilla act are quite realistic and convincing to the untrained eye. George, being a special effects expert for the film industry, is not fooled and can see quite clearly that this is indeed a man in a costume. John, however, is a novice and cannot help but be fooled: he sees this as a genuine gorilla, perhaps escaped from the nearby zoo. In fact, John the novice continues to see this individual as a genuine gorilla even after George the expert assures him that it is in fact a suited man. John may even come to believe George's testimony for he trusts George's expertise, but John cannot shake the impression that it is a real gorilla that is causing a ruckus in the restaurant. There are several key similarities and differences between John and George and Churchland's account of perception helps to explain these similarities and differences. The first similarity is that there is a sense in which both John and George see the same thing. The first difference is that only George sees that thing *as* a man in a suit. The second similarity is that they both know that it is a man in a suit. The second difference is that in spite of his knowledge, John is incapable of seeing it as a man in a suit. The explanation of the first similarity is that John and George both have visual sensations with the same objective intentionality. They both have states of

their visual system that causally co-vary with, for example, the distinctive way that a man in a gorilla suit moves. The explanation of the first difference is that only George is able to automatically (without an intervening inference) apply the concept of a man in a gorilla suit to the thing causing his current visual sensation and thus only George's sensations have the subjective intentionality indicating the presence of a man in a gorilla suit. The explanation of the second similarity depends on nothing peculiar to Churchland: they both know that the thing is a man in a gorilla suit because they have justified true beliefs that it is a man in a gorilla suit. The explanation of the second difference is that, unlike George, John is incapable of automatically (without an intervening inference) applying the concept of a man in a gorilla suit to the thing causing his current visual sensation, and thus John's sensations lack the subjective intentionality indicating the presence of a man in a gorilla suit.

Let us briefly reconsider the example discussed at the beginning of this chapter, namely the distant horse that looks like a speck on the horizon. If the distant speck is indeed a horse and someone were incapable of automatically applying the concept of a horse to the cause of their visual sensation, then even if they knew it was a horse they would be incapable of seeing it *as* a horse. In contrast, if they were able to automatically apply the concept of a horse to the cause of their visual sensation, then they would be seeing the distant speck as a horse: the distant speck would seem like a horse to that person.

The appropriate analogy, then, to introspection would be the following. If a person knew that their mental states were identical to brain states, but was incapable of *automatically* applying the concept of a brain state to a mental state, then in spite of their knowledge they would be incapable of introspecting their brain states *as* brain states. In contrast, if they were able to automatically apply the concept of a brain state to their brain states then they would be introspecting their brain states as such: their brain states would seem like brain states to them.

Thus concludes my initial discussion of Churchland's Introspection Thesis and his defense of it. I turn next to unpack the Transparency Thesis and the opposition between it and Churchland's account of introspection.

THE ALLEGED TRANSPARENCY OF CONSCIOUS EXPERIENCE

Contemporary discussions of the notion that experience is transparent (or diaphanous) frequently trace the idea back to the following G. E. Moore quotation.

> [T]he moment we try to fix our attention upon consciousness and to see what, distinctly, it is, it seems to vanish: it seems as if we had before us a mere emptiness. When we try to introspect the sensation of blue, all we can see is the blue: the other element is as if it were diaphanous. (Moore 1903: 25)

However, for the sake of historical accuracy (at least), it is worth noting that while Moore discusses the Transparency Thesis, he does not actually endorse it. Transparency is introduced in the contemporary literature (and endorsed) by Harman:

> When Eloise sees a tree before her, the colors she experiences are all experienced as features of the tree and its surroundings. None of them are experienced as intrinsic features of her experience. Nor does she experience any features of anything as intrinsic features of her experiences. And that is true of you too. There is nothing special about Eloise's visual experience. When you see a tree, you do not experience any features as intrinsic features of your experience. Look at a tree and try to turn your attention to intrinsic features of your visual experience. I predict you will find that the only features there to turn your attention to will be features of the presented tree . . . (Harman 1990: 667)

Harman's interest is in a defense of functionalism (wherein mental states are type identified in terms of their causal relations, not, *pace* typical qualiaphiles, in terms of their intrinsic properties). Along the way, he defends a kind of representationalism: the objects of experience are intentional objects. Other adherents of the Transparency Thesis who utilize it in the defense of representationalism include Tye (1995, 2000) and Dretske (1995).

Although the metaphor of transparency is visual and thus most appropriate for visual experiences, defenders of the Transparency Thesis intend it to generalize to all conscious experience. So, for example, as Dretske writes, "If one is asked to introspect one's current gustatory experience . . . one finds oneself attending, not to one's experience of the wine, but to the wine itself (or perhaps the tongue or palette)" (1995: 62).

We can get a further understanding of what is being affirmed and denied by the Transparency Thesis by seeing how disagreement over it divides various approaches to understanding consciousness. Advocacy of Transparency (frequently) goes hand-in-hand with First-Order Representationalism and goes against Higher-Order Representationalism. Roughly, First-Order Representationalism explains consciousness in terms of mental representations of aspects of the environment. Thus, according to

First-Order Representationalists, meta-representational states are strictly irrelevant to phenomenal consciousness. As Tye puts the point, "Cognitive awareness of our own feelings itself feels no special way at all. Phenomenal character attaches to experiences and feelings (including images), and not, I maintain, to our cognitive responses to them" (Tye 2000: 36–37). Dretske states his agreement regarding the irrelevance of meta-representational states for phenomenal consciousness as follows:

> Conscious mental states – experiences, in particular – are states that we are conscious *with*, not states we are conscious *of*. They are states that make us conscious, not states that we make conscious by being conscious of them. They are states that enable us to see, hear, and feel, not states that we see, hear, or feel. (Dretske 1995: 100–101)

According to First-Order Representationalism, to have a conscious experience of a blue square on the wall it suffices to have a (certain kind) of mental representation of a blue square on the wall. What kind of mental representation will suffice to give rise to consciousness is something that various First-Order Representationalists may disagree on. But in spite of their differences they agree that the mental representation in question need not itself be represented by any other mental representation in order to give rise to a conscious state.

Higher-Order Representationalism, in contrast, explains consciousness in terms of mental representations of other mental states. A key principle appealed to by Higher-Order Representationalists is the Transitivity Principle.

> In order to have a conscious mental state, one must be conscious *of* that mental state.

Thus, according to advocates of the Transitivity Principle such as Lycan (2001) and Rosenthal (2002), if one has a conscious experience of a blue square, it is insufficient to simply have a mental state that represents a blue square – having only a mental representation of a blue square would be having only an unconscious mental representation of a blue square. One must additionally have a mental representation of the mental representation of the blue square, that is, a second mental representation which represents the first representation, which, in turn, represents the blue square.[5]

The Transitivity Principle gets its name from the fact that consciousness in the intransitive sense of the term (e.g., "Mary's experience was conscious") is being explained by consciousness in the transitive sense of the term (e.g., "Mary was conscious of her experience"). The English word "conscious"

has several uses in the construction of verb phrases, some of which yield transitive verb phrases (e.g., "John was conscious of the smell of coffee") and some of which yield intransitive verb phrases (e.g., "John was conscious" and "John's desire was conscious").

The tension between transparency and transitivity becomes apparent when we note that the higher-order representations must represent aspects of the first-order states themselves. If so-called higher-order states simply had the same contents as their first-order targets, then they wouldn't really be higher-order after all. What makes a mental representation first-order is that it isn't meta-representational – it doesn't represent itself or any other mental representations but instead represents, for example, aspects of the creature's environment or body. If the so-called higher-order state didn't represent aspects of the first-order state itself, but instead represented what the first-order state represents, then the so-called higher-order state would be representing, for example, aspects of the creature's environment or body and would thus itself be a first-order state.

Another way of putting the previous point is in terms of a distinction between representational content and representational vehicle. I may have, at 3 P.M., a memory of something that happened at 2 P.M. – I may remember that at 2 P.M. someone told me a particularly funny joke. Occurring at 3 P.M. is a property of the representational vehicle, it is a property of the memory itself. Occurring at 2 P.M. is a property of the content – it is a property of what was remembered, namely, that a funny joke was told. A second class of examples of the vehicular properties of a mental representation includes the neurophysiological properties of a mental representation. The neurophysiological properties of a first-order representation are typically vehicular properties of that representation. For example, the pattern of neural activation that constitutes my perception of a green bottle three feet away from me is neither green, a bottle, nor three feet away from me. With the distinction between content and vehicle thus in hand, the main point here is as follows. If a representation doesn't represent any of the vehicular properties of some other representation, but simply has similar contents to the second representation, then the first representation isn't a representation of the second representation, and thus isn't a higher-order representation.

The vocabulary of "content" and "vehicle" allows us to formulate the opposing theses of transparency and transitivity as follows. According to the Transparency Thesis favored by First-Order Representationalists, when we have a conscious first-order representation, all we can be conscious of are the contents of that representation, we are thus incapable of becoming

conscious of any of the vehicular properties of that representation. According to the Transitivity Principle, then, when we have a conscious first-order representation we must be conscious of (among other things) vehicular properties of that representation.

These two theses, while exclusive, are not exhaustive. The middle ground left open by merely denying transparency without necessarily affirming transitivity is where we find Churchland's Introspection Thesis, since Churchland's thesis entails that we *can* be conscious of vehicular properties of our first-order representations but does not entail that we *must* be conscious of vehicular properties of our first-order representations. Churchland's thesis states that a suitably educated individual can become aware of their own brain states as such and as I have argued above this means that a suitably educated individual can become aware of the vehicular properties of their first-order representations. It does not, however, entail that everyone who has conscious states must be aware of the vehicular properties of their first-order representations because it leaves open, as it should, that perhaps not everyone is suitably educated in the relevant neuroscience. The Transparency Thesis states that when one has a conscious state one cannot be conscious of the state itself, and as I have argued this entails that one cannot be conscious of the vehicular properties of the state. The Transitivity Principle states that one can have a conscious mental state only if one is conscious of that state, and as I have argued this entails that one must be conscious of the vehicular properties of the state. Thus does Churchland's Introspection thesis occupy a middle ground between Transparency and Transitivity. Transparency entails that you cannot be aware of vehicular properties of conscious states; Transitivity entails that you must be aware of vehicular properties of conscious states; and Churchland's thesis entails that you can, (but don't have to) be aware of vehicular properties of conscious states.

Not only does Churchland's Introspection Thesis conflict with the Transparency Thesis, but it threatens the larger project of representationalism. As Amy Kind (2003) has argued, if the Transparency Thesis is false then representationalism itself is false. If we can have introspective access to conscious states themselves and not just their representational contents, then there must be more to the phenomenal character of a conscious state than its representational contents. To be clear, the representationalism impugned by the falsity of the Transparency Thesis is First-Order representationalism. If we have introspective access to more than the contents of a first-order representation, then there is more to the character of consciousness than those contents. Of course, the possibility remains that the character

of consciousness is still fully determined by representational content, but if transparency turns out to be false, the content in question would include the content of higher-order representations.

It is instructive to see what the transparency thesis looks like when stated in Churchland's vocabulary. It becomes the thesis that while the objective intentionality of a sensation may include information about both itself and states external to it, the subjective intentionality of a sensation is limited to states external to the sensation. The interpretation functions imposed by the conceptual exploitation of sensations may map sensations onto states external to them but can not possibly map sensations onto themselves. Spelling this out further in terms of the analogy to measuring instruments, the claim of the Transparency Thesis becomes tantamount to claiming that while it is possible to calibrate a thermometer so that mercury column heights indicate temperatures, it is impossible to change the marks on the thermometer so that the mercury column heights indicate mercury column heights. That such a reinterpretation of our brain states should be absolutely impossible seems implausible. The implausibility is further heightened when we consider that the Transparency Thesis is supposed to be introspectively and/or pre-theoretically obvious. That something like Churchlandish introspection is impossible seems an odd candidate for something that we would have introspective or pre-theoretic access to.

Once we have the Transparency Thesis stated in a Churchlandish vocabulary, it is apparent that it is less plausible than Churchland's Introspection Thesis. Once we grant Churchland's general view of perception and introspection, namely, that both involve a procedure for mapping sensations onto judgments, it follows quite naturally that, contra the transparency thesis, it would be possible for a suitably educated person to introspect his or her own brain states as brain states. That is, once we grant that sensations carry information about lots of things including themselves, and that perception involves interpreting sensations in ways so that we conceptually exploit the information already contained in the sensations, then there is no reason for it to be impossible to interpret sensations in ways so that we conceptually exploit the information that sensations carry about themselves.

Given the dependence of the introspection thesis on Churchland's views concerning perception and introspection, a natural move for the friend of transparency would seem to be to question such views of perception and introspection. However, it is not clear that such a move would actually be available to the current defenders of transparency. For example, Tye would seem to be hard pressed to deny such views since they seem very close to

his own. Consider, for example, Tye's description of the introspection of our own thoughts and experiences:

> [I]ntrospection of thought contents is a reliable process that takes as input the content of the thought and delivers as output a belief or judgment that one is undergoing a state with that content . . . We acquire introspective knowledge of what it is like to have such-and-such an experience or feeling via a reliable process that triggers the application of a suitable phenomenal concept or concepts. This reliable process . . . takes as input the direct awareness of external qualities (in the perceptual case). . . . (Tye 2000: 53)

The view that introspection involves a mapping process is common to both Churchland and Tye. Of course, whereas Churchland uses the language of "mapping" *x*'s onto *y*'s Tye instead speaks of processes that have *x*'s as inputs and *y*'s as outputs. However, I do not suppose that there is any difference between mapping and input-output processing, so whatever disagreement there must be between Churchland and Tye concerning introspection it must be a disagreement not about the relation involved, but instead about what the admissible relata are. And further, it seems that Churchland and Tye agree that the introspection of sensations would deliver as outputs judgments about sensations. So, whatever Tye could disagree about here would be limited to what the introspective judgments could be about. However, this disagreement simply is the disagreement between the Transparency Thesis and the Introspection Thesis. Therefore, a First-Order Representationalist such as Tye cannot object to Churchland's argument for the Introspection Thesis on grounds concerning the general nature of introspection, that is, whether it is a reliable process that yields judgments about sensory states.

Not only is Churchland's Introspection Thesis more plausible than the thesis of Transparency, but the premises upon which the Introspection Thesis is based can also be used to explain whatever initial plausibility the Transparency Thesis enjoys. The natural explanation that emerges is the following. Transparency is plausible because the mappings of sensations onto propositions that people typically acquire first are mappings that involve judgments about external world objects. Children learn to call objects blue, red, and so on way before (if ever) they learn that there are such things as blue sensations, red sensations, and so on. Further, this kind of mapping is relatively entrenched: it takes a bit of (philosophical?) sophistication for it to occur to any one to map things in any other way, that is, to map sensations onto judgments about sensations as opposed to judgments

about external world objects and their properties. Thus, transparency may seem plausible without being true.

CHURCHLAND IN TROUBLE?

The strength of the Introspection Thesis not only spells trouble for representationalists such as Tye and Dretske, but it may very well spell trouble for Churchland himself, since there is evidence from other parts of his corpus that he himself may be a representationalist. I will here briefly review the *prima facie* evidence, both pro and con, regarding whether Churchland is indeed a representationalist. I will argue that the cons will outweigh the pros: in spite of a few superficial appearances, Churchland is ultimately *not* a representationalist.

One general consideration that favors regarding Churchland as a representationalist is that he is, in general, quite sympathetic to representational (and computational) approaches to cognition and it would thus not be incongruous for Churchland to think that qualia were amenable to a representational/computational analysis. However, this is not the sole consideration that favors reading Churchland as a representationalist. One of the most striking pieces of evidence implicating Churchland's sympathy for representationalism comes from the article "Some Reductive Strategies in Cognitive Neurobiology." In a section entitled "The Representational Power of State Spaces", the bulk of the discussion is concerned with the topic of qualia. The relevant notion of state spaces and their neural implementation is conveyed in the following quotation

> The global state of a complex system of *n* distinct variables can be economically represented by a single point in an abstract *n*-dimensional state space. And such a state-space point can be neurally implemented, in the simplest case, by a specific distribution of *n* spiking frequencies in a system of only *n* distinct fibres. Moreover, a state-space representation embodies the *metrical* relations between distinct possible positions within it, and thus embodies the representation of *similarity* relations between distinct items thus represented. (Churchland 1986: 102, emphases in original)

The upshot of the discussion that ensues can be conveyed in terms of one of Churchland's major examples: color. He endorses Land's view that the perceptual discriminability of reflectances by humans is due to the reception of three kinds of electromagnetic wavelength by three different kinds of cones in the retina. Further, he states that the degrees of perceived similarity

between colors are due to the degrees of proximity between the points in neural state space that represent those colors. Churchland is quite clear in his intent to identify color sensations with the neural representation of colors (the neural representation of spectral reflectance). Thus he endorses, or at least entertains,

> . . . the hypothesis that a visual *sensation* of any specific color is literally identical with a specific triplet of spiking frequencies in some triune brain system. If this is true, then the similarity of two color sensations emerges as just the proximity of their respective state-space positions. And, of course, there are an indefinite number of continuous state-space paths connecting any two state-space points. Evidently, we can reconceive the cube [depicting the three dimensions of coding frequencies for reflectance in color state space] as an internal "qualia cube". (ibid: 105)

Churchland's endorsement of the identification of color sensations with the neural representation of color seems like a straightforward endorsement of First-Order Representationalism, at least as far as color is concerned. Further, the rest of his discussion in "The Representational Power of State Spaces" section of "Some Reductive Strategies in Cognitive Neurobiology" makes it quite clear that he intends the representational approach to generalize to other sensory qualia, since he goes on to discuss gustatory, olfactory, and auditory qualia (ibid: 105–106). I think however, that these remarks can ultimately be read as consistent with the falsity of both First-Order Representationalism and the Transparency Thesis.

When wondering whether Churchland's identification of sensations with neural state space representations is in tension with the Introspection Thesis it is useful to note that in "Some Reductive Strategies . . . " he explicitly portrays the two views as compatible. He writes

> The "ineffable" pink of one's current visual sensation may be richly and precisely expressible as a 95Hz/80Hz/80Hz "chord" in the relevant triune cortical system. The "unconveyable" taste sensation produced by the fabled Australian health tonic Vegamite [*sic*.] might be quite poignantly conveyed as a 85/80/90/15 "chord" in one's four-channeled gustatory system (a dark corner of taste-space that is best avoided). And the "indescribable" olfactory sensation produced by a newly opened rose might be quite accurately described as a 95/35/10/80/60/55 "chord" in some six dimensional system within one's olfactory bulb.
>
> This more penetrating conceptual framework might even displace the commonsense framework as the vehicle of intersubjective description and spontaneous introspection. Just as a musician can learn to recognize the

> constitution of heard musical chords, after internalizing the general theory of their internal structure, so may we learn to recognize, introspectively, the *n*-dimensional constitution of our subjective sensory qualia, after having internalized the general theory of *their* internal structure. (ibid: 106)

Note, however, in the examples in the quotation that what are introspected may very well be conceived of as vehicular properties. The chords in multidimensional systems that he discusses are chords in multidimensional *neural* systems. Another way of putting the point of how Churchland's remarks about qualia and representations can be consistent with the falsity of First-Order Representationalism is by noting how the word "representation" can often pick out only a representational vehicle. That is, while sometimes "representation" picks out the process of *x*'s representing *y*, at other times "representation" can be used to pick out only *x* itself, the thing which represents as opposed to the act of representing. What Churchland is doing is identifying sensations with certain representations. However, this identification is consistent with the view that we have introspective access to aspects of the representations other than their representational contents.

I turn now to briefly consider two considerations in favor considering Churchland as an antirepresentationalist (aside from whatever is implied by his endorsement of the Introspection Thesis). The first consideration is that it seems that Churchland thinks that the identity of a sensation is due to its intrinsic features and that whatever intentionality it has is due to extrinsic features. Churchland illustrates the point in terms of an extended thought experiment concerning beings who perceive temperatures, but in virtue of what in us would be the visual sensations we associate with light and dark (Churchland 1979: 8–14). In his summary of the crucial points of the thought experiment, Churchland notes the following:

> It is clear [. . .] that neither the objective nor the subjective intentionality of a given kind of sensation is an intrinsic feature of that kind of sensation. Rather, they are both relational features, involving the sensation's typical causes in the former case, and its typical (conceptual) effects in the latter. And it is equally clear that both the "of$_o$-ness" and the "of$_s$-ness" of one and the same kind of sensation can vary from being to being, and even over time within the history of a single individual, the variation being a function of differences or changes in sensory apparatus in the case of objective intentionality, and of differences or changes in training and education in the case of subjective intentionality.
>
> [. . .]

> [T]he intrinsic qualitative identity of one's sensations is irrelevant to what properties one can or does perceive the world as displaying. (ibid: 15)

⋮

The tension between this quotation and any representationalist interpretation of the remarks on sensations as state-space representations should be relatively obvious, but in case it requires spelling out, the salient contrasts are two fold: the intrinsic/extrinsic contrast and the intentional/nonintentional contrast. The view expressed in "Some Reductive Strategies . . . " identifies sensations with points in a neural state space and points in a space are individuated extrinsically by their relation to all of the other points in that space. In contrast, the qualitative identity of sensations is characterized as intrinsic in the 1979 quote. Second, and most important, the view expressed in "Some Reductive Strategies . . . " identifies sensations with states having a certain kind of intentionality. In the case of color sensations, the sensations are identified with points in neural state-space that represent spectral reflectance. In contrast, the 1979 quote presents the intrinsic qualitative identity of a sensation as distinct from both its objective intentionality and its subjective intentionality.

A second consideration that casts some doubt on whether Churchland is committed to representationalism comes from an article co-authored with Patricia Churchland, "Functionalism, Qualia, and Intentionality." Of particular interest is a section of that article entitled "The Problem of Distinguishing States with Qualia from States Without" wherein the Churchlands affirm and attempt to explain how it is that a sensation has qualitative character whereas a belief does not. In brief, the Churchlands' explanation of the difference is that, in introspection, "Sensations are identified by way of their intrinsic properties; beliefs are identified by way of their highly abstract structural features" (Churchland and Churchland 1981: 33). According to the Churchlands, "the number of possible beliefs is at least a denumerable infinity" whereas the finite number of continua (and positions on them) that characterize distinct qualia "is sufficiently small" for noninferential discriminatory mechanisms to exploit the intrinsic qualities that define sensations (ibid: 32). The problem for discriminating beliefs, allegedly, is not that beliefs lack characteristic intrinsic properties, but instead that there are too many of them "for us to have any hope of being able to discriminate and identify all of them on such a one-by-one basis" (ibid: 32). Ultimately, the suggestion here seems to be that what makes a sensation a state with qualia is not merely that it has intrinsic qualities (because beliefs have those

too), but that it may be introspectively identified in terms of its intrinsic qualities.[6] While the Churchlands do not explicitly address the question of whether sensations are to be distinguished from beliefs in terms of sensations lacking intentionality, the point that emerges of relevance to Paul Churchland's antirepresentationalism is that if a sensation can be introspectively identified in terms of its intrinsic qualities, this entails that it need not be introspectively identified in terms of its intentionality, since whatever intentionality it has would be extrinsic to it. This is in strict contrast to the view of the introspection of sensations advocated by representationalism, especially adherents of the Transparency Thesis, since on the latter view one may only introspect the representational contents of sensations, and *not* any of their intrinsic properties.

It is worth briefly noting an addendum that closes this section in the version anthologized in Paul Churchland's (1989) *A Neurocomputational Perspective*. In it Paul Churchland writes

> I must now express a loss of confidence in this argument. The problem is that sensations now appear to be decidedly more various than I had originally estimated and to have a much more intricate combinatorial structure than I had earlier supposed (see [the "The Representational Power of State Spaces" section of "Some Reductive Strategies in Cognitive Neurobiology"]). Accordingly, the contrasts on which the preceding argument places so much weight now appear spurious: what seemed a large difference in kind now seems a mere difference in degree. (Churchland 1989: 33)

Whatever, precisely, Churchland is abandoning in this addendum, this much remains clear: Whatever it is that is being trumped, the considerations that are doing the trumping come from the "The Representational Power of State Spaces" section of "Some Reductive Strategies in Cognitive Neurobiology," and as I have already argued above, that material is entirely consistent with viewing Churchland as being opposed to representationalism.

One way of getting a handle on Churchland's position is in terms of a distinction between two meanings of "representationalism" that have emerged in the contemporary philosophical and cognitive scientific literature. The first sense is the broad view that all mental states and processes are representational. The second is the more narrow view concerned with consciousness and has also been the primary focus of this paper: it is the view that the properties of perceptual consciousness are the representational contents about environmental objects and properties. To give more precise labels to these

views we could use "representationalism about the mind" for the first view and "first order representationalism about phenomenal consciousness" for the second view. Churchland is an adherent of the first view but not an adherent of the second.

That Churchland is not an adherent of the second view should be relatively obvious by now, since it has been the major task of this chapter to argue this point. However, it may take further work to make it clear how Churchland is an adherent of the first view, especially in light of his commitment to the possibility of the "direct introspection of brain states." That brain states are *directly* introspectible may very well make it seem like the vehicular properties of experiences are entering into consciousness and thus the overall character of one's mental life contains more than just representational contents, but also includes vehicular properties themselves. I think, however, that this interpretation of Churchland is ultimately in error. To see this most clearly it is useful to consider how, as mentioned previously, Churchland regards introspection as fallible. The possibility of erroneous introspection is explained in terms of the possibility of introspectively misrepresenting sensations. It is natural to suppose, then, that in such cases, what enters into consciousness are not the sensations themselves, but the ways in which the sensations are represented (which may include inaccurate as well as accurate ways of representing them). What follows, then, from the Introspection Thesis is not that the vehicular properties of first-order states enter into consciousness, but that what enters into consciousness in the direct introspection of brain states are the *contents* of higher-order representations (which are, of course, representations of the vehicular properties of the first-order representations).

What, then, are we to make of the "direct" in "the direct introspection of brain states"? I think that, for Churchland, the directness here is that the introspective judgments are *noninferential*. That is, they are not inferred from introspective evidence but *directly* caused by the occurrence of their target sensations.

I close, then, by briefly summarizing what I take myself to have shown in this chapter. My primary aim was to unpack Churchland's Introspection Thesis and pit it against the ultimately inferior Transparency Thesis of the representationalists. I then considered the question of whether the tension between the Introspection Thesis and representationalism plays itself out in Churchland's corpus. I argued that the evidence that Churchland is a representationalist can be easily explained away and is further overwhelmed by contrary evidence. The position that emerges from these considerations is a kind of representationalism about consciousness but not in a sense

equivalent to either First-Order or Higher-Order Representationalism. As I read Churchland, the qualitative character of consciousness is always identical to the content of some representation or other but, contra First-Order Representationalism, it need not always be the content of a first-order representation and contra Higher-Order Representationalism, it need not always involve the presence of some higher-order representation. If there is, however, a tension remaining in Churchland's work concerning the qualitative character of consciousness it is a tension concerning whether he ultimately thinks that the intrinsic properties of neural states themselves can enter into consciousness or whether it is the *representation of* intrinsic properties that enters into consciousness. I offer that the latter option is the superior view, and if it isn't what Churchland explicitly has in mind, then it should be.

I have pitted Churchland's bold and surprising Introspection Thesis against the allegedly obvious yet opposing Transitvity Principle and the Transparency Thesis. Both the Transitivity Principle and the Transparency Thesis are supposed by their proponents to be pre-theoretically intuitively obvious, but once we see what they entail, it is not clear how they can be pre-theoretically intuitively obvious. Transitivity entails that it is *necessary* that we are aware of vehicular properties of our conscious states and Transparency entails that it is *impossible* for us to be aware of such properties. Both claims seem too strong to be accessible to pre-theoretic intuition. In contrast, Churchland's Introspection Thesis, in spite of its bold and surprising content, turns out to be the most plausible of the three.

ACKNOWLEDGMENTS

This work was supported in part by grants from The National Endowment of the Humanities and The James S. McDonnell Foundation's Project for Philosophy and the Neurosciences. I am grateful for feedback on this material from Brian Keeley and Josh Weisberg and from members of the William Paterson University Cognitive Science Reading Group, especially Mike Collins, Thomas Hogan, and Alexander Vereschagin.

Notes

1. Churchland's most extended treatment of the Introspection Thesis is Churchland (1979) but see also Churchland (1985, 1986).
2. See, for example, Tye (1995, 2000) and Dretske (1995).
3. See, in particular, Tye (2000: 45–51).

4. Although, as I'll argue later, just because representational contents *of the experience* do not figure into phenomenal character, this doesn't mean that phenomenal character is anything besides representational content. It may involve, for instance, the representational contents of higher-order states.
5. While the transitivity principle is typically taken to be satisfied by a second representation, it is at least *prima facie* possible for the principle to be satisfied with a single representation that is, in part, self-representational.
6. This remark perhaps suggests a *dispositional* higher-order representational theory of qualia insofar as a state has qualia in virtue of certain dispositions there are for its uptake by the higher-order representations employed in introspection. I will not explore this possibility further beyond noting it here.

References

Churchland, P. M. & Churchland, P. S. (1981). "Functionalism, qualia and intentionality." *Philosophical Topics* **12**: 121–32. Reprinted in *A Neurocomputational Perspective* (Cambridge, MA, MIT Press, 1989).

Churchland, P. M. (1979). *Scientific Realism and the Plasticity of Mind*. Cambridge, Cambridge University Press.

Churchland, P. M. (1989). *A Neurocomputational Perspective: the Nature of Mind and the Structure of Science*, Cambridge, MA, MIT Press.

Churchland, P. M. (1985). "Reduction, qualia and the direct introspection of brain states." *Journal of Philosophy* **82**: 8–28. Reprinted in *A Neurocomputational Perspective* (Cambridge, MA, MIT Press, 1989).

Churchland, P. M. (1986). "Some reductive strategies in cognitive neurobiology." *Mind* **95**: 279–309 Reprinted in *A Neurocomputational Perspective* (MIT Press, 1989).

Dretske, F. (1995). *Naturalizing the Mind*. Cambridge, MA, MIT Press.

Harman, G. (1990). "The Intrinsic Quality of Experience." In *The Nature of Consciousness*, Ned Block et al, eds. Cambridge, MA, MIT Press, 663–75.

Kind, Amy. (2003). "What's So Transparent About Transparency," *Philosophical Studies* **115**: 225–244.

Lycan, W. G. (2001). "A simple argument for a higher-order representation theory of consciousness." *Analysis* **61**: 3–4.

Moore, G. E. (1903). "The Refutation of Idealism." *Philosophical Studies*. Totowa, NJ, Littlefield, Adams & Co., 1–30.

Rosenthal, D. (2002). "Explaining Consciousness," in *Philosophy of Mind: Classical and Contemporary Readings*, ed. David J. Chalmers, New York, Oxford University Press, 406–21.

Tye, M. (1995). *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. Cambridge, MA, MIT Press.

Tye, M. (2000). *Consciousness, color, and content*. Cambridge, MA, MIT Press.

4 Empiricism and State-Space Semantics

JESSE J. PRINZ

For some time now, Paul Churchland has been defending a connectionist theory of mental content, alliteratively labeled state-space semantics. The theory is at once holistic, prototype-based, and neurally reductionist. If you mix those ideas into a single theory, one thing is guaranteed: you will hear from Jerry Fodor and Earnest Lepore. The exchange between Churchland and the Rutgers duo has spanned several papers, and several voices have been chiming in from the sidelines. Both sides have claimed victory, but these declarations are premature. I think Churchland has been successful in addressing some of the objections levied by Fodor and Lepore, but others remain. Fodor and Lepore recognize that there may be a way out for Churchland. He could overcome some of their more pressing concerns if he embraced an extreme form of concept empiricism. In pointing this out, they intend to highlight the fatality of their objections. Empiricist theories of concepts are unacceptable on other grounds. No one, not even Paul Churchland, wants to resurrect Hume. If empiricism is the only way to navigate state space, it's time to give up on that program and set course for another semantic theory. This is a rare point of agreement between Churchland and his critics.

I will argue that empiricism isn't as unworkable as it appears. If I am right, then state-space semantics may be able to withstand the objections of Fodor and Lepore. But that does not mean Churchland wins the debate. Empiricism raises further concerns about state-space semantics. These concerns suggest that an adequate semantic theory will borrow more from Hume than from Hebb.

1. SPACE WARS

1.1 State-Space Semantics

Connectionism is undeniably seductive. Networks learn, and they often behave in ways that are psychologically realistic. For example, they

spontaneously form prototypes. And they do this in a way that is neurally plausible, broadly speaking. Networks show that we can get mind-like performance out of brain-like models, and that is a big step forward in solving the mind-body problem. Turing showed that mentation could be physical but he hadn't a clue about how it was actually implemented in naturally evolved brains.

Some critics of connectionism concede that it provides a promising strategy for explaining certain aspects of perceptual pattern recognition, but they suspect that it will shed little light on higher-level cognition. For cognitive tasks, such as forming thoughts, networks must implement language-like symbols. Once we get to the level of symbols, the argument continues, we can drop network talk entirely. Churchland is a long-time critic of this view. He thinks that symbolic cognition is a bad theory. Connectionism can be used to explain high-level cognition and illuminates the path to radical new theories of mentation. Thinking is vector transformation, not symbol crunching.

But what exactly are vectors and how do they bear on mental content? Churchland's basic proposal begins with the observation that we can represent connectionist networks geometrically. Networks achieve their magic by assigning weights to the connections between units. Input units and output units are customarily hand-coded by experimenters, and they usually correspond to external sensory inputs or behavioral outputs. The hidden units are where the action is. They correspond to how the network "understands" the input. If you want to know what's on a network's mind, look to see what its hidden units are doing. A pattern of activity across a group of hidden units is a vector. Vectors can be represented as points in a multidimensional space, with dimensions corresponding to activity levels in units. The total space corresponds to the total range of possible activations. The space can be subdivided into partitions by tracing boundaries around all the vector points corresponding to a range of inputs that produce the same output. Any point within such a boundary (or hyperplane) may be active when the network recognizes that an input falls into a given category. But there will also be an average. Within a hyperplane one can find subvolumes corresponding to the central tendency of the network when it responds to a category. That subvolume is comparable to a prototype for the category. Concepts can be identified with these multidimensional prototypes.

Churchland's theory is holistic, because prototypes are defined by their position within an entire network. The location of a prototype depends on average activation in every hidden unit of the network, and those activations depend, in turn, on the weights throughout the network, which depend, in

turn, on how the network has come to respond, through training, to every category that it can recognize.

Churchland's theory is also reductionist. He often assumes that the units in a network correspond to neurons. The position of a concept in vector space is, thus, also its position in neural activation space. Prototypes are average patterns of neural response. Strictly speaking, Churchland does not need to go this route. He could say that units reside at a level of analysis higher than individual neurons. This, however, would detract somewhat from the allure of connectionist semantics. Once we ascend to higher levels, we open up the possibility of shifting to a theoretical vocabulary that is not based on neural network architecture.

In this formulation, Churchland's theory is about content. The content of a concept can be specified by giving its conditions of individuation, and, for Churchland, concepts are individuated by their geometrical position within a network. If Churchland wants to show that two networks share concepts, he will have to avail himself of internal similarities in their state spaces. Churchland recognizes that exact identity across networks is unlikely, but he isn't bothered by this. He thinks we can relax the requirement that concepts are widely shared across individuals, and replace it with the claim that individuals often have similar concepts. People can have similar state spaces. Fodor and Lepore think this proposal is fraught with serious problems.

1.2 Similarity in State Space

Fodor and Lepore (1996a, 1996b, 1999) launch a multipronged attack against state space semantics. The attack is largely an application of objections that they have marshaled against other holistic semantic theories. They think that the connectionist framework does nothing to allay familiar difficulties with holism, and it introduces difficulties of its own. There are three main worries that I want to discuss. In a recent paper, Fodor and Lepore 1999) add a fourth worry, saying that state-space semantics cannot do the work for which semantic theories are needed. Most of their arguments for that allegation concern the adequacy of state-space semantics for explaining linguistic meaning. I will ignore these arguments, because I am mainly interested in whether state-space semantics can explain *mental* content. I will, however, touch on issues raised in that latest round of critique. The three worries that I will consider are as follows.

First, Fodor and Lepore (1996a) argue that Churchland does not have an adequate account of how state-space dimensions are to be individuated.

To establish the location of a concept in state-space, one must specify the dimensions of that space. If state-space semantics is a semantic theory, dimensions must be equivalent to semantic features: they are meaningful. But any feature-based theory of meaning must indicate which features are fundamental. Which features are the primitives from which more complex concepts may be derived? Churchland sometimes presents cases for which this question has a relatively plausible answer. For example, he talks about mental representations of colors. These may be built up from sensory primitives. The dimensions can be equated with ranges of wavelengths that are registered by the three types of color sensitive cells in the retina. But what are the primitives for more abstract representations? What dimensions make up the semantic space containing the concept UNCLE or the concept CLEOPATRA? One might think that these concepts reside in state spaces that contain the dimensions: BROTHER and POLITICIAN, respectively, but that just raises the question of how these two concepts get their meaning. If BROTHER and POLITICIAN are defined relative to state spaces, then Churchland needs to tell us what the dimensions of those spaces are. There must be a ground floor. Churchland needs a principled way of specifying which dimensions are primitive. If he can't deliver, he will not be able to compare networks. Similarity must always be defined relative to features that are fixed. Here Churchland faces a serious dilemma. He might go the way of classical empiricism, and argue that the primitive are all sensory. This, for reasons discussed below, seems hopeless to Fodor and Lepore, as well as Churchland. Alternatively, he might accommodate nonsensory primitives by showing that some concepts can get their content in a way that does not depend on their position is state space. For example, he might invoke an informational semantics. But once he does this for some concepts, he might as well drop the whole business of state spaces and do it for all (Fodor and Lepore 1999). If empiricism is false, what principled difference is left between the concepts that are assigned contents by relation to other concepts, and the concepts that are assigned contents by relation to the world?

Second, even if Churchland could specify the primitives, Fodor and Lepore (1996a) think he will face a problem of semantic relevance. Concepts are located in a multidimensional space with a dimension corresponding to every primitive in the system. Fodor and Lepore think that it is implausible that *each* of these dimensions is semantically relevant. If they are right, then Churchland must distinguish the dimensions relevant to the meaning of each concept from those that are not. That is, he must draw something like an analytic/synthetic distinction. Fodor and Lepore are persuaded by Quine

(1953) that such a distinction will not be forthcoming. Roughly, definitions of analyticity are either circular because they invoke semantic concepts that presuppose analyticity, or they are epistemically implausible, because they presuppose a notion of unrevisability that cannot be reconciled with the holistic nature of confirmation.

The third problem raised by Fodor and Lepore (1996a) concerns a similar worry. No matter how Churchland slices it, the identity of a concept will depend on its location relative to other features represented in the system. There is nothing to guarantee that any two people will have systems that contain the same features – indeed, biographical differences guarantee just the opposite. Any two people know different things about dogs, because they have experienced different dogs. If the content of a concept depends on its location is a mutidemsional space shaped by all experience, then no two people will share DOG concepts. People have different collateral information. There is nothing to guarantee that two people have the same number of dimensions in their state spaces, let alone that they have concepts occupying the same location relative to those dimensions that are in fact shared. To get around this problem, Churchland would need to invoke a notion of analyticity again, to find dimensions that are widely shared and highly consistent across individuals.

The collateral information problem and the relevance problem are two sides of the same coin. One focuses on the semantic intuition that some features are not relevant to meaning, and the other focuses on the fact that, were this not the case, meanings would be unsharable. In response to both worries, Churchland might try to circumvent Quine's critique of analyticity by invoking prototypes. On this approach, the dimensions that matter for individuating any given concept are the ones that have values that are highly typical, salient, and diagnostic for the category designated by that concept. If Churchland goes this route, he can avoid Quinean worries, because prototypes are determined statistically, not semantically or epistemically. But this takes Churchland from pan to fire. Prototypes are not compositional, and concepts must be compositional. This is a very familiar argument to anyone who has spent any time reading Fodor (see, e.g., Fodor 1998). The basic idea is that a feature can be prototypical for a compound concept without being prototypical parts (old saws are typically rusty, though neither old things nor saws are typically rusty). If concepts were not compositional, we could not entertain new concepts by simply combing familiar concepts. Every thinkable thought would have to be learned from scratch.

In sum, Churchland's theory seems to be saddled with some very serious problems. With no account of which dimensions are primitive and

which are semantically relevant, and little to guarantee dimensional alignment across individuals, he provides little hope for thinking that we can individuate concepts, much less individuate them in a way that allows for concepts to be shared. If concepts are unsharable, the whole enterprise of explaining behavior by appeal to psychological generalizations will collapse. Fodor himself is pessimistic about the future of psychology in any case, and Churchland thinks traditional psychological explanation will give way to neuroscientific theories. But, for the rest of us, losing psychological generalizations sounds like a disastrous outcome. Is there a way out?

1.3 Churchland's Response

There is, as we will see subsequently, an empiricist answer to the challenges put forward by Fodor and Lepore. Indeed Fodor and Lepore think that Churchland has been blind to the problems with his theory because of his covert commitment to empiricism. Churchland, as we will also see below, adamantly denies this charge. He has other strategies to offer.

Churchland's (1996a) initial response to Fodor and Lepore makes two key moves, as I read it. First, in response to the problem specifying primitive features, he says that networks will have highly idiosyncratic primitives that do not encode features that can be easily specified linguistically. Churchland (1996b) illustrates this with an example from a neural network that was trained by Gary Cottrell to recognize faces from photographs. If you look at the response profile of hidden units in the network, you discover that they do not encode simple features like nose size or mouth length. Instead, each unit responds to entire faces. If you were to depict the input that produces maximum response, each unit would represent a blurry phantom face, morphed together from all the training inputs. Following Janet Metcalfe, Churchland calls representations of this type "holons." If hidden units in networks are holons, and the dimensions of a state space correspond to hidden units, then there will be no hope for developing an account of primitives. The primitives will be weird, ineffable, and variable from network to network.

This move is a bit puzzling. Rather than addressing the concern about primitives, Churchland seems to compound it. If primitives are like that, then it is hard to see how networks can be compared. Two networks may form prototypes for the same face (average activation vectors across hidden units), but these will not be defined with respect to the same constituent features. Churchland's next move looks helpful at first, but it is even more puzzling on closer analysis. He says that sharing a representation is not a

matter of sharing prototypical hidden unit vectors. Rather, it is a matter of "rich causal and computational *sequelae*" of such prototypes (Churchland 1996a: 276). In particular, Churchland invokes the motor responses and perceptual expectations that people form. You and I share a concept that represents cats if we have connectionist prototypes that cause similar perceptual expectations and cat-related behaviors.

This "sequelae" reply is multiply mysterious. First, it is a departure from state-space semantics. Content is determined by functional role, not location in activation space. Second, Churchland offers no account of how perceptual expectations and motor responses are to be compared. Presumably two people don't enact the exact same motor response when they see cats, and perceptual expectations are contentful mental representations, so Churchland owes us a tractable accound of how these are shared (functional roles again? state spaces?). Identification by functional roles faces problems comparable to Fodor and Lepore's intial objections: what primitive representations pin down roles? Are global functional rules semantically relevant or just partial roles? Are functional roles shared in their entirely across individuals? And finally, the idea that the content of a concept could be fixed by perceptual expectations and motor commands stretches the imagination. There do not seem to be content-determining images and behaviors for CAT, let alone DEMOCRACY.

Churchland has since defended a more promising reply (Churchland, 1998). He calls on an innovation by Laakso and Cottrell (1999, 2000). They developed a mathematical technique for quantifying similarity across networks that does not rely on networks having the same hidden units. The technique can be informally summarized as follows. First, one assesses the distance between vectors in each network, considered individually. Pairwise comparisons of vectors yields a description of the basic structure of the vector space. Then one can compare the structure across networks by looking for similarities in the distances between points. Different networks, trained on similar stimuli, will have different weights, and hence different vector patterns for any input, but the overall similarity space for a given network may be very much like the similarity space for another network. It is important to note that this technique works even when networks have a different number of hidden units. Relative distances between vectors can be the same across networks even if those vectors have different parts.

This technique for comparing networks goes part of the way toward answering Fodor and Lepore's objections. It shows that strict identity of features is not necessary for measuring similarity. Similarity does not entail partial identity. The technique may also prove useful for neuroscience. No

two brains have the same number and configuration of cells. The Laakso and Cottrell technique suggests that researchers can abstract away from such differences to reveal isomorphisms. But does the proposal go far enough to save state-space semantics?

Fodor and Lepore (1999) are not convinced. They have two main counterarguments. First, they say that Churchland has confused neural state space with semantic state space. They accuse Churchland of slipping back and forth between brain talk and semantic talk. The Laakso and Cottrell technique looks like a way of comparing patterns of neural activation across systems, and Churchland gives no argument, they complain, for inferring similarities in content from similarities in neural activation.

This complaint can be answered by appeal to connectionist theory. Fodor regards concepts as arbitrary, innate symbols in a language of thought. There is nothing to guarantee that semantically related symbols are implemented by similar patterns of neuronal activation. But suppose concepts are learned, and suppose learning works by strengthening connections between neurons that are caused by our encounters with category instances. We see a bunch of dogs, and connections between neurons are shaped by those encounters. Then we see a bunch of cats, and the same thing happens. Chances are the resulting patterns of activation will be similar. This is supported by decades of neuropsychological studies on category specific disorders showing that semantically related concepts tend to cluster in the same brain regions. It also finds resounding confirmation in connectionist modeling research, where similarities in inputs reliably produce similarities in hidden unit vectors. It begs the question against Churchland to deny that representations of similar things have similar neural implementations. Fodor and Lepore say Churchland has no argument for that conclusion. To the contrary, his entire philosophical oeuvre practically oozes with examples that support similarity-preserving isomorphisms between mental representations and brain states.

Fodor and Lepore's second objection targets the Laakso and Cottrell technique more directly. They pick up on Churchland's suggestion that networks with fewer nodes can be embedded in networks with a larger number of nodes to establish similarity. They argue that embedding does not establish similarity in content (nor even similarity in neural activation). Here is their "counterexample":

> Suppose I have only three semantic dimensions Hard-Soft, Black-White, Heavy-Light. And suppose my concept ROCK is identified with a vector that specifies a region in the space that these dimensions define. Likewise for

> you . . . The difference between us is that, whereas these dimensions define the whole of my semantic space, they define only an embedded subspace of yours. Now, does it follow that our concepts ROCK are similar? We shouldn't have thought so. For, perhaps your space has a dimension ANIMACY which, by assumption, mine lacks. And suppose that you think that rocks are actually quite animate . . . [S]uppose also that your space contains a dimension for abstractness, and that you think that rocks are pretty abstract . . . Is your concept ROCK still similar to mine? If there are any such cases where the right answer is 'no', Churchland loses.

There is an obvious response to this. Churchland can simply insist, without trampling any intuitions, that the two ROCK concepts *are* similar. They are similar in just those dimensions where they overlap. I cannot fathom why Fodor and Lepore would deny this. One of the nice things about similarity, as opposed to identity, is that concepts can be both similar and different. The dimensional approach has resources for capturing both. Such similarities and differences are extremely valuable for psychological explanation. Both people in the rock example will identify a typical rock as a rock. The person with the deviant concept, however, will expect it to move around, and will try (hopelessly and hypocritically) to convince others that it is an abstract object. Moreover, if Churchland wanted to screen off some differences between networks as semantically irrelevant, or rule that certain similarities are more important than others, he could always appeal to prototypes. If animacy were part of one person's ROCK prototypes, and not part of another person's ROCK prototype, he could say there is a difference in what the two people mean by rock. If animacy and inanimacy were peripheral features for these individuals, Churchland could say they have the same ROCK concept and divergent rock beliefs. If the two individuals had entirely different prototypes, Churchland could say that their concepts are not semantically similar at all. His account has considerable flexibility for dealing with examples of this kind.

The rock case can actually be turned against Fodor's own semantic theory. It is likely that the person with the deviant rock concept is nevertheless reliable at detecting rocks. Her rock concept may be under the nomic control of rocks, despite the fact that she is prepared to make weird predictions when she encounters one. For Fodor, such nomic relations confer intentional content, and there is no other dimension of content on this theory. Thus, it could work out that the two individuals in the rock case have ROCK concepts that are not merely similar in content, but exactly identical, even if they have very different rock prototypes. If semantic content is restricted

to reference, radical differences in rock beliefs have no direct bearing on content comparisons. If there are any two concepts that are co-referential but semantically distinct, Fodor loses.

In sum, I think Fodor and Lepore have not refuted Churchland's proposal for comparing concepts across networks. That does not mean that Churchland is out of the water, however. Other objections to his proposal are imaginable. Calvo Garzón (2003) has argued that the Laakso and Cottrell technique fails to address one of Fodor and Lepore's initial objections. It does not address worries about collateral information. In demonstrating that their technique works, Laakso and Cottrell trained networks using the same training stimuli. They did not dramatically vary the experiences of the networks. In the case of human concepts, things are quite different. We all have different experiences, and we bring different beliefs to bear when we encounter category instances. These differences in experience, which Calvo Garzón likens to differences in collateral information, can result in very different state spaces. Laakso and Cottrell can cope with variation in network dimensionality, not with differences in network biography.

There is another concern that I have with the Laakso and Cottrell proposal. In effect, Churchland wants to use their method to bypass the worry about primitive features. He wants to say we do not need to specify what the dimensions are or how they should be semantically interpreted, because we can ascertain similarity across networks without taking this extra step. It could turn out that each dimension of comparison is as indescribable as the holons in the face recognition network mentioned above. Laakso and Cottrell prove that networks can be compared even if their units are holons. Thus, my initial worry about holons has been answered. There is, however, another worry. If representational similarity across networks were a matter of isomorphism in holon space, there would be no way to accurately describe the respects in which representations are similar. We could not say, for example, that Jones has a DOG prototype with pointy ears and Smith has a DOG prototype with hanging ears, because this would not be a relevant feature of comparison. Ear shape would be inextricably bound to primitive representations that each correspond to the shapes of entire dogs, rather than their parts.

In psychological explanation, it is often useful to compare representations by appeal to their component features. Psychologists measure prototype similarity by giving subjects feature listing tasks. Performance on those tasks correlates well with typicality rating and categorization efficiency, suggesting that the features that people list are psychologically real and causally efficacious. The very fact that people list features suggests

that they are extractable from our prototypes. Features also seem to be independently manipulable. We can imagine earless dogs, for example. In conceptual combination, we often modify individual features, leaving others in place. We imagine pink elephants by swapping the feature PINK for the default feature GRAY.

This worry is not the same as Fodor and Pylyshyn's (1987) widely discussed allegation that connectionist networks are not compositional. The compositionality objection has to do with phrasal concepts: it alleges that connectionist networks are incapable of combining two conceptual representations together to get a conceptual representation of the compound, from which the original two concepts can be recovered. The worry that I have would stand even if Fodor and Pylyshyn's worry could be answered. There are compositional methods of integrating connectionist representations, such as Smolensky's (1990) tensor products. These methods address Fodor and Pylyshyn's objection (despite their claims to the contrary), but they leave my worry unanswered. I am concerned about connectionist representations of lexical concepts, not phrasal concepts. I am concerned that connectionist nets do not, as a matter of course, represent concepts such as DOG or CAR or CHAIR by means of features, such as FURRY, or HAS WHEELS, or HAS A BACK. The new techniques for measuring similarity across networks – like the older techniques for combining connectionist representations – are deliberately designed to work on representations that do not decompose into discrete features. This is supposed to be an advantage, but, on scrutiny, it merely italicizes the fact that connectionist representations of categories are quite unlike the category representations used in human cognition. *Our* category representations have manipulable features, and representational similarities in human category space can be defined with reference to these features.

An adequate account of concept similarity must avail itself of interpretable features. Indeed, any adequate theory of concepts must do that. Features play a central role in psychological explanation. Holons may play a part in mental representation at some level of processing, but they cannot be the whole story. If we are to work within a connectionist framework, we need to work with one that is sufficiently localist to support intelligible comparisons of representations across networks. It is not enough to quantify similarity; we must be able to describe it.

The Laakso and Cottrell technique could be used to *discover* meaningful features that are implicit in the network. Consider the closely related technique of cluster analysis. Here pairwise similarity comparisons of hidden unit vectors are used to form hierarchically organized dendograms that

reveal meaningful groupings in a network's activation space. Sejnowski and Rosenberg's (1988) NETtalk was, most famously, analyzed using this technique, and it was discovered to have spontaneously separated vowels from consonants in converting written words to speech. In a similar vein, the Laakso and Cottrell technique could be used to identify clusters of similar vectors that share semantic significance. There are, however, three problems with this. First, the dimensions that emerge may be uninterruptible. As with multidimensional scaling and other statistical techniques that depend on similarity judgment, nothing guarantees that the dimensions used to plot similarities will have any significance. As remarked, the dimensions could be comparable to holons. Second, in plotting similarities, the number of dimensions can be arbitrary. Much depends on the algorithm used and the interests of the experimenters. Third, there is no reason to think that the emergent dimensions are causally important for the system. Indeed, even if they were interpretable and nonarbitrary in number, there would be a further question about whether the system can perform processes that are sensitive to one dimension and not others. The dimensions *might* be causally implicated at a higher level of analysis, and they *might* be extracted by other networks and used in feature-sensitive processing. The point is, nothing about the comparison technique gives us any reassurance that we are getting at the dimensions that play any psychologically important role. Of course, the Laakso and Cottrell technique could be applied to pairs of *localist* networks that include some units with the same interpretation. In that case, dimensions could correspond to causally relevant meaningful parts, but then the technique would not be a solution to the primitive feature problem. The problem would have already been solved.

Let me take stock. We need a way to label dimensions in semantic space. Once we have that, we can do straightforward similarity assessments across networks by counting off shared features (or weighted shared features). But finding labeled dimensions depends on having an adequate response to Fodor and Lepore's worry about primitive features. The Laakso and Cottrell technique is designed to bypass questions about what the primitive features are in a network, but now we see that there is an independent reason for thinking the question must be answered. The technique might be construed as a discovery tool by which primitive features can be identified, but, even if it could succeed in doing that, it would not guarantee that those features are available to the system for feature-sensitive processing. Thus, the primitive feature challenge remains. Indeed, the challenge splits in two. First, the plausibility of state-space semantics cannot be fully assessed until we have a nonarbitrary way of determining what semantic features a

network uses. (If network primitives turn out to be perceptual in nature, state-space semantics would prove to be a version of empiricism – something Chuchland wants to avoid.) Second, if semantic features are to play any role in cognition, networks must be sensitive to them. Nothing in the Laakso and Cottrell method of aligning networks guarantees that the emergent features are accessible in a way that allows them to be independently modified. So we don't know if the technique is getting at the features that really matter for developing cognitive and semantic theories.

2. HUME'S RAFT

2.1 How Empiricism Might Help

Recall that Fodor and Lepore think Chruchland is a covert empiricist. They think that he fails to notice the problem about primitive features because he is tacitly assuming that all concepts can ultimately be explained in terms of sensory primitives. We can appeal to the senses to find a principled basis for identifying primitive features. This is what Hume sought to do, when he distinguished simple and complex ideas. It is an elegant and appealing solution to the problem.

If concepts are built up from sensory features than we can compare concepts by sensory overlap. We can say that two DOG prototypes differ in the shape of the ears. We can also explain how representations of pink elephants are formed. Some sensory features will be difficult to verbally described. We have no established term for the particular shape of a curve in a representation of a dog's belly, but assemblies and equivalence classes of perceptual features can be readily named. We can describe a belly as concave, convex, or S-shaped (a horizontal sigmoid curve). We can think about belly shapes in isolation without thinking about other features. We can draw them, and we can transform them in our images. Imagine a pregnant Chihuahua.

The empiricist view also helps with the problem of collateral information. If two people have been exposed to different dogs, their DOG prototypes will nevertheless inhabit a space of perceptual features that can be readily compared. If you think dogs are vicious and I think they are docile and adorable, your dog prototype might include growling sounds and exposed teeth while mine depicts a wagging tail and a floppy tongue. You might experience shivers down your spine when you think of dogs, while I grin happily. These are patently describable differences. When state spaces have ineffable dimensions, we have little to go on but an abstract similarity

space: points and their relative locations. If you regard dogs as similar to snakes, and I regard them as similar to bunnies, the shapes of our state spaces will differ dramatically. If state spaces were the only basis of comparison, it would be hard to say what, if anything is shared by us. If we have interpretable sensory features to go on, we get explanatory purchase on the situation. You liken dogs to snakes because of their teeth and your tingling spine. I liken dogs to rabbits because of the goofy sentimentality they instill in me. Despite these space warping differences, there is plenty of overlap. We both think dogs look pretty much the same.

To drive this point home, consider two ways in which prototypes are represented in the psychological literature. Sometimes they are represented as points in a multidimensional space, and sometimes they are represented as lists of features. The spatial representations leave out a lot of information, especially when the dimensions are not labeled. The very same pattern of points could correspond to different similarity spaces; fruit space may be geometrically akin to mammal space. Individual differences are inexplicable. Feature lists are much more informative. They tell us what information is encoded in concepts, and when two people list different features, direct contrasts can be drawn. Feature lists can be used to plot a similarity space, so they retain information about conceptual relations without discarding information about the basis of those relations. I'm not suggesting that concepts are feature lists, by the way. Rather, I am suggesting that concepts are collections of sensory features, which can be labeled, in principle, but need not be. Feature lists are just a tool for describing perceptual prototypes.

Perceptual feature prototypes handle the problem of semantic relevance in much the way that Churchland would handle it. If concepts are prototypes, there is a principled, though statistical, contrast between features that are conceptually constitutive and those that are not. I noted previously, however, that Fodor and Lepore don't think that concepts can be prototypes, because prototypes are not compositional. I think the defender of perceptual feature prototypes is in a good position to address this problem.

There are ways of combing prototype representations that satisfy the compositionality requirement. That requirement says that there must be some way of combining familiar concepts to generate representations of arbitrary new novel concepts. If prototypes are distributed representations in connectionist networks, compositionality can be accommodated using one of the available techniques for vector combination. Smolensky's (1990) tensor products might do. If prototypes are sets of features (feature lists or collections of perceptual features), compositionality can be accommodated by pooling features together. Properly formulated, these methods are

sufficient for supplying concepts with all the compositionality that they need to explain the productivity of thought. As long as we can pool features or multiply vectors, we have a mechanical procedure for comprehending things that we have never thought about before.

For this reason, I think the compositionality objection is a pseudoproblem (see Prinz 2002). The real challenge is not explaining how concepts compose; it is rather explaining departures from compositionality. How is it that we often come up we features for compounds that are not features of the prototypes for component parts? I'd like to suggest that connectionist networks are not in a position to answer this question *unless* they have a way of assigning meanings to features. If networks were based on holons all the way down, emergent features would be impossible to explain. The reason for this is that emergent features are often portable. They can arise on the scene leaving other features intact. They also often come about by a process of reasoning, whereby individual features are taken into consideration. When we think of an old saw as rusty, it is because we know that saws are metal. We then recall that metal rusts with age. To arrive at this point, we must apply the adjectival feature OLD to individual features of the saw prototype. If the saw prototype were comprised of holons acquired from prior saw encounters, local application of the adjectival feature would be difficult to explain.

One could get around this problem by chaining neural networks together. One could have networks deigned to extract features from other networks at earlier stages of processing. The feature METAL could be extracted, in this way, from a holonic saw prototype. If that were always done as a matter of course, then all of our prototypes would come along with representations of their constituent features, and feature-sensitive processes, such as concept combination, could be performed. I have no qualms with this proposal. It is a variant of the empiricist view. Features are acquired on this model by an abstraction process. The real primitives (where decomposition bottoms out) may be holons, but there must also be more readily describable features that function as if they were conceptual building blocks when we perform conceptual tasks. Indeed, the describable features extracted from holons would probably be the lowest level features over which feature-sensitive processes would take place. Holon specific processes are unlikely. Thus, for all intents and purposes, the non-holonic features would serve as primitives for the system. In connectionist jargon, holons may be microfeatures, but psychologically important features are likely to be more discrete.

The main point that I have been trying to make is that any adequate account of conceptual similarity must ultimately allow for discrete

feature-based comparisons. I have also claimed that there is reason to think that the primitive features are perceptual. This empiricist proposal offers an answer to the primitive feature problem raised by Fodor and Lepore. If concepts are defined in terms of other concepts (meaningful dimensions in state spaces), things must bottom out somewhere. Sensory features are a natural place to bottom out (along, perhaps, with motoric features, used in controlling action). If primitives were all sensory, we would have a principled basis for establishing which features are primitives, akin to Hume's method of identifying simple ideas. But the question of which features are primitive is ultimately empirical. To defend the empiricist proposal, it would be nice to have some empirical support.

I think such support can be found (see Prinz 2002, Barsalou 1999). This is not the place to make a full empirical argument for empiricism. Much of the necessary laboratory work has yet to be done. But let me mention a couple of intriguing findings. The first is a general point about functional neuroanatomy. The majority of the neocortex is dedicated to sensory processing and motor response. There is some cross talk between sensory areas, and there are low-level bimodal cells that coordinate the senses, but one can fairly well circumscribe areas as a function of which input modality they serve. The senses are differentiated within the brain. As one moves downstream away from primary sensory areas and in towards higher-level "association" areas, there is some evidence that representations remain specific to sensory modalities. In frontal cortex, for example, many researchers believe that there are modality-specific working memory areas. Some areas are dubbed polymodal association areas, but polymodal does not mean amodal. Representations in these areas may coordinate and correlate information from the senses. If we thought in amodal codes, we would expect to find two things. First, we would expect to find large brain regions that are not modality specific. This is not the case. There are large areas associated with emotion, with attention control, and, of course, with action, but little neural real estate looks especially likely to contain vast repertoires of amodal representations. Second, we would expect functional imaging to reveal that cognitive tasks engage these amodal areas, and leave sensory specific areas unperturbed. That is not what is found. The entire brain contributes to cognition, and anterior areas, thought to contain polymodal representations probably play a role of reactivating modality specific regions during cognitive tasks. Contrary to popular belief, frontal cortex is not the locus of thought. Frontal regions maintain, manipulate, reactive, and encode modality specific activations that take place elsewhere in the brain. It looks, in other words, like the brain specializes in perception, perceptual reenactment, and perceptual reasoning.

The second finding that I want to mention comes from psychology. The suggestion that the brain re-engages modality specific regions during the performance of cognitive tasks is supported by psychological research. Consider a study by Pecher et al. (2003). Subjects were given the simple task of confirming that familiar objects have certain features. For example, they were asked "Are blenders loud?" or "Are cranberries tart?" or "Do leaves rustle?" These are cognitive judgments. They tap into our conceptual and collateral knowledge of categories. If thought were couched in an amodal code, we should be able to answer such questions without forming sensory specific mental images. Suppose, however, that we think in modality specific codes. If so, should verify features by forming sensory images. To test between these two possibilities, Pecher et al. availed themselves of a well known fact about perception. When we shift from a task that uses one sense modality to another, there are temporal switching costs. We are faster at shifting from one visual task to another than, say, switching from a visual task to an auditory task. Thus, if modality specific representations are used, there should be switching costs. And this is exactly what the researchers found. For example, subjects who were first asked whether blenders are loud and then asked whether leaves rustle were faster at answering that second question that subjects who were first asked whether cranberries are tart and were then asked if leaves rustle. There is no reason why we should be slowed down if we are using representations in that are amodal.

This study, and others like it, provide prelimary support for the hypothesis that we think by means of modality specific representations. If so, it follows that representational primitives are modality specific. Such experimental findings do not prove that empiricism is right. For that we need more evidence and a more complete understanding of the nature and organization of representations in the brain. But the Pecher et al. study shows that empiricism has some predictive power. Nonempiricist explanations of their findings are likely to be ad hoc. We should, therefore, be receptive to the possibility that empiricist theories of concepts are right. Receptive, that is, unless there are knock-down arguments against empiricism.

2.2 Fodor and Lepore Against Empiricism

Fodor and Lepore reject the empiricist solution to the problem of primitives for a familiar reason. It's just obvious, they think, that many of our concepts cannot be built up from sensory features. To make the point, they give some examples. The concept AUNT and the concept UNCLE clearly have something in common, but that commonality cannot be described in terms of sensory

features. Aunts and uncles have no characteristic appearances. Likewise, there is something shared by Cleopatra and the current President of the United States; they are both politicians. This is not a sensory feature. More strikingly, WATER and ICE are conceptually related, but their commonality is not sensory, even though they are concrete concepts. How on Earth can empiricism cope with these and myriad other concepts that do not seem to decompose into simple percepts?

Before attempting an answer, I want to make a general point. In considering such concepts, philosophers seem to be susceptible to a Grand Illusion. If concepts are not made up of sensory features, then they are made up of amodal features. Perhaps they are words or descriptions in a language of thought. Now, we can ask, how does *this* help to explain our possession of abstract concepts? To merely stipulate that there is a SIBLING symbol or a POLITICIAN symbol doesn't get us very far. The question is, how do these symbols get their meaning? Two answers are possible: they get their meaning by relation to other symbols or by relation to the world. If we go the first route, we will be stuck in a hermeneutic circle. Symbols relating to symbols gives us nothing more than an abstract causal/inferential structure. Labeling the symbols would be a cheat. We can talk only about symbol *n* being linked to symbol *m*. The relations are too abstract to explain meaning. The second route is to explain meaning by relating symbols to the world. This is Fodor's (1990) preferred route. A Mentalese term, such as UNCLE, gets its meaning from the fact that tokens of that term are lawfully caused when we encounter uncles (or, more precisely, UNCLE is under the lawful control of unclehood and that law does not depend on any of the other things that cause UNCLE to be tokened). Now we can ask, How does UNCLE get caused by uncles? For this to occur, we must have a way of recognizing when an uncle is present. We must be able to use our senses to detect the presence of uncles. So, the way a symbol gets meaning, on Fodor's view, presupposes that we have a capacity to identify uncles by means of perception (Prinz 2002). The Grand Illusion is the belief that amodal symbols are somehow able to explain abstract concepts in virtue of being amodal. In reality, amodal symbols may depend on perceptual representations to get their meaning. Amodal symbols are not an advantage when it comes to abstract concepts; they simple delay the inevitable task of explaining abstract concepts in sensory terms.

So how do we recognize uncles and aunts? Here's one technique. Suppose we paint a magenta dot on the navel of every uncle, and a chartreuse dot on the navel of every aunt. We could perceive uncles and aunts by looking at the colors of their navel dots. Our perceptions would be of colored navels,

but these perceptions would be under the causal control of the properties of unclehood and aunthood. Our perceptions of magenta and chartreuse dots would, under these conditions, represent those kinship properties. Of course, we don't do this. We do something else. All societies have kinship terminology. Uncles and aunts are called "uncles" and "aunts." Rather than gaping at midriffs, we are told that Harry is an uncle and Henrietta is an aunt. These words are perceivable, and, more importantly systematic linguistic practices ensure that they are reliable indicators for the presence of unclehood and aunthood.

In addition to these terms, we have a rich kinship vocabulary, and we know the inferential relations between words in that vocabulary. We also can identify culturally specific kinship roles, and we have some core kinship terms, such as "mother," "father," "brother," and "sister," that are learned early on by ostension in societies that have nuclear families. We learn culturally specific kinship roles by observing our immediate family members. When someone is called an "aunt," we can gain further insight into what that means by relating her to mothers and sisters (individuals whose social roles are more fixed). Of course, we also come to learn about biological relatedness, about sex and birth, about Oedipal complexes, adoptive parents, and much else besides. All of these concepts are understood by complexes of words and descriptions (in perceivable public-language symbols), and they are pinned down by a host of helpful images (we can image sex, gestation, and birth, for example). This may all sound simplistic, even obvious. But that's just my point. When we start to probe into how people realy do understand kinship concepts, the puzzle seems to go away. Kinship concepts can be pinned down by inter-related words, recognizable social roles, familiar individuals, and reproductive imagery.

Other abstract concepts can be handled by empiricists as well. How do we know what a politician is? Well, we know a good deal about how the word "politician" is used, and we can point to specific individuals who fall under that term and belong to a complex web of social practices that we can describe, recognize, and conform to. Practices of voting and of obeying authority are symbolically mediated, but perfectly concrete. They are things that we do.

What do water and ice have in common? Well, we have expectations about how one becomes the other, represented using image transformations and embodied in kindergarten-learned skills. We fill the ice tray and, Voila! We put an ice cube in our mouths and, Presto! If we explicitly conceptualize a commonality, that may take the form of words ("H_2O") and chemistry diagrams. These facts are almost too obvious to mention. The crucial point

is that amodal symbols add nothing to this picture. Interrelated images do the trick, and they are needed on any account.

Here someone might protest that I have abandoned the spirit of empiricism. I have invoked representations of words, and, hence conceded that abstract concept must be represented in a non-sensory way. To this I respond by reminding the critic of the dialectic. Fodor and Lepore argue that empiricism is hopeless because we cannot explain concepts by appeal to features represented in sensory specific codes. I have just been showing that this widespread assumption (a dogma of anti-empiricism) is false. Words are sensible. The proposals that I have been sketching are consistent with the assumption that the primitive representations used in thought are modality specific. The knee jerk text book dismissals of empiricism are simply misguided. Once we give up the Grand Illusion, the main objection to empiricism evaporates, and Hume's raft can coast gently along.

2.3 Churchland Against Empiricism

I believe that concept empiricism is entirely defensible (Prinz 2002). Therefore, it is frustrating that the label is used as a pejorative by so many philosophers. Worse still, proving that someone is an empiricist is treated as a *reductio ad absurbum*. Fodor and Lepore use it this way in arguing against his Churchland. I think Churchland should just embrace the label. Instead he goes to great lengths to distinguish his approach from the "antiquated" empiricist picture. Toward this end, he lists five "points of divergence from Hume" (Churchland 1996b: 281). I want to suggest that Churchland's account may be closer to Hume's than he is willing to acknowledge, and where he departs from Hume, he oughtn't.

Churchland's first contrast is that connectionist networks are ampliative in nature. They are not passive slaves to inputs. Rather, they process information in a way that is prejudiced by prior learning. Churchland contrasts this with Hume, and with Fodor's informational semantics. I have two things to say in response. First, neither Hume nor Fodor need to deny the ampliative nature of perception. Once knowledge has been obtained, it can mediate in perceptual recognition and judgment. The conclusions we draw from experience depends on prior knowledge. What looks like a simple rash to me, may look like an early stage of measles to a doctor. Nothing about this requires a departure from Hume's credo that all ideas are copies of impressions. Second, Churchland should not be so quick to dismiss Fodor's informational semantics. As remarked above, Churchland does not have an adequate account of how representations get their intentional content.

He appeals to inferential and motor sequele. As a connectionist, there is nothing to prevent him from saying that the distal causes of internal states contribute to their content. Lacking a good theory of intentionality, he could simply appropriate Fodor's theory. Fodor's theory is an empiricist theory. Fodor himself traces it to Skinner, and similar ideas were proposed by Locke (1690, see Cummins 1991).

Churchland's second point of departure from Hume gets to the heart of empiricism. He says that he is not an empiricist because he does not believe that mental representations are primarily sensory in nature. To illustrate this point, he uses the Cottrell face recognition network, and reminds the reader that the hidden units therein are holons. This example is baffling. The face network takes images as inputs and holons are analyzed as images (albeit ghost-like blurry images). Thus, holons in this network are sensory. They are visual. They are derived from visual inputs, and they are couched in a code that is proprietary to the face network, rather than being couched in a generic amodal code. Churchland's point may be that holons are not simple ideas in Hume's sense. They are not simple. But, as I have already argued, simple ideas must be extracted by the network at some point, if it is to participate in feature-sensitive processing. In any case, the claim that connectionism does not use sensory representations requires further support. It requires an argument to the effect that the networks underlying human cognition are amodal.

Churchland draws another contrast with Hume that is intended to provide such an argument. He says that networks embedded in systems that have entirely different senses can have the same "internal economy." A system that had touch sensors could have a network that is isomorphic with a system that has visual sensors. True enough, but this is no help to Churchland. Why would we say that these networks are representing the same thing? After all, mere isomorphism couldn't be enough for content identity. By freak coincidence, the dendogram showing similarity clusters in NETtalk could be isomorphic with the dendogram for a face recognition network. Those two dendograms could be isomorphic with a dendogram showing distances between Canadian cities (grouped by province). To avoid preposterous alignments, content similarity must depend on similarity of inputs (or, in the case of motor systems, outputs).

Churchland could establish that connectionist nets use amodal codes by showing that networks in sensory pathways feed into one grand network that has activation patterns that are indifferent to the modality of input. This would be an amodal network. A system could certainly be designed this way; no one said amodal codes are impossible. The key question is whether

naturally evolved mammals have brains (or minds) like that. Churchland gives no reason to think so, and, earlier, I tried to give reasons against this amodal architecture. Most networks in the brain are, as a matter of fact, proprietary to specific sensory modalities, and those that are not proprietary may be primarily involved in coordinating activity in those that are (see also Damasio 1989).

Churchland's fourth contrast with Hume (switching his order) is that networks all have distinct representational regimes, rather than shared simple ideas. Similarity is a measured by isomorphism, and isomorphism depends on state space partitions that depend on a network's outputs, as well as its inputs. I have already addressed these points. Networks need identifiable simple ideas at some causally relevant level of processing or analysis. Otherwise, they couldn't be used to explain feature-sensitive processing. The bit about ampliative effects is, once again, a red herring.

Churchland's final point is that empiricists have traditionally been concerned with explaining the meaning of linguistic items and their direct mental analogues. He wants his account to subsume infralingual creatures. I don't see this as a point of contrast. Locke certainly wanted his account to accommodate non-human animals (and not just the talking parrot that he discusses in his chapter on personal identity). Grounding cognition in perception is a way of showing that thinking has a basis in something that's more rudimentary – something that sprouted on the phylogenetic tree long before language came on the scene. I have argued that some concepts may make use of language in an important way (a point missed by Locke and Hume), but I do not think Churchland would disagree. Poodles probably don't have a concept of uncles or politicians. So the question is, Do empiricists allow for *some* concept in infralinguals? The answer is: absolutely.

In drawing this last contrast, Churchland makes a revealing remark. He says that his account does not commit to the existence of beliefs and desires. In that respect, he surely differs from Hume. But faith is propositional attitudes is not essential to empiricism. If Churchland delivers on his promise to develop an alternative theory of thinking, he will not have refuted empiricism; he will have situated an empiricist theory of representation in a novel theory of how mental representations are used. Personally, I agree with Churchland that the constructs of belief and desire are vulnerable to elimination. More exactly, I think they are vulnerable to fractionation. Scientific psychology will divide these concepts into separate processes. Belief will fractionate into such things as semantic and episodic memory, mental models, and lexical networks. Desire will fractionate into reward

registration, various specific motivational states, and somatic markers. It is time to move beyond belief/desire psychology.

What I deny is that fractionation of beliefs and desires would be a departure from concept empiricism. None of the successor constructs that I just mentioned is incompatible with the idea that mental representations are stored records of perceptual states (or combinations thereof). Empiricism must mature as the data come in.

I conclude that Churchland's contrasts with Hume are either inadequately defended or, more often, wholly consistent with the basic tenets of concept empiricism. Rather than resisting the label, Churchland should embrace empiricism. He should climb aboard Hume's raft.

3. CONCLUSION: HOLONS AND HUME

I have argued that state-space semantics is vulnerable to some serious objections. The problems of collateral information and primitive features, advertised by Fodor and Lepore must be taken seriously. Churchland's attempt to address these problems by appeal to the Laakso and Cottrell technique is inadequate. Fodor and Lepore, however, misdiagnose the problem with that attempt. The problem is not that Churchland confuses mind and brain. The real problem is that Churchland desperately needs an account of semantic features, and he shies away from traditional empiricism, which may offer a solution. I have not offered a serious defense of empiricism here, but I gave some reasons for taking the position seriously, and I suggested that the standard objections may be misguided.

Where does all this leave state-space semantics? Once we have an account of primitive features, I don't think the geometrical methods of content individuation will be necessary. I also think that some kind of informational semantics will play an important role in explaining intentional content. The semantic theory that will emerge at the end of the day will have two components: an account of the perceptual primitives that make up our concepts, and a theory of how primitives and complexes get semantically grounded to the world through causal relations. This is just good old fashioned empiricism. Connectionist networks may play a role in this story. Connectionism offers exciting avenues for explaining perceptual recognition. Holons may play a role in this story too. Primitive features may be extracted (or "abstracted") from holonic microfeatures. And beliefs and desires may end up playing no role in the story. So, there may be much to preserve in Churchland's new vision of how the mind works. But, in

adopting brave new ideas, we should not lose sight of good old ideas. We should not abandon Hume's program entirely.

References

Barsalou, L. W. (1999). "Perceptual symbol systems." *Behavioral and Brain Sciences* **22**: 577–609.

Cummins, R. (1991). *Meaning and mental representations* Cambridge, MA, MIT Press.

Churchland, P. M. (1996a). Fodor and Lepore: State space semantics and meaning holism. In R. McCauley (ed.), *The Churchlands and their Critics* (pp. 273–7). Oxford: Blackwell.

Churchland, P. M. (1996b). Second reply to Fodor and Lepore. In R. McCauley (ed.), *The Churchlands and their Critics* (pp. 278–283). Oxford: Blackwell.

Churchland, P. M. (1998). "Conceptual similarity across sensory and neural diversity: The Fodor/Lepore challenge answered." *Journal of Philosophy* **95**: 5–32.

Damasio, A. R. (1989). "Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition." *Cognition* **33**: 25–62.

Fodor, J. (1990). *A theory of concept and other essays*. Cambridge, MA, MIT Press.

Fodor, J. (1998). *Concepts: Where cognitive science went wrong*. Oxford, Oxford University Press.

Fodor, J. & Lepore, E. (1996a). Churchland on state space semantics. In R. McCauley (ed.), *The Churchlands and their Critics* (pp. 145–58). Oxford: Blackwell.

Fodor, J. & Lepore, E. (1996b). Reply to Churcland. In R. McCauley (ed.), *The Churchlands and their Critics* (pp. 159–62). Oxford, Blackwell.

Fodor, J. & Lepore, E. (1999). "All at sea in semantic space: Churchland on meaning similarity." *J. Phil*. **96(8)**: 381–403.

Fodor, J. & Pylyshyn, Z. (1988). "Connectionism and cognitive architecture: A critical analysis." *Cognition* **28**: 3–71.

Garzón, Francisco Calvo (2003). "Connectionist semantics and the collateral information challenge." *Mind & Language* **18**: 77–94.

Laakso, A. & Cottrell, G. (1998). How can I know what you think?: Assessing representational similarity in neural systems. In M. A. Gernsbacher and S. Deny (eds.), *Proceedings of the 20th Annual Cognitive Science Conference*. Mahwah, NJ, Lawrence Erlbaum.

Laakso, A. & Cottrell, G. (2000). "Content and cluster analysis: Assessing representational similarity in neural systems." *Philosophical Psychology* **13**: 47–76.

Pecher, D., Zeelenberg, R., & Barsalou, L. W. (2003). "Verifying properties from different modalities for concepts produces switching costs." *Psychological Science* **14**: 119–124.

Prinz, J. J. (2002). *Furnishing the mind: Concepts and their perceptual basis*. Cambridge, MA, MIT Press.

Quine, W. V. (1953). "Two dogmas of empiricism." In *From a logical point of view* (pp. 20–46). Cambridge, MA, Harvard University Press.

Sejnowski, T. J. & Rosenberg, C. R. (1988). "NETtalk: A parallel network that learns to read aloud." In J. A. Anderson and E. Rosenfeld (eds.) *Neurocomputing: Foundations of research* (pp. 661–72). Cambridge, MA, MIT Press.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist networks. *Artificial Intelligence* **46**: 59–216.

Tiffany, E. (1999). Semantics San Diego Style. *Journal of Philosophy* **96**: 416–29.

5 Churchland on Connectionism

AARRE LAAKSO AND GARRISON W. COTTRELL

INTRODUCTION

Paul Churchland cemented his appointment as Ambassador of Connectionism to Philosophy with the 1986 publication of his paper "Some reductive strategies in cognitive neurobiology." However, as Churchland tells the story in the preface to his collection of papers, *A Neurocomputational Perspective*, his relationship with connectionism began three years earlier, when he became acquainted with the model of the cerebellum put forward by Andras Pellionisz and Rodolfo Llinas (1979). The work of Pellionisz and Llinas foreshadows many of the arguments that Churchland makes. They argue that functions of the brain are represented in multidimensional spaces, that neural networks should therefore be treated as "geometrical objects" (323), and that "the internal language of the brain is vectorial" (330). The Pellionisz and Llinas paper also includes an argument for the superiority of neural network organization over von Neumann computer organization on the grounds that the network is more reliable and resistant to damage, a theme to which Churchland often returns.

Over the years, Churchland has applied connectionism to several areas of philosophy, notably: philosophy of mind, epistemology, philosophy of science, and ethics. Churchland's arguments in these areas have a common structure. First, he shows that the predominant positions in the field are (a) based on an assumption that the fundamental objects of study are propositions and logical inferences, and (b) have significant internal difficulties largely attributable to that assumption. Second, he presents a re-construal of the field based on connectionism, giving a "neurocomputational perspective" on the fundamental issues in the field. Finally, he argues that his connectionist alternative fares better than the predominant position on a number of criteria, and explores its further consequences. This is certainly not a formula, since Churchland always considers the particulars of each field, but it is a pattern.

In this paper, we explicate these arguments in a little more detail, and try to give some indication of what we think Churchland has right and what he has wrong. The explication is brief because we are attempting to cover, in just a few pages, topics about which Churchland has written hundreds of pages. The idea is simply to give enough context that the reader may understand the gist of Churchland's arguments and the role that certain problematic claims play in his larger program.

In evaluating Churchland's relation to connectionism, we attempt to skirt the more obviously and centrally *philosophical* issues and concentrate more on the *empirical* issues. Hence, we focus more on what Churchland claims about connectionism (and, in part, about cognitive neuroscience and cognitive psychology) than on what he claims about philosophy of mind, epistemology, philosophy of science, or ethics. This is partly an attempt to balance the perspectives taken on Churchland's work in this volume, and partly a broader attempt to balance the perspectives taken on Churchland's work in the literature as a whole. The distinction between philosophical and empirical issues is indeterminate, however. Churchland is a naturalist – he believes that philosophy and science are continuous – and this is evident in his writing. A critical part of his thesis is that certain claims about connectionism (about its properties, its uses and applications, and its consequences) are in themselves claims about philosophical issues. Moreover, we agree with him – and with Quine (1951) – that there is no fundamental distinction between philosophy and science, that rather there is a continuum of issues from those most peripheral and amenable to change to those most central and resistant to change. So, while we attempt to remain near the edges of the web of belief, we sometimes inevitably slip closer to the center.

We do not cover the basic principles of connectionist networks, such as what a hidden unit is, or what backpropagation means. These topics are dealt with in depth in many other works – see, for example, Ballard (1999), Bishop (1995), Hertz, Krogh, and Palmer (1991) – and Churchland himself often gives explanations that are engaging and clear enough for the layperson. Instead, we save our comments for those cases where we feel that Churchland has the facts wrong about connectionism, or has treated it incompletely.

THE PHILOSOPHICAL CONTEXT

In this section, we discuss Churchland's use of connectionism to ground novel approaches to the philosophy of mind, epistemology, philosophy of science, and ethics. In each case, we first present the "classical" or

"sentential" theory, then Churchland's connectionist alternative. We make an effort to highlight those features of Churchland's argument that we will return to later for a more critical view. At the same time, we try not to be too judgmental at this stage. The goal is to get Churchland's ideas on the table. We then "dig in" to a few choice bits in later sections.

Philosophy of Mind

Churchland contrasts his novel, connectionist philosophy of mind with views that he variously calls "orthodox," "classic," "propositional," and "sentential." Churchland covers a variety of different views with this umbrella, but the most central of them is Fodor's (1975) "language of thought" hypothesis, viz., that thoughts and beliefs are sentences in an internal "language of thought"; thoughts are occurrent sentences (those being actively considered), while beliefs are stored sentences. A second, related, view is that thinking (which Churchland sometimes refers to more generally as "cognitive activity") is performed by updating the occurrent belief sentences by a process of logical inference. A third is that learning is the rule-governed revision of a set of stored (belief) sentences. We will call this view "Propositionalism."

Churchland achieves a certain generality in his arguments by limiting his discussion of the opposing views to these general claims, which (at least directly) imply nothing about the computational architecture of a thinking machine. However, one of Propositionalism's strengths is that it is well-suited for a computational implementation. According to the view that we will call "Computationalism," the sentences that express particular thoughts are actually sequences of tokens in the registers of a von Neumann computer. Likewise, the sentences that express beliefs are sequences of tokens in the memory of such a computer. Thinking and learning are computational processes (programs) that operate on thought and belief tokens in virtue of their form and syntax. The programs implement algorithms that transform thoughts and beliefs according to their form and syntax, while preserving their semantics (meaning and truth values).

For Propositionalists, the affinity of Computationalism and Propositionalism is one of the principle virtues of Propositionalism. It explains how it is possible for machines (including, potentially, biological machines such as human beings) to think. To account for thinking, we need not suppose that there is some sort of nonmaterial substance, or that thought is identical to behavior, or to neural activity, or any of a thousand other problematic things. Rather, thinking is running a program, and we all understand more or less what that is.

For Churchland, on the other hand, the affinity of Computationalism and Propositionalism is one of the principle vices of Propositionalism. The brain is organized very differently from a von Neumann machine. Computationalism is the best implementation account that Propositionalism has to offer, and Propositionalism is therefore completely disconnected from any detailed, neuroscientific account of how the brain actually functions. While the mind might be organized like a computer, the brain is not. There have been attempts to show how the compositional structures essential to symbolic computation might be implemented in a more biologically plausible architecture. Some notable examples are: Cottrell's (1985) implementation of Reiter's default logic for inheritance hierarchies with exceptions in a spreading activation network; Touretzky's (1990) proposal for a method to implement the fundamental Lisp data structure in Boltzmann machines; Smolensky's (1990) proposal for implementing compositional representations as tensor products; Pollack's (1990) recursive auto-associative memory architecture; and Elman's (1991) demonstration that recurrent networks can process complex embeddings in grammatical structure. These and other examples of "implementational connectionism" (e.g., Plate 1995, Derthick 1987, Touretzky & Hinton 1985, Ballard 1986) can be viewed as attempts to demonstrate that an essentially Computationalist model of the mind can be implemented in a connectionist network.

Churchland offers a more radical view: connectionism as an "alternative cognitive paradigm" (Churchland 1989e: xiv), not merely a biologically plausible implementation mechanism for a Computationalist model of the mind but a truly novel model of the mind itself. Where Computationalism takes the computational architecture of cognition to be the von Neumann computer, Churchland takes it to be a connectionist network. The claims of his view, which we will call "Connectionism,"[1] fall out of this fundamental change. Where the Computationalist takes thoughts to be instantiated as sequences of tokens in the central processor of a computer, the Connectionist takes thoughts to be instantiated as patterns of activation in the units of a neural network. The Computationalist takes thinking to be instantiated as the transformation of sets of thought tokens according to a program that is sensitive to their structure, whereas the Connectionist takes thinking to be instantiated as the transformation of patterns of activation in the units of a neural network according to the weighted connectivity between them.

The analogy between Computationalism and Connectionism is somewhat more complicated for belief and learning. We have seen that the Computationalist takes beliefs to be instantiated as sequences of tokens

in the memory of a computer. Some Connectionists take beliefs to be instantiated in the weighted patterns of connectivity between the units in a neural network. Churchland, for example, embraces this view of Connectionism in "On the nature of theories" (1990). Other Connectionists take beliefs to correspond to the partitions of activation patterns that the connection weights determine, or, in recurrent networks, the stable patterns of activation – attractors – that are determined by the weights. Churchland reconsiders the question and adopts this latter view in "Learning and conceptual change" (1989a: 234–234). We have also defended the "partitioning of activation space" interpretation of belief in previous work (Laakso & Cottrell 2000). The interpretation of learning in Connectionism also depends on the position one takes with respect to belief. As we have seen, the Computationalist takes learning to be instantiated as the transformation of sets of belief tokens according to a program that is sensitive to their structure. The Connectionist sympathetic to the beliefs-are-weights view takes learning to be instantiated as the updating of the weighted patterns of connectivity between the units in a neural network according to an algorithm that is sensitive to their values. The Connectionist sympathetic to the beliefs-are-partitions view takes learning to be instantiated as the transformation of the partitions (or attractors) in activation space according to an algorithm that is sensitive to the weights that determine those partitions.

There is also an analogy between Propositionalism and what we will call "Vectorialism." Vectorialism is to Connectionism what Propositionalism is to Computationalism. Propositionalism asserts that thoughts are occurrent sentences in an internal "language of thought," whereas Vectorialism asserts that thoughts are vectors in an internal neural activation coding. Propositionalism asserts that beliefs are stored sentences in the "language of thought," whereas Vectorialism asserts that beliefs are matrices of connection weights – or equivalence classes of activation vectors, depending on one's view of what constitutes belief in a connectionist network – in an internal neural connectivity coding. Propositionalism asserts that thinking is logical inference, whereas Vectorialism asserts that thinking is the changing of activation vectors by matrix multiplication and nonlinear transformations. Propositionalism asserts that learning is the rule-governed revision of a set of beliefs, whereas Vectorialism asserts that learning is the mathematically governed revision of a matrix of connection weights – or of a set of equivalence classes of activation vectors, again depending on one's view of what constitutes belief in a connectionist network. Vectorialism can also be stated in an alternative geometric and kinematic language, one that Churchland sometimes uses. That is, thoughts (activation vectors) may also

be conceptualized as points in activation space; beliefs (weight matrices) may also be conceptualized as points in weight space; thinking (updating activation vectors) may also be conceptualized as motion in activation space; and learning (updating weight matrices) may also be conceptualized as motion in weight space.

We have coined the term Vectorialism because there is no widely used term for the view we have just described. It is possible to be a Connectionist without being a Vectorialist, as the examples of "implementational connectionism" that we mentioned above demonstrate. Churchland sometimes uses the term "state-space semantics" to encompass this and other parts of his view, and we have followed him in previous work (Laakso & Cottrell 2000). However, the term "semantics" arguably does not apply at this level – Vectorialism is a claim about the form of mental representations, not about their contents. Prinz (this volume) is one of many who have pointed out that "state-space semantics" is a misnomer in the absence of an adequate theory of how vectors in state space get their contents. Hence the need to use another term. Whether Churchland offers an adequate theory of content for state-space semantics is a distinctly philosophical question that we do not address here. For the same reason, we do not consider here the broader question of whether it is possible to offer an adequate theory of content for state space semantics independent of Churchland (but see Cottrell, Bartell, & Haupt 1990, for one example).

These features of the different accounts are presented in Table 5.1, which thus provides a brief summary of Churchland's position and distinguishes it from the Computationalist orthodoxy.

Table 5.1 *Comparison of Propositionalism, Computationalism, Vectorialism, and Connectionism as Approaches to the Philosophy of Mind*

	Orthodox View		*Churchland's View*	
	Propositionalism	*Computationalism*	*Vectorialism*	*Connectionism*
Thoughts	sentences	symbolic tokens	numeric vectors	activations
Beliefs	sentences	symbolic tokens	numeric matrices/ classes of vectors	connectivity weights/partitions
Thinking	logical inference	algorithmic updating	vector transformations	changing activations
Learning	rule-governed revision	algorithmic updating	matrix transformations/ class changes	weight changes/ partition changes

One of the principal virtues that Churchland sees in Connectionism is its biological plausibility. It seems natural to think of units in a connectionist network as simplistic models of neurons, and connections as simplistic models of synapses. As Churchland writes, Connectionism "puts cognitive theory firmly in contact with neurobiology, which adds a very strong set of constraints on the former, to its substantial long-term advantage" (1990: 98). In a future section, we consider just how biologically plausible connectionism really is, but for now it is safe to say that intuitively it seems quite plausible, certainly much more plausible than the declarative and procedural mechanisms characteristic of "good old fashioned" artificial intelligence (GOFAI).

Churchland sees many virtues in Connectionism besides biological plausibility. One is its natural account of categorization. He notes that connectionist networks such as the rocks-from-mines detector (Gorman & Sejnowski 1988) and NETtalk (Sejnowski & Rosenberg 1987) develop rich, structured internal representations that both enable them to exhibit impressive behavior and correspond to real structure in their input. Churchland often explains these categorization feats in terms of "prototype representations" in the hidden unit activation space.

Another virtue that Churchland sees in Connectionism is its natural account of similarity as corresponding to proximity in state space. He writes, with evident gusto, "a state-space representation embodies the *metrical* relations between distinct possible positions within it, and thus embodies the representation of *similarity* relations between distinct items thus represented" (1986: 299, emphasis in original). He uses this feature of Connectionism to give an account of qualia, considering the example of color in depth, but also with reference to taste, olfaction and audition (1989d: 221).

Speed is another virtue that Churchland sees in Connectionism. Connectionist networks can operate very quickly, because of their massive parallelism. The declarative and procedural programs of GOFAI, on the other hand, can be very slow. Moreover, they are usually not amenable to parallelization. (This is of course part and parcel of their biological implausibility.) When such programs do achieve real-time speeds, it is generally in virtue of exploiting the remarkable speed of modern computing hardware. Neurons do not operate nearly as quickly as transistors, so the rapidity of cognition is achieved by parallelism. Connectionism models this computational strategy more closely than GOFAI.

Another virtue of Connectionism is "functional persistence" (as Churchland usually calls it) or "graceful degradation" (a more common term that Churchland also sometimes uses). The brain is remarkably resilient

to trauma, including injury and disease. There are limits, of course, but the anecdotes of famous clinical cases, like Phineas Gage for example, are remarkable not only because of the highly specific and unusual deficits that they document but also because of the remarkable amount of function that is preserved despite large-scale trauma. Connectionist networks can exhibit a similar resilience: their function is often largely preserved despite a simulated "loss" of some of their units or connections.

Churchland also praises Connectionism for being applicable to nonhuman animals. This is a consequence of its biological plausibility; since connectionist networks are plausible models of the operation of biological neural networks, and since the fundamental computational principles in biological neural networks are the same across species, Connectionism not only explains human cognition, but also explains cognition in other species. This is a claim that Propositionalism cannot make. As implausible as it is that human beings think by manipulating sentential representations in an internal language of thought, it is even more implausible that nonhumans do so. For some Propositionalists, this is a virtue of their account, because it provides a theory of cognition on which thought is uniquely human (see, for example, Bickerton 1995). For most cognitive scientists, the notion is ludicrous.

Epistemology

Although Churchland mentions epistemology frequently, it is almost always in the context of a broader discussion of either philosophy of science or philosophy of mind. For Churchland, epistemology is essentially a bridge between philosophy of mind and philosophy of science. That is, Churchland's attack on traditional epistemological theories follows immediately from his views on the philosophy of mind; and his views on the philosophy of science are, in turn, grounded in his epistemology. Hence, it is possible to summarize Churchland's Connectionist epistemology rather quickly.

Recall that, on Churchland's Connectionist philosophy of mind, beliefs are not sentences in an internal language of thought but vectors in a high-dimensional connectionist weight space. It follows immediately that *knowledge* is not a set of stored sentences (that happen to be true and justified, or something to that effect), but rather a set of stored connection weights. Similarly, on Churchland's Connectionist philosophy of mind, learning is not a process of rule-governed updating of stored belief-sentences, but a process of mathematically governed updating of stored belief-weights.

Table 5.2 *Comparison of Deductivism and Connectionism as Approaches to the Philosophy of Science*

	Deductivism (Orthodox View)	*Connectionism (Churchland's View)*
Knowledge	sets of sentences	prototypes
Learning	logical inference	changing weighted connectivity
Theories	sets of sentences	prototypes
Explanatory Understanding	logical inference	categorization

Again, it follows immediately that knowledge is not acquired by the rule-governed updating of stored belief-sentences, but by the mathematically-governed updating of stored belief-weights.

Philosophy of Science

As he did in the domain of philosophy of mind, Churchland also contrasts his novel, Connectionist epistemology and philosophy of science with views that he variously calls "orthodox," "classic," "propositional," and "sentential." This is another big umbrella, but the most important of the theories covered by it is the deductive-nomological or hypothetico-deductive view, according to which (a) a theory is a set of propositions, and (b) scientific inference and understanding proceed by logical inference. For brevity, we will refer to this view as Deductivism (see Table 5.2).

Deductivism has a number of well-known logical weaknesses. Among them are the paradoxes of confirmation, the problem of explanatory asymmetry, the problem of irrelevant explanation, and the problem of accidental universals. Specific versions of Deductivism also have certain logical problems, such as the indeterminacy of falsification on Popperian theories and the fact that laws were assigned negligible credibility on Carnapian accounts. Churchland discusses these and other problems with Deductivism in detail (1989d, 1990), so we will not dwell on them here.

Churchland's criticism of Deductivism focuses on its empirical implausibility, above and beyond its logical problems. One of the most important empirical issues with Deductivism is timing. People often come to understand an explanation in a very rapid flash of insight. The nearly instantaneous speed of learning and explanatory understanding seems inconsistent with the hypothesis that explanatory understanding is the outcome

of a lengthy process of logical inference. One part of the inconsistency stems from the fact that, on a Deductivist account, grasping a new concept requires first looking up the *relevant* laws or facts. The relevant laws or facts are presumably discovered by some sort of a search, and searches are notoriously slow. The second part of the inconsistency stems from the fact that, even once the relevant basic laws have been retrieved, the cognizer must then deduce the appropriate conclusion. Logical inference is also a computationally intensive operation, one that frequently requires a great deal of backtracking. In fact, logical inference can itself be viewed as a kind of search. So, on the Deductivist account, understanding and learning require *two* searches: one to locate the relevant premises in the space of all known facts, and another to locate the relevant deduction in the space of all possible inferences from those facts. This makes a mystery of our frequent experience of learning and explanatory understanding as rapid and effortless.

Another empirical issue with Deductivism is that the laws and inferences so painstakingly (and yet so rapidly) found are, often, completely inaccessible to the cognizer. People are generally unable to articulate the laws that underlie explanations of phenomena that they appear to understand. They also are generally unable to perform or recite logical inference to anywhere near the degree of rigor and completeness that Deductivism requires. Nonhuman animals also appear to be capable of some forms of causal understanding, but are presumably incapable of storing propositional representations of laws and performing logical inference on them, let alone articulating the laws and the inferences.

The same is true of young infants, and this gives Deductivism a kind of bootstrapping problem. If learning and understanding are characterized by applying the rules of logical inference to propositional premises, and if young infants can neither store propositional premises nor use the rules of logical inference, then how do they *learn* to do so? Evidently, there must be some other account of learning or development that explains our coming to have the abilities that Deductivism requires. Deductivism, however, gives no clues as to what the other account might be. Even if it did, the idea that there should be two different kinds of learning and understanding (a Deductivist account for adults and a – so to speak – *pre*-Deductivist account for infants and perhaps nonhuman animals) seems inelegant at best.

A final empirical issue with Deductivism is that it provides no account of learning or understanding skills, as opposed to facts. However, knowing-*how* is as much a part of our cognitive armory as knowing-*that*. They are, in fact, interrelated in complex ways, as shown by many studies of context-dependent learning. An explanation of skill learning is particularly

important for the philosophy of science in light of Kuhnian observations that implicit knowledge of appropriate practice is an important part of science (Kuhn 1962). While Kuhn may have overstated the importance of skills, it is now widely acknowledged that some part of scientific understanding consists of acquiring appropriate skills.

Some of the most significant problems with Deductivism are neither logical nor psychological per se, but normative. One of the goals of an account of explanation is to determine when changes to a theory are justified; similarly, one of the goals of an account of knowledge is to determine when learning produces *justified* beliefs. However, Deductivism does not meet these criteria.

For one thing, Deductivism cannot justify massive conceptual change. According to Deductivism, all explanation occurs within a framework limited by basic laws and the rules of inference. The laws of inference justify drawing novel conclusions from the basic laws, but they do not warrant changes to the laws themselves. However, fundamental shifts in the basic explanatory axioms often accompany major advances in explanatory understanding (in science) and learning (in individuals).

Nearly everyone who cares about science agrees that scientific theories should be "simple" and "elegant," but almost no one agrees about what those terms really mean or *why* they are important. While Deductivism can perhaps give an account of what "simplicity" means, it does not explain why it is important in a scientific theory. A simple definition of simplicity in Deductivist terms would be the total number of propositions that are required to state the laws governing some field; a slightly more sophisticated view might consider the total number of terms and logical operators in the laws. Regardless, Deductivism does not provide a means for justifying claims that one theory is superior to another on the grounds of simplicity; it does not explain why simplicity is important.

A corollary of the problem about justifying massive conceptual change is that Deductivism cannot give a realist account of scientific progress. Formally speaking, false premises can form just as good a basis for inference as true ones – no amount of inference alone can distinguish false premises from true. (The same is not true for inconsistent premises, but consistency is a very weak normative standard.) However, Deductivism offers no grounds for justifying one set of laws (premises) over another above and beyond their capacity to generate (by logical inference) statements that are true by observation. It was just this property that led us to say that Deductivism provides no means for justifying massive conceptual change, that is, no means for justifying revision of the premises that serve as laws. In much the same way, Deductivism provides in itself no grounds for preferring one

set of fundamental laws (premises) over another. As far as Deductivism is concerned, a false set of premises that is consistent with observation is just as good as a true set of premises. Deductivism alone provides no grounds for preferring true theories over false ones.

Much as Churchland offered Connectionism as an "alternative cognitive paradigm" in the philosophy of mind, so he offers Connectionism as (what we might call) an alternative explanatory paradigm in the philosophy of science. The idea is that explanatory understanding should be thought of not as a product of arriving at a new logical inference but as a product of learning a new category – that a person's grasping a scientific explanation can be modeled by a connectionist network categorizing its input or, equivalently (as Churchland sees it) activating a prototype vector. Of course, coming to understand a scientific theory is *more* than just making a category judgment. It is, among other things, learning to understand a wide variety of things in a certain way and coming to see commonalities among those things, including ones you have never seen before, that you would not otherwise have grasped. In learning a new category, however, a connectionist network does more than simply label the things that it has already been exposed to; it develops an internal representation that can be used to classify new things it has never seen before, and that can potentially be used in other ways in other sorts of computations. Connectionist representations "amplify" knowledge by supporting generalization and complex, context-sensitive processing (1989d: 212).

Churchland argues that Connectionism in the philosophy of science overcomes many of the problems with Deductivism. We have seen that Deductivism offers no explanation of why simplicity is an explanatory virtue. Connectionism, by contrast, has a natural account of why explanatory simplicity is a virtue: an overly complex connectionist network (one with many more hidden units than are required to categorize inputs effectively) will "memorize" the mapping between its inputs and outputs, and fail to generalize to novel inputs. A sufficiently simple connectionist network (one with just enough hidden units to categorize inputs effectively) will achieve both acceptable performance on known inputs *and* effective generalization to novel inputs. An overly simple connectionist network (one with too few hidden units) will be unable to learn to categorize its inputs effectively. Hence, Connectionism can explain why simplicity is a virtue in a scientific explanation: it allows for better generalization to future observations. Connectionism can also explain why too much simplicity is undesirable: there is a natural tradeoff between accurately describing known observations and accurately predicting new observations.

Connectionism also applies to many more types of explanation than Deductivism. We have seen that the Deductivist account of explanatory understanding does not fit scientific (causal) explanations particularly well, but there are many other types of explanations that it does not fit at all. Deductivism offers no account of inductive explanation, for example, or of moral, ethical, or legal explanation. Connectionism, on the other hand, provides a very general account of explanation as a process of concept formation, and therefore applies just as well to these other sorts of explanations as it does to scientific explanation.

A Connectionist account of explanation has other virtues as well. It accounts for our nearly instantaneous grasp of new explanations by the rapidity of parallel processing. It explains our inability to articulate laws or appreciate extended deductive arguments (because we are not using them). It also avoids many technical difficulties with the Deductivist account of scientific explanation – such as the problems of explanatory asymmetry, irrelevant explanations, and accidental universals – which have puzzled philosophers of science for decades.

Churchland admits that his Connectionist account of explanatory understanding does not provide a full account of what explanation itself means. That, however, is not his goal. For Churchland, the challenging question is how cognitive beings come to understand scientific explanations, not what explanations "really are" in some metaphysical sense.

Churchland also draws some broader morals from his Connectionist account of explanatory understanding. He claims that viewing explanatory understanding as vector processing rules out the possibility of finding unassailable "observational" foundations on which to ground the rest of science: all observation, indeed, all perception, is conceptual in the sense that it involves the same sort of vector processing operations. There is no "raw input" to the nervous system that has not been transduced by some part of the sensory system, which is a neural network. This also explains the remarkable plasticity of human beings and cognizers in general – because they are neural networks, they can adapt and change the very means by which they conceive things.

Ethics

Churchland also endeavors to draw *moral* conclusions from Connectionism, specifically to use Connectionism to ground an ethics that neither dismisses moral "knowledge" as bias nor grounds it in abstract rationality (1989b, 1995). Conceptually, social and moral knowledge consists in knowing how

to behave in social situations and how to react to moral dilemmas. It develops by learning to categorize social situations and moral questions appropriately using the pattern classification abilities of a connectionist network. Training consists in coming to react appropriately to social situations (to exhibit socially acceptable, if not advantageous, behaviors), according to the lights of the society in which one grows up. This is not merely becoming socialized to the currently prevailing moral platitudes, because there is room for disagreement (activation of different moral categories in different individuals) and for improvement over time (not only on an individual basis but also on a societal basis, as laws are codified and so on).

As we might expect, Churchland contrasts his Connectionist position with an orthodoxy that explains moral knowledge in terms of a set of sentential rules. The major traditions in ethics may be distinguished by the nature of the rules that they posit and the sources of moral authority that they acknowledge. All such traditions both prescribe and proscribe behavior according to some set of laws. On Churchland's view, by contrast, moral behavior is not prescribed by a set of laws but is is caused by a set of prototypes of "good" and "bad" behavior. Of course, to represent it as a single binary opposition is to oversimplify it drastically. Still, the point remains: for Churchland, ethical and social guidelines are prototypes, not rules. For Churchland, moral disagreements are typically not disagreements over what set of moral rules to follow, but rather, over which moral prototype most closely matches the present situation.

SOME EMPIRICAL ISSUES

Having laid out the basic philosophical context in which Churchland positions his work and explained the main uses to which he puts Connectionism, we now turn to some empirical issues raised by Churchland's claims. We begin with his claim that Connectionism has the virtue of providing a natural account of semantic similarity, in terms of proximity in activation space. In the following section, we consider Churchland's identification of volumes in the hidden-unit activation space of connectionist networks with "prototypes." Finally, we consider whether connectionism really is as biologically plausible as Churchland claims it is.

Similarity

As we noted in the previous section, one of the principle virtues that Churchland sees in Connectionism is a natural account of similarity. On

Churchland's account, perceptual and conceptual similarity is distance between activation vectors in a connectionist network. To determine how similar A is to B, we measure the hidden unit activations used to represent A and the hidden unit activations used to represent B, and then we calculate the distance between them.

A natural first question about this approach is: distance according to what metric? There are many ways of measuring distance, and even more ways of measuring dissimilarity of vectors. The "standard" Euclidean distance between two n-dimensional vectors $\vec{x}, \vec{y} \in \mathbb{R}^n$:

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{1}$$

is only one of a family of norms known as the Minkowski metrics:

$$d = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} \tag{2}$$

for $p = 1, 2, \ldots$, each one of which defines a different possible measure of dissimilarity. Note furthermore that for the special case where $p = \infty$:

$$\lim_{p \to \infty} \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} = \max |x_i - y_i| \tag{3}$$

which is (qualitatively) yet another measure of dissimilarity.

Each one of these satisfies the mathematical definition of a *metric*, which is a general formalization of what we intuitively call "distance." Specifically, a metric must satisfy the triangle inequality (the distance from X to Y to Z is never shorter than from X directly to Z), be symmetric (the distance between X and Y must be the same as the distance between Y and X), be nonnegative, and satisfy the identity of indiscernibles (if the distance between X and Y is 0, then X and Y must be the same) and its converse (the distance between X and itself must be 0). The latter three conditions (nonnegativity, the identity of indiscernibles, and its converse) are sometimes together called *minimality*.

Besides the common examples we have given in Equations (1)–(3), there are many other less common metrics that could be defined as well. Consider, for example, the trivial example where we define distance as 0 for identical points and 1 otherwise. We know of no *a priori* reason to prefer any one of these metrics over any other for the purpose of measuring representational

similarity. To be empirically adequate, the choice of metric would need to be based on psychological considerations, that is, the results of experiments probing the properties of the cognitive similarity metric.

Those experiments have been done, and it turns out that human semantic similarity judgments do not satisfy the conditions on *any* metric meeting the definition above (Tversky 1977, Tversky & Gati 1978). For one thing, semantic similarity is not symmetric: human subjects reliably judge less prominent things to be more similar to more prominent ones than the reverse (e.g., North Korea is more similar to China than China is to North Korea). People judge their friends to be more similar to themselves than they are to their friends (Holyoak & Gordon 1983). It is also possible to find exceptions to minimality. For example, subjects find that the letter *S* is more similar to itself than the letter *W* is to itself, judging by reaction time in a same-different task (Podgorny & Garner 1979). Subjects also find the letter *M* to be more similar to the letter *H* than it is to itself, judging by inter-letter confusions in a recognition task (Gilmore, Hersh, Caramazza, & Griffin 1979).

Semantic similarity also violates the triangle inequality. An apple is similar to a banana (their "similarity distance" is short, because they are both edible fruits), and a banana is similar to a boomerang (their "similarity distance" is also short, this time because they have similar shapes). Hence, by the triangle inequality, the "similarity distance" between an apple and a boomerang (the "direct route" in this case) should also be short – less than the sum of the distances between apple and banana, on the one hand, and banana and boomerang, on the other. However, the "similarity distance" between an apple and a boomerang is quite large, because they have very little in common. In human subjects, similarity is always judged "with respect to" something – apples are similar to bananas with respect to edibility but not shape, and bananas are similar to boomerangs with respect to shape but not edibility. Humans are able to adjust the features on which they base their similarity judgments depending on the context. Equating similarity with simple distance between activation vectors in a connectionist network affords no analogous ability for adjusting the relative salience of features depending on context: the distance just is what it is.

There is also a question as to which activation vectors should be included in assessing similarity. There are often three levels of activations in a connectionist network (inputs, hidden units, and outputs), and similarity may be assessed at any of these levels, or any combination of them, including all of them simultaneously. The question is even more acute for biological neural networks, which have many layers of processing. Churchland

Table 5.3 *Activations of All Units in a Hypothetical XOR Network*

Input 1	*Input 2*	*Hidden 1 (OR)*	*Hidden 2 (AND)*	*Output (XOR)*
0	0	0	0	0
0	1	1	0	1
1	0	1	0	1
1	1	1	1	0

often writes as though the relevant activations are *all* of the activations in the network. This can lead to some counterintuitive results. Consider, for example, a feedforward network that computes the Boolean *XOR* function by combining one hidden unit that computes *OR* and one hidden unit that computes *AND*. The activations of the units in this hypothetical example are shown in Table 5.3. The Hamming distances (the sums of the number of different bits in each vector – a metric based on the Minkowski 1-norm described by Equation (2) above, for the case where $p = 1$) between the activations of *all* of the units for each pair of inputs is shown in Table 5.4. Note in Table 5.4 that the network activations for input pattern (0, 1) and the network activations for input pattern (1, 0) are Hamming-distance 2 apart, whereas the network activations for input pattern (0, 0) and the network activations for input pattern (1, 1) are Hamming-distance 4 apart. However, the input patterns (0, 1) and (1, 0) are the same Hamming-distance from each other, as are the input patterns (0, 0) and (1, 1) (i.e., 2 in every case), and the output patterns for (0, 1) and (1, 0) are identical (both 1) as are the output patterns for (0, 0) and (1, 1) (both 0). So, in this case, two pairs of patterns that are equally dissimilar at the inputs and equally similar at the outputs have different overall similarities. This suggests that it is important to consider the layer at which the patterns are compared. If we want to use distance between activation patterns as a similarity metric, then we need to

Table 5.4 *Hamming Distances Between Activations of All Units for All Possible Pairs of Input Patterns in a Hypothetical XOR Network*

	Input Pattern B			
Input Pattern A	*(0, 0)*	*(0, 1)*	*(1, 0)*	*(1, 1)*
(0, 0)	0	3	3	4
(0, 1)	3	0	2	3
(1, 0)	3	2	0	3
(1, 1)	4	3	3	0

specify which patterns are to be compared; comparing all of them is likely to lead to uninformative results.

There are also differences that do not strictly violate the metric axioms but nevertheless conflict with the properties of common metrics like Euclidean distance. For example, most metrics strictly limit the number of points that can have a common nearest neighbor, whereas human similarity judgments often rate many items as most similar to a single item (Tversky & Hutchinson 1986). In Euclidean space, the maximum number of points that can have the same nearest neighbor[2] i in $1D$ is 2 (a third point will either have one of the other two as its nearest neighbor, if it falls outside them on the line, or be the nearest neighbor of one of the other two, if it falls inside them on the line). If we disallow ties, then the maximum number of points that can have the same nearest neighbor i in $2D$ is 5. (The vertices of a regular pentagon with i at the center will all be closer to i than to each other, whereas some of the vertices of a hexagon with i at the center will be at least as close to each other as to i.) In human similarity judgments, by contrast, many items often have the same nearest neighbor (most similar item); in particular, people often associate all (or nearly all) exemplars of a basic-level category most closely with the category itself (Tversky & Hutchinson 1986). For example, in data reported by Rosch and Mervis (1975), subjects rated the category name "fruit" as most related to all but 2 of 20 common instances of fruit (the exceptions being "lemon," which was more related to "orange," and "date," which was more related to "olive"). The fact that human similarity judgments exhibit this sort of dense nearest-neighbor structure, which metric models of similarity cannot capture, suggests that the metric models are incorrect or, at the least, incomplete.

There are nonmetric theories of semantic similarity, as well as more sophisticated metric theories. The non-metric theories include models based on matching features, such as the "contrast model" proposed by Tversky (1977), models based on aligning representations, such as Goldstone's SIAM (1994), and models based on transforming representations, such as that recently advocated by Hahn, Chater, and Richardson (2002). This is not to say that we should give up entirely on accounting for semantic similarity in terms of distance. There are several proposals on offer for basing an account of semantic similarity on distance *with some additional apparatus*, such as spatial density (Krumhansl 1978) or attentional bias (Nosofsky 1991). Hence, it may be possible to defend the claim that semantic similarity corresponds to proximity in activation space in some sense. However, doing so requires some account of how proximity is augmented

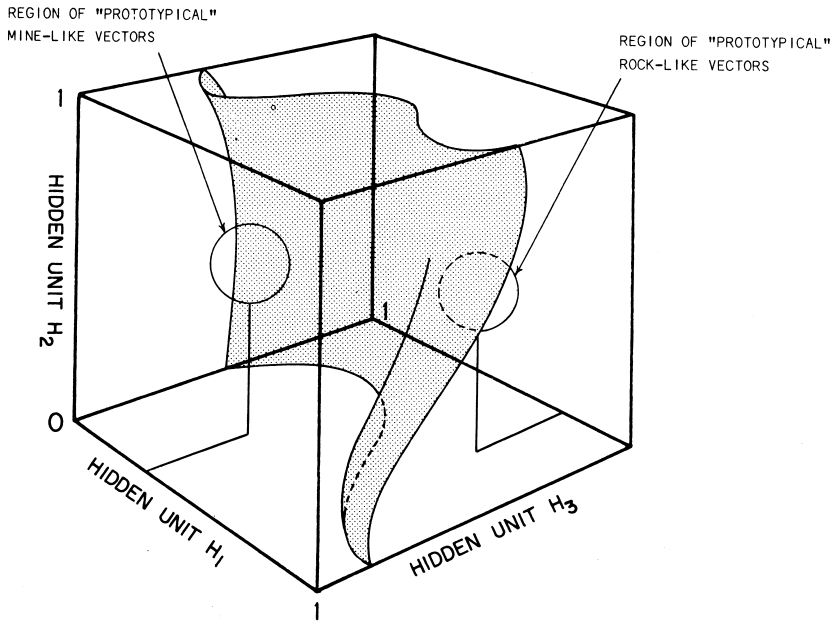


Figure 5.1: Churchland's impression of prototypes in the activation state space of a trained neural network (from "The nature of theories").

in order to adequately model the empirical data. For an excellent review of historical developments and outstanding issues in the study of conceptual similarity, see Goldstone and Son (2005).

Prototypes

We noted several times in the previous section that Churchland often writes about regions in activation state space and *prototypes* as if they are identical. Perhaps the earliest example of this is the following: "under the steady pressure of a learning algorithm . . . the network slowly but spontaneously generates a set of internal representations [that] take the form of a set or system of similarity spaces, and the central point or volume of such a space constitutes the network's representation of a *prototypical* [category member]" (1988, p. 123). The description is consistent with a diagram that Churchland often uses, beginning in "On the nature of theories" (1990), and shown here as Figure 5.1.

It seems clear from Figure 5.1 that Churchland intends us to take his description of a prototype as a "central point or volume" in activation space

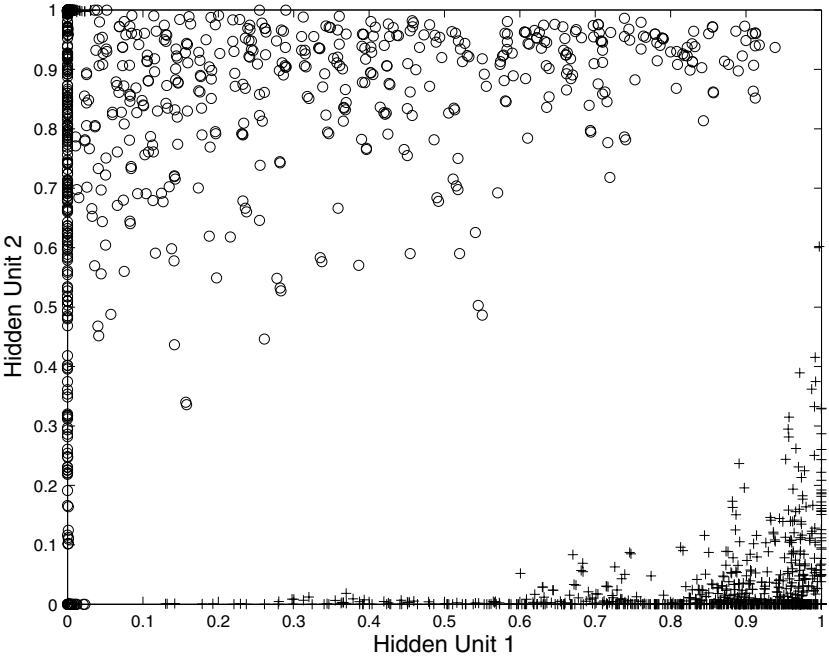


Figure 5.2: Actual distribution of hidden unit activations in a trained connectionist network.

as literally true. The volumes that he labels as prototypes are indeed at or close to the centers of the regions separated by the hypersurface depicted in the figure. Churchland is not alone – it has become part of philosophical lore that connectionist networks naturally learn and use prototype representations of this kind. Prinz (this volume), for example, asserts that connectionist networks spontaneously form prototypes in activation space of just the sort that Churchland depicts in Figure 5.1.

However, this is not really how connectionist networks work, at least not feedforward networks trained by backpropagation. To demonstrate this, we trained a feedforward network by backpropagation on a classification problem and plotted the actual locations of points in its activation state space, shown in Figure 5.2. The problem was to discriminate poisonous from nonpoisonous mushrooms in a hypothetical sample corresponding to 23 species in the Agaricus and Lepiota family, based on 22 nominally valued physical attributes such as the shapes of their caps and the colors of their stalks (Schlimmer 1987a, 1987b). The 22 nominally valued attributes were represented locally in the input by converting them to real values uniformly

distributed in the interval [0, 1]. For example, the "cap shape" attribute, which could have values of "bell," "conical," "convex," "flat," "knobbed," or "sunken" – six possible values – was represented in our inputs by values from the set {0, 0.2, 0.4, 0.6, 0.8, 1.0}. The targets were represented by 0 (edible) or 1 (poisonous). The network had two layers, with two units at the hidden layer and one unit at the output layer, with logistic activation functions at both layers. The training data consisted of 1624 randomly selected patterns, the validation data consisted of a different 1624 randomly selected patterns, and the test data consisted of a third non-overlapping set of 1624 randomly selected patterns. The network was trained by backpropagation on the training data until the mean squared error on the validation data fell below 0.01. In the example shown, this happened after about 80 epochs, and the mean squared error on the test set was 0.0162 after training.

Figure 5.2 shows the activations of the hidden units on the test patterns after training, with the activation vectors representing edible mushrooms marked with circles (o) and those representing poisonous mushrooms marked with plusses (+).

The actual hidden unit activations in Figure 5.2 don't look very much like the hypothesized prototypes in Figure 5.1. Clearly the network depicted in Figure 5.2 has learned to distinguish edible from poisonous mushrooms by distributing their respective hidden-unit activations in such a way that it is easy to separate them. It has, so to speak, "pushed" the edible mushrooms into the upper-left corner of activation space and the poisonous mushrooms into the lower-right corner of activation space, enabling it to "draw a line" between the hidden unit activations, separating the two categories. If there are prototypes in this space, they are in the *corners* of the activation state space, not in the centers of the spaces separated by the discriminating line that the output units draw between, roughly (0.1, 0) and (1.0, 0.6). This is not an accidental artifact of a single renegade run starting with unfortunate random connection weights and winding up in a local minimum. It happens every time the network is trained on this problem, even when the network is afforded even more "extra" room in activation space by giving it three or more hidden units.

In any case, even though the corners of the hidden unit activation space are where backprop "tries to" represent data, it stretches the imagination to construe the corners of such activation space as prototypes. In its ordinary usage in psychology, a prototype is a template for a concept, such that putative exemplars of the concept can be judged according to their similarity to the prototype. It is commonly assumed in the psychological literature that

prototypes are the central tendencies (in the statistical sense, e.g., averages) of their category instances, not only physically but also psychologically, much as Churchland depicts them in Figure 5.1. The prototype represents the best (i.e., most central) instance of the category, and other instances of the category are nearer to or farther from the prototype in psychological space as a function of how similar they are to the prototype.

However, there is nothing to indicate that the network depicted in Figure 5.2 either (a) represents the "best" edible mushroom – the central or average edible mushroom – at or near (0, 1), or (b) interprets distance from (0, 1) as indicating the "degree" of edibility. In general, backpropagation adjusts the weights to the output layer to discriminate the inputs by means of a linear transformation of the hidden unit activations, and adjusts the weights to the hidden layer to maximize the (output layer) discriminant by non-linearly transforming the input patterns into hidden-unit patterns (Bishop 1995: 228). To the extent that activations in the corners of hidden unit activation space are semantically interpretable, then, we could consider them to be the most discriminable exemplars – the ones that are easiest to tell apart. They are psychological *extremes* rather than psychological prototypes.

There is a kind of network, called a *radial basis function network* (RBFN), that more closely resembles Churchland's idea of prototypes in hidden-unit activation space (Bishop 1995). In the standard connectionist models that Churchland usually uses for his examples, units compute a non-linear function (normally a threshold or sigmoid) of the scalar product of the sum of their inputs with a weight. The computation performed by each unit in a such a typical feedforward connectionist network is very straightforward. Each unit j computes a function from $\mathbb{R}^n$ (a vector of the activations of the n units $i_1, \ldots, i_n$ feeding into j) into $\mathbb{R}$ (the activation of unit j), of the form:

$$z_j = g\left(\sum_{i=1}^{n} w_{ij} x_i + w_0\right) \tag{4}$$

where $x_1, \ldots, x_n$ are the activations of the input units $i_1, \ldots, i_n$; $w_{1j}, \ldots, w_{nj} \in \mathbb{R}$ are the weights on the connections from the input units to j; w_0 is a "bias"; and $g : \mathbb{R} \to \mathbb{R}$ is the "activation function," usually the logistic function $g(x) = 1/(1 + e^{-x})$ or the hyperbolic tangent function $g(x) = \tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$. Still simpler models take the activation function to be a threshold function and output a single bit. Most modern models use sigmoid activation functions, however, both because sigmoidal gates have real-valued outputs, giving multilayer networks of sigmoidal gates more computational power than those with threshold gates,

and because the backpropagation learning algorithm requires a differentiable function. Normally, all of the units at a given level of the network compute this function simultaneously. In practice, the activations are calculated serially because of the limitations of simulators running on ordinary hardware. This is irrelevant, however, because the *model* is that they are calculated in parallel, and this is how they are actually calculated when parallel hardware is available.

In a radial basis function network, by contrast, the activation of a hidden unit is calculated by comparing the vector of input activations to a prototype vector. Each hidden unit j in an RBFN computes a function from $\vec{x} \in \mathbb{R}^n$ (an input vector) into $\mathbb{R}$ (the activation of unit j), of the form:

$$z_j(\vec{x}) = \phi(d(\vec{x}, \vec{\mu}_j)) \tag{5}$$

where $\vec{\mu}_j$ is a vector determining the center of the basis function for hidden-layer unit j, and the function $d(\cdot) : \mathbb{R}^n,\ \mathbb{R}^n \rightarrow \mathbb{R}$ is a distance function, usually the Euclidean distance (1), between the input vector $\vec{x}$ and the center of the basis function for hidden-layer unit j at $\vec{\mu}_j$. The basis function $\phi(\cdot) :\ \mathbb{R} \rightarrow \mathbb{R}$ is usually a spherical Gaussian:

$$\phi(d_j) = \exp\left(-\frac{d_j^2}{2\sigma_j^2}\right) \tag{6}$$

where σ_j is a "width" parameter determining the smoothness of the basis function for hidden-layer unit j.

The activation y_k of an output unit k in an RBFN is a simple linear combination of the basis functions:

$$y_k(\vec{z}) = \sum_{j=1}^{n} w_{jk} z_j \tag{7}$$

In the first stage of training an RBFN, the parameters of the basis function (6) – the centers $\vec{\mu}$ and the widths $\vec{\sigma}$ – are set by an unsupervised learning technique, usually by considering the basis functions to be components of a Gaussian mixture model and optimizing them using the expectation maximization (EM) algorithm. Once the basis function parameters have been fixed, the weights to the output units can be quickly determined using singular value decomposition.

There is a lot to like about RBFNs. They are fast and easy to train. They have nice mathematical properties, such as the fact that their hidden unit activations can be interpreted as the posterior probabilities of the presence of features in the input space, and their output layer weights can

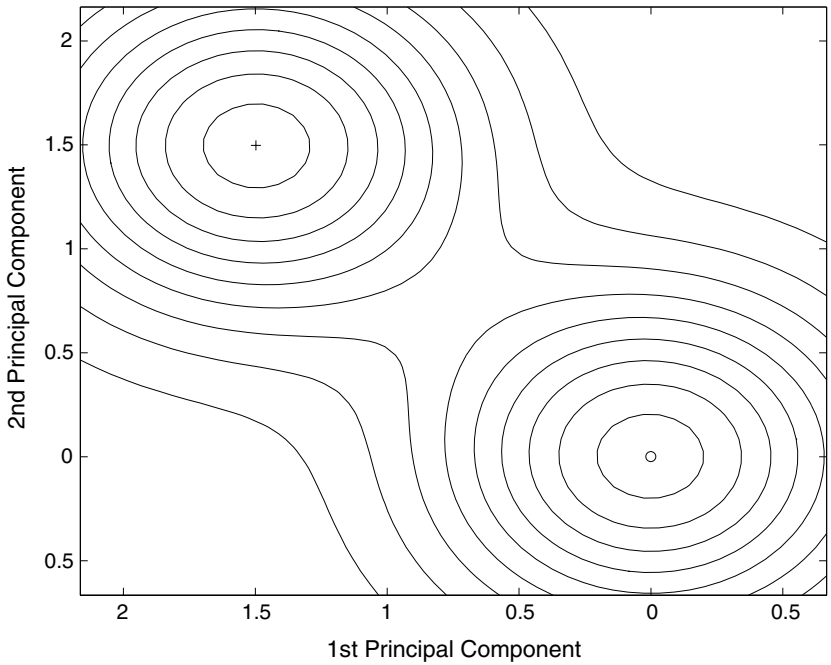


Figure 5.3: Isoprobability contours for the basis functions in a radial basis function network.

be interpreted as posterior probabilities of membership in the output class given the presence of the features represented at the hidden layer (Bishop 1995).

RBFNs also have nice "psychological" properties. Although it is difficult to find a way to interpret the hidden units of a backprop network as representing prototypes, it is natural to interpret the hidden units of an RBFN as representing prototypes. To demonstrate this, we trained an RBFN with two hidden units and one output unit on the same data used to train the backprop network whose hidden unit activations are shown in Figure 5.2. Because the basis functions are fit to the input data, they have 22 dimensions. To visualize them, we found the first two principal components of their centers and plotted two-dimensional isoprobability curves around them. The resulting graph is shown in Figure 5.3.

Intuitively, the basis functions shown in Figure 5.3 are much more like the "prototypes" that Churchland depicts in Figure 5.1 than the hidden unit activation distributions in Figure 5.2. Their centers represent the "best" examples of edible and poisonous mushrooms, respectively. Furthermore,

distance from the centers represents systematic differences in the system's certainty: the closer an exemplar is to the "edible" center, the more probable that it is edible, and the closer an exemplar is to the "poisonous" center, the more probable that it is inedible.

However, the performance of RBFNs using basis units corresponding to intuitive prototypes may not match what one would like. For example, the RBFN depicted in Figure 5.3 achieved mean squared error of only about 0.4, even after 1000 iterations of EM during the training stage. (Recall that the backpropagation network depicted in Figure 5.2 achieved mean squared error of less than 0.01 after fewer than 100 epochs!) Adding more basis units would certainly improve the performance of the RBFN. However, doing so would entail giving up our interpretation of the basis units as prototypes for edible and poisonous mushrooms. Since each basis unit corresponds to a prototype, adding more basis units means adding more prototypes. It is reasonable to suppose that a human mycologist might have multiple prototypes for edible and poisonous mushrooms (an Edible-A prototype, an Edible-B prototype, a Poisonous-A prototype, and so on). At the extreme, an Agaricus-Lepiota expert probably would have a prototype for each species (a prototype for *Lepiota clypeolaria*, a prototype for *Lepiota procera*, and so on). An RBFN with many basis units might be suitable not only for classifying mushrooms on these sorts of fine-grained distinctions but also for modeling the subordinate (species-level) prototype structure of the human expert. However, there would be no natural way to aggregate the many basis units in such an RBFN into just two prototypes corresponding to the superordinate categories, edible and poisonous. It seems that neither a standard feedforward network trained by backpropagation nor an RBFN is a particularly good model of a prototype theory of human categorization.

On the other hand, prototype theory may not be a good explanation of human categorization. Reaction times in human categorization experiments decrease with the distance in psychological space from the stimulus representation to the decision bound that separates exemplars of contrasting categories; subjects are quicker to categorize stimuli the further they are from the category boundary (Ashby, Boynton, & Lee 1994). The natural prediction of prototype theory, by contrast, is that reaction time should increase with the distance between the stimulus and the nearest prototype. Barsalou (1985, 1991) has pointed out that, at least in the case of some goal-directed categories, the "ideal points" in psychological space may be at the extremes, rather than at the prototypes. That is, people sometimes categorize things based on extremes rather than on

central tendencies. For example, the best exemplar of "drinking water" is one that has zero contaminants, rather than one that has the average amount of contaminants found in a wide sample of potable water. Furthermore, Palmeri and Nosofsky (2001) have shown that prototypes in physical feature space (i.e., the central tendencies of the features of stimuli presented as category instances) may sometimes behave as if they were at extreme points in psychological space relative to other category instances. Although Churchland may have been wrong to assume that the hidden-unit activations of connectionist networks trained on categorization tasks are like prototypes, this evidence suggests at least the possibility that he may have been right to argue that they are good models of human categories.

Biological Plausibility

As discussed in the previous section, there are certain kinds of "connectionist neural networks" that are completely implausible biologically, even though they have excellent mathematical and computational properties, not to mention nice conceptual glosses. However, many philosophers, following Churchland, believe that ordinary feedforward connectionist networks consisting of sigmoidal gating neurons are at least somewhat biologically plausible, if highly abstract. Doubts about the biological plausibility of *backpropagation* are legion in the philosophical literature about Connectionism. So are doubts about the *psychological* plausibility of connectionism. However, just about everyone seems to accept the claim that the networks themselves (if not the supervised training algorithms) are biologically plausible, at least "broadly speaking" (see, for example, Prinz, this volume). In this section, we take a close look at just how broadly we must speak in order to justify the claim that connectionism is biologically plausible.

Churchland himself discusses many biologically implausible aspects of connectionism in some detail (1990: 86–92). The differences that Churchland notes include the fact that biological neural networks are not fully interconnected (as many connectionist models are) and the fact that individual biological neurons generally do not have both inhibitory and excitatory postsynaptic effects (while individual units in connectionist networks may have inhibitory connections to some units and excitatory connections to others). While none of the biologically implausible factors that Churchland recognizes subvert his claim that activation vectors are the fundamental units of representation, there are reasons to believe that this is not true in biological neural networks.

A biological neuron emits short ($\approx 1 - 2$ milliseconds) voltage pulses (with amplitudes of about 100 millivolts) called "action potentials" or "spikes." A spike in a neuron *j* is triggered by a complex process that starts when spikes in other neurons reach synapses that connect to *j*. A synapse transforms the spike in the presynaptic neuron *i* into a postsynaptic potential of about 1 millivolt in *j*, lasting from 1 millisecond to several seconds or longer. The postsynaptic potential may be excitatory (tend to increase the probability that the postsynaptic cell *j* fires) or inhibitory (tend to decrease the probability that the postsynaptic cell *j* fires), depending on whether it increases or decreases the membrane potential of *j*. The neuron *j* emits an action potential when some threshold, typically a time-dependent function of membrane potential, is crossed.

A very detailed model of a single neuron – the most famous is the Hodgkin-Huxley model of the giant axon of the squid – attempts to capture this process with as much detail and accuracy as possible. Such models often take into account the equilibrium potential, the specific properties of numerous ion channels (channels with sodium, potassium, calcium, and other ions, most with multiple sub-types operating at different time scales, and some with multiple sub-types sensitive to different voltage thresholds), the different types of synapses (with different neurotransmitters, different receptors, different time scales, and so on), the spatial layout of the dendritic tree (which results in a non-uniform distribution of membrane potential, inducing additional current along the membrane as well as across it), and the specific shape, amplitude, and duration of both postsynaptic potentials and action potentials. (See Gerstner & Kistler 2002 for an excellent review.)

There are simpler, so-called "phenomenological" models of the neuron. Philosophers are likely to find themselves either amused or shocked to find the term "phenomenological" applied to a mathematical model of a single neuron, but they can rest assured that the use of the term in this context is intended neither to beg any important questions (about how consciousness might arise from neural activity) nor to stipulate any sort of philosophical methodology (be it Husserlian or otherwise) toward answering that question. In this usage, which is common in the sciences, "phenomenological" means merely "relating to a phenomenon." A phenomenological model simulates some phenomenon without attempting to capture its underlying causes. Phenomenological models of the neuron attempt to reproduce some aspect of a neuron's behavior (e.g., the timing of spikes) while abstracting away from the biophysical and biochemical details. The sacrifices in accuracy and detail are balanced by gains in simplicity and comprehensibility.

The connectionist model of the neuron as a sigmoid gate described by Equation (4) is a very simple phenomenological model. On the standard interpretation of the correspondence between this model and a biological neuron, the weights $w_1, \ldots, w_n$ are identified with the "coupling strengths" of the presynaptic neurons (the efficiency of synapses impinging on j), and the activation of a unit is identified with the firing rate of the neuron. The theory behind this model is that, in the neural code, the "signal" is not carried by the specific times at which individual spikes occur, but rather by the mean rate at which they occur, that is, by the number of spikes generated in some relatively long window. It is important to emphasize that this principle, known as "rate coding" is a *hypothesis* about the neural code.

The rate coding hypothesis has come under significant scrutiny, and several alternatives have been proposed (e.g., Gerstner 2001, Gerstner & Kistler 2002, Maass 1998, Hopfield & Brody 2001, Shastri & Ajjanagadde 1993). In most such models, action potentials are considered to be "stereotyped" events (i.e., they are all equivalent – a spike is a spike is a spike) and are therefore modeled simply as formal events that mark points in time. Coding is hypothesized to take place by the specific timing of spikes, by their phase difference with respect to some periodic signal, or by their temporal cooccurrence or synchrony. The best-known model is called the "leaky integrate-and-fire" model, because it models a neuron as summing its postsynaptic potentials over time (integrating) with some decay (leaking) and "firing" or spiking when its membrane potential exceeds some threshold. In a generalization of the leaky integrate-and-fire model known as the "spike response model" (Gerstner 2001), the membrane potential of neuron j is modeled as:

$$u_j(t) = \sum_{k=1}^{m} \eta\left(t - t_j^k\right) + \sum_{i=1}^{n} \sum_{l=1}^{o} w_{ij}\varepsilon\left(t - t_i^l\right) \tag{8}$$

where $t_j^1, \ldots, t_j^m \in \mathbb{R}$ are the previous firing times of neuron j; $w_{1j}, \ldots, w_{nj} \in \mathbb{R}$ are measures of the efficiency of the synapses impinging on j; and $t_i^1, \ldots, t_i^o \in \mathbb{R}$ are the previous firing times of the n presynaptic neurons $i_1, \ldots, i_n$. The function $\eta(t - t_j^k) : \mathbb{R} \to \mathbb{R}$ determines the voltage contribution to the membrane potential of j at time t that is due to the previous spike of j at time t_j^k. It characterizes the reset of the membrane potential to a resting level immediately after each spike, causing a period of "refractoriness" after each spike that tends to prevent another spike from happening for some time. The function $\varepsilon(t - t_i^l) : \mathbb{R} \to \mathbb{R}$ determines the voltage contribution to the membrane potential of j at time t that is due to

the presynaptic spike on neuron i at time t_i^l. It therefore characterizes the response of j to incoming spikes – the postsynaptic potential.

The neuron j emits a spike when its membrane potential reaches some threshold ϑ Hence, its spiking history $\vec{t}_j$ is updated as follows:

$$\vec{t}_j = \begin{cases} [\vec{t}_j, t] & \text{if } u_j(t) = \vartheta \\ \vec{t}_j & \text{otherwise} \end{cases} \tag{9}$$

In other words, whenever a neuron spikes, the time is added to the neuron's spiking history.

The difference in complexity between Equation (4), on the one hand, and Equations (8) and (9) on the other, is obvious. However, the complexity of the mathematics is not itself an issue. The question is: does the difference really matter with respect to Churchland's position? We believe that it does. In a connectionist network consisting of sigmoidal gating units acting according to Equation (4), the information that a unit contributes to the network is accurately and completely characterized by its current state – its activation z_j. Hence, in a large network consisting of many such neurons (i.e., a connectionist network), it is fair to say that the informational state of the network consists of a vector of the current activations of each of the units. By contrast, in a neural network consisting of spiking neurons acting according to Equation (8), the information that a neuron contributes to the network is *not* accurately and completely characterized by its current state – its membrane potential $u_j(t)$. Rather, the information that the neuron contributes to the network is characterized by its spiking history $\vec{t}_j$, the vector of times at which it emitted an action potential. One could consider the spiking history to be part of the state of a neuron, for example by characterizing it by set of differential equations. However, this entails attributing a greater complexity to the unit itself (the state of the unit consists of the values of at least two equations instead of the one that characterizes a PDP neuron), and therefore jeopardizes the idea that the state of a network can be captured by a single vector.

It is possible to calculate a firing rate from the spiking history by counting the number of spikes in some time window. Hence, the spiking response model embodied in Equation (8) is consistent with the rate coding hypothesis, and, therefore, consistent with Connectionism. However, the rate coding hypothesis itself has come under attack. One of the main reasons for this is that a code based on an average over time is slow, because it requires a sufficiently long period of time to accumulate enough spikes for the average of their intervals to be meaningful. While firing rates in peripheral neurons are relatively fast, the typical firing rates of cortical neurons are

under 100 Hz. If we assume that the rate is 100 Hz, then we would need 50 ms to sample 5 spikes, a reasonable lower bound for a meaningful average. If we assume that classifying a visual stimulus requires 10 processing steps, then it would require 500 ms. However, empirical data shows that human beings can actually classify complex visual stimuli in about 200 ms (Thorpe, Fize, & Marlot 1996). There is also evidence that stochastic fluctuations in the timing of primate cortical action potentials are not simply due to random noise within individual cells, and that the cortical circuitry preserves the fine temporal structure of those fluctuations across many synapses (Abeles, Bergman, Margalit, & Vaadia 1993; Bair & Koch 1996). It has been established physiologically that connections between cortical neurons are modified according to a spike-timing dependent temporally asymmetric Hebbian learning rule (synapses that are active a few milliseconds before the cell fires are strengthened, whereas those that are active a few milliseconds after the cell fires are weakened), and modeling studies have established that this mechanism implements a form of temporal difference learning that could be used for predicting temporal sequences and detecting motion (Rao & Sejnowski 2001, Shon, Rao, & Sejnowski 2004). Finally, there have been theoretical arguments that temporal synchrony is required for binding representations of features together when appropriate (Malsburg 1995), and experimental evidence supports this hypothesis (Singer & Gray 1995). All in all, it seems unlikely that human visual cortex uses rate coding exclusively.

It might be possible to salvage rate coding by interpreting the rate not as the rate at which a single neuron spikes but as the mean rate at which a population of neurons spike. This is sometimes called *population rate coding*. The idea is to determine the mean rate of firing of all the neurons in a small spatial neighborhood over a short time interval. By increasing the number of spikes per unit of time, this approach solves the "slowness" problem with the naive rate coding approach that brings it into conflict with empirical results. In fact, it turns out that neurons coding by a population rate can respond nearly instantaneously to changes in their inputs, under reasonable assumptions (Gerstner 2001).

However, even if population rate coding is a reasonable hypothesis about networks of biological neurons, it is by no means clear how it should be mapped onto connectionist models. If unit activations model population firing rates, then it is no longer reasonable to assume that units correspond to neurons; instead, units must correspond to populations of neurons. Then, connections between units cannot correspond to synapses between (individual) neurons. Perhaps we could consider connections between units to

correspond to overall average connection strength between two populations. However, the populations of the population rate coding hypothesis are defined spatially (by their physical proximity to each other) not topologically (by their connections to each other). Hence, the "connection strength between two populations" must be considered merely a statistical regularity, not a causal mechanism. (Of course, in some cases – namely, when two populations under consideration are not only spatially proximal but also topologically connected – there will be a causal mechanism; our point is only that there need not be one in every case.) As a consequence, it is no longer reasonable to assume (in general) that weights correspond to synaptic efficiencies. It is possible that the analogy between connectionist networks and neural networks could be reconstructed along these lines, but this would require giving a detailed account of how the key elements of connectionist networks (minimally: units, connections, activations, and weights) *do* correspond to features of biological networks using population rate coding. This is an interesting problem, but not one that we can take up here.

In the absence of a defensible mapping between connectionist networks and biological neural networks, it seems only fair to say that connectionist networks are simply *not* biologically plausible. This can be difficult to accept, on two counts: first, the intuitive plausibility of a network of interconnected units modeling a network of neurons; and second, the enormous success that connectionist networks have displayed in modeling, generating, and predicting complex cognitive skills. With respect to the first objection (the intuitive plausibility of connectionism), we can only remind the reader that – as discussed in detail in a previous section – our best current understanding of biological neural networks is potentially inconsistent with some fundamental principles of connectionist modeling. At least in cortex, it appears that neurons may not use rate coding. Instead, their operation may be crucially dependent on the timing of individual spikes, and the history of such timings. With respect to the second objection (the success of connectionism), it is important to note that connectionism's success in modeling complex cognitive abilities is consistent with its biological implausibility. It might be, for example, that both connectionist networks and biological neural networks are calculating statistical properties of their inputs and performing complex probabilistic inferences on them. The fact that connectionist networks perform such calculations in a way that is biologically implausible does not hinder their abilities to do so.

It is still tempting to say that connectionist networks are somehow *more* biologically plausible than Computationalist models. Computationalist

models are primarily declarative and procedural programs executed on von Neumann digital computers. These intuitively *seem* further removed from the brain than connectionist networks. It was possible for a long time to tell a plausible and consistent story about exactly how connectionist networks map to biological neural networks. Even now it is possible to tell a somewhat implausible but still consistent story about how connectionist networks map to biological neural networks. Computationalist models, on the other hand, have never had such a story, and show little promise of generating one anytime soon. So perhaps there is something to the view that, while connectionist networks may not be the most biologically plausible models available, at least they are more biologically plausible than Computational models. It seems not only unreasonable but unnecessary to divide models of cognition into two categories, plausible and implausible, and assert that a model is either one or the other. Rather, biological plausibility is a spectrum, with a wide range between the most plausible models and the most implausible. Computationalism, we might say, is closer to the implausible end of things. While we might once have thought that connectionism was quite clearly on the plausible end of things, we might say that it is somewhere in the middle. Phenomenological models from computational neuroscience, such as the leaky integrate-and-fire model and the spiking response model, are more on the plausible side. Finally, detailed biophysical models, such as the Hodgkin-Huxley model of the giant axon of the squid and more recent models in the same vein, are as plausible as we can be right now.

Churchland's more abstract thesis, which we have dubbed Vectorialism, may fare somewhat better. Recall from Table 5.1 that Vectorialism is the theory that thoughts are vectors that beliefs are matrices, that thinking is transformation of thought-vectors, and that learning is transformation of belief-matrices. In the spiking response model, the state of a neuron is a vector consisting of the times of its previous firings. On a simplistic model of "spike coding," these times themselves would carry information. On one interpretation, the time between the stimulus and the first spike encodes information; on another ("phase coding"), information is carried by the phase of a spike with respect to some other periodic signal; on a third ("correlation coding"), information is encoded by the intervals between the firings of two or more neurons (Gerstner & Kistler 2002). It is not yet known whether or when biological nervous systems use these or other possible coding schemes; deciphering the neural code is an ongoing research project. It is entirely possible that biological neural networks use all of these codes and others not mentioned here and even not yet imagined, and that

the code used might vary from one organism to another, from one system to another within the same organism, and even from one task to another within the same system (Maass 1998). What is important here is that *all* of these coding schemes carry information by quantities that can be represented by numeric vectors. The actual history of spike times for a neuron is a vector, as we saw in Equation (9), so the state of the system could be captured by a matrix of such vectors. Time-to-first-spike is a real number, so the state of the system could be captured by a vector of time-to-first-spikes for all the relevant neurons. Likewise for phase. The correlation coding hypothesis is particularly interesting, because it surmises that information is not encoded in the spikes themselves but rather in their relations. Presynaptic neurons that fire simultaneously communicate to a postsynaptic neuron that they belong together. Of course, this too could be encoded numerically and expressed as a vector. Indeed, Paul Smolensky has shown how to encode Lokendra Shastri's temporal synchrony model of variable binding as a tensor (Tesar & Smolensky 1994).

Reducing the information content of all of these diverse coding schemes to a raw description as "vectors" obscures their important differences and unique properties. There is first of all the fundamental difference between all of the various spike coding hypotheses and the rate coding hypotheses: that the phenomena of interest are points in time rather than rates. Then there is the difference between population coding hypotheses, single-neuron coding hypotheses, and correlational hypotheses, with respect to "how many" neurons are relevant to determining the signal. Finally, there are all of the fine differences between the various spike coding hypotheses. These are important differences, not only for neuroscience *per se* but also for any theory of cognition that wants to make a claim of biological plausibility. Surely, it matters whether the computational units are single neurons, topologically connected combinations (groups) of neurons, or whole (spatially contiguous) populations of neurons. Similarly, it matters whether the phenomena of interest are time series, rates, or some other kind of quantity. Finally, whatever the computational units are, whatever the quanta of information are, it matters how the information is encoded.

Since Vectorialism encompasses all these alternatives, we must ask whether it is too general a view to be of much value. After all, nearly anything can be reduced to a vector or a matrix by a suitable interpretation, and any vector or matrix can be transformed by many mathematical operations. It is even possible to construe Computationalism as Vectorialism, by taking bits in the machine's memory to be vectors of truth values and the logical operations of the CPU to be transformations between such

vectors. Churchland's particular brand of Vectorialism got its teeth from its association with Connectionism. The only charitable way to interpret Churchland's flavor of Vectorialism without making it vacuous is to suppose that the core hypothesis is not that thoughts are vectors per se, but that thoughts are vectors-of-activations. Likewise for beliefs (not matrices *per se*, but matrices-of-connection-weights), thinking (not vector transformations *per se*, but transformations-of-activation-vectors) and learning (not matrix transformations *per se*, but transformations-of-weight-matrices). We could call this view "Connectionist-Vectorialism."

The problem with interpreting Vectorialism as Connectionist-Vectorialism is that it then becomes subject to the fate of Connectionism. Specifically, since Connectionism is not all that biologically plausible, neither is Connectionist-Vectorialism. So, if biological plausibility is a desideratum of our theory of mind, Connectionist-Vectorialism does not fit the bill all that well.

How should we proceed from here? One possible approach would be to wait for a single clear victor in the neural coding debate currently being waged in computational neuroscience. One could then build a theory of mind around that hypothesis, much as Churchland has built one around connectionism, and go on to explore its ramifications in other areas like epistemology, philosophy of science, and ethics, again much as Churchland has done with Connectionism. One problem with this sort of "winner take all" strategy for determining our ultimate theory of mind is that it is entirely possible, even likely, that there will be no single winner in the neural coding debate. As we mentioned earlier, it is possible, if not likely, that it will turn out that different organisms, different systems, and even different tasks evoke fundamentally different neural coding schemes.

An even larger problem with the "winner take all" strategy for determining our ultimate theory of mind is that there are all kinds of other models of cognition besides computational neuroscience. Connectionism, for one, shows no signs of going away – despite widespread acknowledgment within the modeling community that it is not very realistic biologically, it continues to be widely used to model cognitive phenomena. Connectionism has also contributed to the explosion of interest in machine learning, statistical learning theory, and information theory in recent years. Machine learning is the study of algorithms for programming machines (computers) to learn to solve pattern recognition, categorization, classification, approximation, estimation, regression, and generalization problems, without concern for biological plausibility. Statistical learning theory is the study of the mathematical nature of those learning problems, in the absence even of any specific

concern for machine implementation. Information theory is the study of the fundamental properties of information, and turns out to have important and interesting links with statistical learning theory. Each of these fields (computational neuroscience, connectionism, machine learning, statistical learning theory, information theory) has made important contributions to the others, and these contributions have flowed in all directions.

To take a single example, Bayesian belief networks (also known as graphical models) arose out of probabilistic extensions to the binary or three-valued logics commonly used in early expert systems. They have since received penetrating statistical analysis, resulting in a solid mathematical foundation. They have also engendered an intense interest in the machine learning community, resulting in efficient exact inference algorithms for special cases and relatively fast approximation algorithms for the general case (see Russell & Norvig 2003 for a review). Psychologists have used them to model many cognitive phenomena. At least one respectable philosopher, Clark Glymour (2001), has asserted that they are a fundamental part of our cognitive architecture and are the mechanism behind our grasp of scientific explanations. These are, of course, much the same claims that Churchland has made about connectionism.

We do not believe that there is any reason to think that Glymour is any more right *or any more wrong* about graphical models than Churchland was about connectionism. Connectionist networks are an instance of graphical models, and *both* frameworks provide useful models of significant domains of cognition. So too do many other models from many other areas, including many other models from psychology, computational neuroscience, connectionism, machine learning, statistical learning theory, and information theory that we have not discussed. The obsession with finding a single theory that "explains the mind" seems to be a peculiarly philosophical affliction. Other fields – including those that are arguably the most centrally engaged in "explaining the mind," cognitive psychology and computational neuroscience – seem quite at home with having multiple models. The driving force behind the philosophers' affliction seems to be a fondness for unity, specifically the unity of science and the unity of explanation. None of the models that we have discussed are nonmaterialistic, nor do they challenge the unity of science as a whole in any other way. Considering them together, in all of their diversity, it is tempting to say that they do not provide a unified explanation of "mind." Taking this to mean that they do not provide a *single* explanation, it is clearly true. Rather than concluding that they must all be subsumed into some higher-level, "more unified" explanation, however, we would argue that the proper response is to conclude that "the mind"

is not a unitary phenomenon. Not only are multiple levels of explanation required to explain cognitive phenomena, so too (at least in some cases) are multiple models required at the same (e.g., computational) level. There does not seem be a single "privileged" perspective (those worshipping at the Church of Bayes notwithstanding). Although the resulting explanation is not unitary, it is consistent.

This is not to say that any and all models are equally good. Freud's model of cognition, for example, was quite bad (even though, like Fortran, it is surprisingly resilient). Models can and must be evaluated on the basis of many criteria, including precision, accuracy, falsifiability, consistency, simplicity, comprehensibility, plausibility, and utility. Some will certainly fare better than others. New and better models will be developed and older and worse models will fall out of use. In the end, there is no particular reason to think that just one will triumph. We think that Churchland would be happy with this conclusion, since it is consistent with both his scientific realism (his view that science is converging on the truth) and his pragmatism (his view that the truth is what works best).

CONCLUSIONS

Churchland's use of connectionism to support novel theories in the philosophy of mind, epistemology, the philosophy of science, and ethics is highly original and thought provoking. It has also had a lasting effect on the field of philosophy, generating many intense exchanges between parties of all philosophical persuasions. In this chapter, we outlined how Churchland has applied connectionism in a variety of philosophical areas, and then discussed several empirical issues with Churchland's interpretation of connectionism. Specifically, we showed that: (1) Churchland's claim that semantic similarity corresponds to proximity in activation space is contradicted by some experimental findings in psychology; (2) Churchland's claim that ordinary connectionist networks trained by backpropagation represent categories by prototype vectors is ill founded, although there are other sorts of connectionist networks that can be interpreted as representing categories by prototypes; and (3) in light of recent developments in computational neuroscience that call the rate coding hypothesis into question, it may turn out that connectionist networks are not very biologically plausible after all.

While making an effort to present Churchland's use of connectionism in context, we have avoided making too much of the more specifically philosophical issues that Churchland addresses. There is certainly

enough criticism of Churchland's philosophy around. At the same time, it is possible that someone will try to use the empirical results we have discussed to advance a more philosophical criticism against Churchland. While Churchland might have his own reservations – either about the empirical conclusions we have drawn here, or about any philosophical uses to which they might be put – we would expect that he would be more excited than dismayed by a philosophical criticism based on empirical data. After all, Churchland is the first truly natural epistemologist. Quine (1951) opened the doors by arguing that natural science *does* matter to philosophy (and vice versa). Churchland was the first to boldly step through those doors and demonstrate how naturalized epistemology could, and should, be done. Even if everything Churchland ever wrote about connectionism and neuroscience should turn out to be utterly wrong, which is unlikely, that legacy will remain.

EPILOGUE

Although this chapter has been critical of some aspects of Churchland's position, the authors would like to end on a personal note.

GWC

Paul Churchland and I found ourselves at the same AI workshop in Austria in 1990. At that conference, Paul gave a talk about how one might model the notion of someone changing their mind. He saw current beliefs as a pattern of activation, and that new activation entering the network would change where the network settled to. He pointed out that it would have to be a recurrent network or this wouldn't work. This is one time when I got an idea from a philosopher that I could act upon, and this idea led to one of my students, Dave Noelle, doing his thesis on learning by being told. I would like to acknowledge Paul for seeding that thesis in my mind!

AL

Paul's work on Connectionism and the reaction to it from other quarters in the philosophical community first grabbed my interest while I was an undergraduate, and they have held it ever since. Paul also inspired my thesis, which was in large part a defense of Churchland's Connectionism against certain objections that Fodor and Lepore had raised.

Notes

1. In this chapter, we use "Connectionism" (with a capital *C*) to refer to the philosophical position that the fundamental architecture of cognition is something like a connectionist network. We continue to use "connectionism" (with a lower-case *c*) to refer to the practice of using such networks in general, where the practitioners are agnostic about the philosophical claim. This distinction parallels our use of the term "Computationalism" to refer to the philosophical position that the fundamental architecture of cognition is something like a digital computer.
2. The question we ask here (a) "How many points can have the same nearest neighbor?" is different from the question (b) "How many points can be each other's nearest neighbors?" to which the answer is 2 points on a line in 1*D*, the 3 vertices of an equilateral triangle in 2*D*, the 4 apexes of a tetrahedron in 3*D*, and so on. It is also different from the question (c) "How many points can be the nearest neighbor of a given point?" to which the answer is 2 points in 1*D* and an infinite number in any higher dimension, arrayed around a circle, a sphere, or a hypersphere. The reason that (a) and (c) are different is that the nearest neighbor relation is not symmetric: the fact that *i* is the nearest neighbor of *j* does not entail that *j* is the nearest neighbor of *i*.

References

Abeles, M., Bergman, H., Margalit, E., & Vaadia, E. (1993). "Spatiotemporal firing patterns in the frontal cortex of behaving monkeys." *Journal of Neurophysiology* **70**(4): 1629–38.

Ashby, F. G., Boynton, G., & Lee, W. W. (1994). "Categorization response time with multidimensional stimuli." *Perception & Psychophysics* **55**(1): 11–27.

Bair, W., & Koch, C. (1996). "Temporal precision of spike trains in extrastriate cortex of the behaving macaque monkey." *Neural Computation* **8**(6): 1185–202.

Ballard, D. H. (1986). Parallel logical inference and energy minimization. *Proceedings of the 5th National Conference on Artificial Intelligence (AAAI-86)* (Vol. 1, pp. 203–9). Philadelphia, Morgan Kaufmann.

Ballard, D. H. (1999). *An introduction to natural computation*. Cambridge, MA, MIT Press.

Barsalou, L. W. (1985). "Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories." *Journal of Experimental Psychology: Learning, Memory, & Cognition* **11**(1–4): 629–54.

Barsalou, L. W (1991). Deriving categories to acheive goals. G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory*, (Vol. 27, pp. 1–64). San Diego, Academic Press.

Bickerton, D. (1995). *Language and human behavior*. Seattle, University of Washington Press.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford, Clarendon Press.

Churchland, P. M. (1986). "Some reductive strategies in cognitive neurobiology." *Mind* **95**(379): 279–309.

Churchland, P. M. (1988). Folk psychology and the explanation of human behavior. *A neurocomputational perspective: The nature of mind and the structure of science* (pp. 111–35). Cambridge, MA, MIT Press/Bradford Books.

Churchland, P. M. (1989a). Learning and conceptual change. *A neurocomputational perspective: The nature of mind and the structure of science* (pp. 231–53). Cambridge, MA, MIT Press/Bradford Books.

Churchland, P. M. (1989b). Moral facts and moral knowledge. *A neurocomputational perspective: The nature of mind and the structure of science* (pp. 297–303). Cambridge, MA, MIT Press/Bradford Books.

Churchland, P. M. (1989c). *A neurocomputational perspective: The nature of mind and the structure of science*. Cambridge, MA, MIT Press/Bradford Books.

Churchland, P. M. (1989d). On the nature of explanation: A PDP approach. *A neurocomputational perspective: The nature of mind and the structure of science* (pp. 197–230). Cambridge, MA, MIT Press/Bradford Books.

Churchland, P. M. (1989e). Preface. *A neurocomputational perspective: The nature of mind and the structure of science* (pp. xi–xvii). Cambridge, MA, MIT Press/Bradford Books.

Churchland, P. M. (1990). On the nature of theories: A neurocomputational perspective. C. W. Savage (Ed.), *Scientific theories* (Vol. 14). Minneapolis, University of Minneapolis Press.

Churchland, P. M. (1995). *The engine of reason, the seat of the soul: A philosophical journey into the brain*. Cambridge, MA, MIT Press/Bradford Books.

Cottrell, G. (1985). Parallelism in inheritance hierarchies with exceptions. *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, Los Angeles, CA.

Cottrell, G. W., Bartell, B., & Haupt, C. (1990). Grounding meaning in perception. H. Marburger (Ed.), *Proceedings of the German Workshop on Artificial Intelligence (GWAI)* (pp. 307–21). Berlin, Springer-Verlag.

Derthick, M. A. (1987). A connectionist architecture for representing and reasoning about structured knowledge. *Proceedings of the Ninth Annual Conference of the Cognitive Science Society* (pp. 131–42). Hillsdale, NJ, Lawrence Erlbaum Associates.

Elman, J. L. (1991). "Distributed representations, simple recurrent networks, and grammatical structure." *Machine Learning* **7**: 195–225.

Fodor, J. A. (1975). *The language of thought*. Cambridge, MA, Harvard University Press.

Gerstner, W. (2001). What's different with spiking neurons? H. Mastebroek & H. Vos (Eds.), *Plausible neural networks for biological modeling* (pp. 23–48). Boston: Kluwer.

Gerstner, W., & Kistler, W. M. (2002). *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge, UK, Cambridge University Press.

Gilmore, G. C., Hersh, H., Caramazza, A., & Griffin, J. (1979). "Multidimensional letter similarity derived from recognition errors." *Perception and Psychophysics*, **25**: 425–31.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge, MA, MIT Press.

Goldstone, R. L. (1994). "The role of similarity in categorization: providing a groundwork." *Cognition* **52**: 125–57.

Goldstone, R. L., & Son, J. (2005). Similarity. K. Holyoak & R. Morrison (Eds.), *Cambridge handbook of thinking and reasoning*. Cambridge, UK, Cambridge University Press.

Gorman, R. P., & Sejnowski, T. J. (1988). "Analysis of hidden units in a layered network trained to classify sonar targets." *Neural Networks* **1**: 75–89.

Hahn, U., Chater, N., & Richardson, L. B. (2002). "Similarity as transformation." *Cognition*, **87**: 1–32.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the theory of neural computation*. New York, Addison-Wesley.

Holyoak, K. J., & Gordon, P. C. (1983). "Social reference points." *Journal of Personality and Social Psychology* **44**: 881–7.

Hopfield, J. J., & Brody, C. D. (2001). "What is a moment? Transient synchrony as a collective mechanism for spatiotemporal integration." *Proceedings of the National Academy of Sciences* **98**(3).

Krumhansl, C. L. (1978). "Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density." *Psychological Review* **85**: 450–63.

Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago, University of Chicago Press.

Laakso, A., & Cottrell, G. W. (2000). "Content and cluster analysis: Assessing representational similarity in neural systems." *Philosophical Psychology* **13**(1): 77–95.

Maass, W. (1998). On the role of time and space in neural computation. *Proceedings of the Federated Conference of CLS'98 and MFCS'98* (Vol. 1450, pp. 72–83). Berlin: Springer.

Malsburg, C. von der. (1995). "Binding in models of perception and brain function." *Current Opinion in Neurobiology* **5**: 520–6.

Nosofsky, R. M. (1991). "Stimulus bias, asymmetric similarity, and classification." *Cognitive Psychology* **23**: 94–140.

Palmeria, T. J., & Nosofsky, R. M. (2001). "Central tendencies, extreme points, and prototype enhancment effects in ill-defined perceptual categorization." *Quarterly Journal of Experimental Psychology: Human Experimental Psychology* **54A**(1): 197–235.

Pellionisz, A., & Llinas, R. (1979). "Brain modeling by tensor network theory and computer simulation. The cerebellum: Distributed processor for predictive coordination." *Neuroscience* **4**: 323–48.

Plate, T. A. (1995). "Holographic reduced representations." *IEEE Transactions on Neural Networks* **6**(3): 623.

Podgorny, P., & Garner, W. R. (1979). "Reaction time as a measure of inter-intraobject visual similarity: Letters of the alphabet." *Perception and Pyschophysics* **26**(1): 37–52.

Pollack, J. B. (1990). "Recursive distributed representations." *Artificial Intelligence* **46**(1–2): 77–105.

Quine, W. V. O. (1951). "Two dogmas of empiricism." *Philosophical Review* **60**: 20–43.

Rao, R. P. N., & Sejnowski, T. J. (2001). "Spike-timing-dependent Hebbian placticity as temporal difference learning." *Neural Computation* **13**(10): 2221–37.

Rosch, E., & Mervis, C. B. (1975). "Family resemblances: Studies in the internal structure of categories." *Cognitive Psychology* **7**(4): 573–605.

Russell, S. J., & Norvig, P. (2003). *Artificial intelligence: A modern approach* (2nd ed.). Upper Saddle River, NJ, Prentice-Hall.

Schlimmer, J. S. (1987a). *Concept acquisition through representational adjustment*. Unpublished doctoral dissertation, University of California, Irvine.

Schlimmer, J. S (1987b). Mushrooms dataset. The UCI Machine Learning Repository. (Retrieved August 11, 2004, from ftp://ftp.ics.uci.edu/pub/machine-learning-databases/mushroom).

Sejnowski, T. J., & Rosenberg, C. R. (1987). "NETtalk: Parellel networks that learn to pronounce english text." *Complex Systems* **1**, 145–68.

Shastri, L., & Ajjanagadde, V. (1993). "From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings using temporal synchrony." *Behavioral and Brain Sciences* **16**: 417–94.

Shon, A. P., Rao, R. P. N., & Sejnowski, T. J. (2004). "Motion detection and prediction through spike-timing dependent plasticity." *Network: Computation in Neural Systems* **15**: 179–98.

Singer, W., & Gray, C. M. (1995). "Visual feature integration and the temporal correlation hypothesis." *Annual Review of Neuroscience* **18**: 555–86.

Smolensky, P. (1990). "Tensor product variable binding and the representation of symbolic structures in connectionist systems." *Artificial Intelligence* **46**(1–2): 159–216.

Tesar, B., & Smolensky, P. (1994). Synchronous-firing variable binding is spatio-temporal tensor product representation. A. Ram & K. Eiselt (Eds.), *Proceedings of the 16th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ, Lawrence Erlbaum Associates.

Thorpe, S., Fize, D., & Marlot, C. (1996). "Speed of processing in the human visual system." *Nature* **381**: 520–2.

Touretzky, D. S. (1990). "BoltzCONS: dynamic symbol structures in a connectionist network." *Artificial Intelligence* **46**: 5–46.

Touretzky, D. S., & Hinton, G. E. (1985). Symbols among the neurons: Details of a connectionist inference architecture. *Proceedings of the 9th International Joint Conference on Artificial Intelligence (IJCAI 85)* (pp. 238–43). San Mateo, CA, Morgan Kaufmann.

Tversky, A. (1977). "Features of similarity." *Psychological Review* **84**(4): 327–52.

Tversky, A., & Gati, I. (1978). Studies of similarity. E. Rosch & B. Lloyd (Eds.). *Cognition and categorization* (pp. 79–98). Hillsdale, NJ, Lawrence Erlbaum Associates.

Tversky, A., & Hutchinson, J. W. (1986). "Nearest-neighbor analysis of psychological spaces." *Psychological Review* **93**: 3–22.

6 Reduction as Cognitive Strategy

C. A. HOOKER

I

The name 'Paul Churchland' is synonymous in many minds with eliminative materialism, the view that a materialist theory of mind will triumph, not by showing how the mental ascriptions of common sense or folk psychology reduce to neurophysiological conditions, but simply by eliminating the former as a bad theory and substituting its own, different concepts, principles, and data in its place. This looks nothing like reduction, which precisely saves the succeeded within the successor. Yet Churchland understood it as a process akin to reduction. To understand why, his treatment of reduction must be set within his larger vision of cognition and cognitive strategy. In his enthusiasm for a science-led revolution in our reconception of our world, ourselves included, Churchland is pursuing an agenda inspired by early-Feyerabend and his (Churchland's) contribution to reduction theory lies primarily in what he consequently taught us about its place among our epistemological strategies rather than any specific internal technical detail.[1]

Churchland is an unrestrained radical scientific naturalist (see note 3), pushing hard the consequences of understanding this world as a single whole, avoiding the imposition by intellectual fiat of both divisions and unities (e.g, imposing a priori either a mind/body dualism or monism) and accepting only what can be scientifically grounded. That Churchland was an enthusiastic and stylish presenter of it cannot detract from the good intellectual sense with which he explored it and argued its merits. He remains an enthusiastic supporter, without slipping into the tired or triumphal dogmatism that dogs seniority, and we need to take his arguments seriously.

His openness and good sense was what chiefly attracted me to him 30 years ago when I was as an equally young, equally radical naturalist, learned through my prior physics training, and was searching for mutual support within a sea of pervasive, often unconscious, antinaturalism. Physics taught me how easily we form erroneous conceptions of things, even of the seemingly everyday and obvious (like the sun's rising, or the uniqueness of

simultaneity), and of how radically we can re-learn improved conceptions if we apply a powerful enough learning process like science – with enough hard work, and luck! But he, and his equally radically naturalist wife Pat, found (fought?) their way to naturalism through their philosophical training, and to an equal love and respect for science.[2]

II

Within the Western analytic tradition reduction is typically posed as a formal relation between theory-pairs: T_0 reduces to $T_n = T_n$ and 'bridge laws' $B \rightarrow T_0$. (Here T_0 is typically thought of as a cruder predecessor at time t_0 to T_n at later time t_n, and I shall do so, but this is only the common epistemic situation, not a logical requirement, and is not always so.) Under this requirement T_n explains T_0, the laws of T_0 becoming special cases of T_n patterns (not necessarily T_n laws) that arise under the conditions specified in T_n. B is confined to stating correspondences and is logically required to connect the differing properties and principles of T_0 and T_n and should be at least as strong as nomic (lawlike). But since the point of reduction is to achieve explanation in a way that leads to ontological economy (T_0's ontology becomes part of T_n's) the desirable status for B is that of an identity. In that case science can be understood as an historical progression through successively more powerful theories, with past theories being absorbed, explanatorily and ontologically, into their successors, where they live on as special cases. This framework engenders a raft of consequent concerns and forms the dominant locus of debate.

However, it is narrow and formal. One dimension of this concerns the nature of the relations between relevant theories, as we actually find it in science. In the most thoroughly worked out cases, those of physics, the relevant inter-theory relations are all specified as asymptotic relations: $Lim_{p\rightarrow 0}(T_n) = T_0$, for some parameter **p**. This is ultimately a dynamically specified relationship, and one that shows rich variety and complexity (see Hooker 2004). The single abstract formal logical relation of the traditional statement abstracts this character away, and thereby loses key purchase on its nature. A second, related dimension concerns the larger context in which the intertheoretic relations occur. A suite of difficulties with B focuses on theoretical changes that are evidently too radical to capture within B as either nomic or identity correspondence conditions. But these are not the rare exception; rather, a neat, clean reduction relation is the rare exception. Phlogiston chemistry is not a special case of oxygen chemistry; the two

attempt to model chemical reactions in such different ways that we conclude that phlogiston does not exist. And while some concepts and simpler laws of thermodynamics are mirrored in statistical mechanics, much is not and what is, is so only under narrow or idealised statistical assumptions (cf. Hooker 1981, Part I, Sklar 1967, 1993), and so on.

Churchland's wider approach to reduction was already well developed by the time of his first book, *Scientific Realism and the Plasticity of Mind* (1979) and there is still no clearer guide to his thinking, so let us start there. (Hereafter page references to this book are given as '[RPM, n]'.)

The book's overall problem is that of "reckoning just what the objectivity/integrity/ rationality of the intellectual process consists in" [RPM, 3]. By page 3 he has already canvassed the difficulties of the classical empiricist solution – that it consists in incorrigible observation and logical inference within a commonsense conceptual framework – and we soon discover that he intends a radical solution along Feyerabendian lines that will involve re-conceiving most of philosophy, including reduction – indeed, reduction plays a key role in his scheme of re-thinking how intelligence works. Speaking of the traditional privileged-commonsense/ parasitic-theoretical dichotomy and the accounts of perception, scientific knowledge, reason, and the like that derive from it, he says:

> This [general position] now appears to be impossible [to maintain]. The failure of the various accounts proposed forms part of the reason, but the main consideration is rather more interesting. The premiss on which the older account is based . . . appears to be false. . . . If viewed warily, the network of assumptions and principles constitutive of our common-sense [sic] conceptual framework can be seen to be as speculative and as artificial as any overtly theoretical system. It even displays fluidity if one takes a decently long view. Comprehensive theories, on the other hand, prove not to be essentially parasitic, but to be potentially autonomous frameworks in their own right. In short, it appears that all knowledge (even perceptual knowledge) is theoretical; . . . Our common-sense conceptual framework stands unmasked as itself being a theory, or a battery of related theories. And where before we saw a dichotomy between the theoretical and the non-theoretical, we are left with little more than a distinction between freshly minted theory and thumb-worn theory whose cultural assimilation is complete. [RPM, 1–2]

This is indeed a radical shift. Behind the traditional view stands a conception of a priori normative philosophy setting out the primacy of the commonsense life-world and logic-based accounts of metaphysics, reason

and knowledge, and so on, which work out from it to show our science and other activities as instrumental extensions of that life-world, useful in practice but subordinate to commonsense, in turn subordinate to a priori normative truth. In vast contrast, behind Churchland's alternative vision is a learning natural human mammal, finite, ignorant, and fallible, but hugely creative, actively exploring its world, spinning out trial conceptual frames as it goes with which to construct the trial methods and theories that are its cognitive tools, the better to organize its activities for survival and flourishing. These speculative constructions include perceiving and reasoning practices and all philosophical theories of these, and thus turn the traditional conception on its head. Rather than being driven from high a priori philosophical knowledge 'down,' it is the naturally embodied creature exercizing its finite, fallible cognitive capacities that drives the whole historical exfoliation of our speculations, from practical exploration 'up' to philosophical theorizing. And it is the proper fluidity of all and every aspect of these in the face of our manifold kinds of environmental interactions that is the powerfully creative plasticity of mind confronting the real world, of which his book title speaks.

"The function of science, therefore, is to provide us with a superior and (in the long run) perhaps profoundly different conception of the world, even at the perceptual level" [RPM, 2]. This raises "the matters of cross-theoretical comparison and intertheoretic reduction. These raise in turn the more general issues of meaning, translation and conceptual change" [RPM, 4]. So, reduction is nested within translation across conceptual frames generally as these are successively transformed by science. Reduction arises in the context of the general relations between whole working understandings, including their concepts, principles, theories, ontologies, methods (including perception), and data.

In these circumstances it must be expected that some agreements between older and newer frames will be closer than others. While in the analytic approach reduction requirements are typically focused solely on intertheoretic derivability, this discussion makes it clear that much more is involved, for successor frameworks will often involve new concepts and principles that must be internally coherent and explanatorily comprehensive, they will often re-describe the so-called data in very different ways and invalidate some older data, and they will often disable preceding methods and require the development of new ones (cf. Aristotelian and Galilean/Newtonian frameworks, see e.g., Feyerabend 1962, Brown 1988). In all cases, however, we seek an explanation of why one frame should be preferred over another.

The general answer, to a first rough approximation, is that preference goes by explanatory power: frame $F(t_n)$ is preferred to frame $F(t_0)$ exactly when $F(t_n)$ either explains $F(t_0)$'s explanatory capacity, or explains $F(t_0)$'s explanatory failures, or both, but not vice versa. In an older terminology, $F(t_n)$ either explains $F(t_0)$ or explains $F(t_0)$ *away* by explaining why it worked as well as it did (see Sellars 1962, [RPM, 45]). Explaining $F(t_0)$ is most complete when $F(t_0)$'s laws and data patterns are wholly reconstructible in $F(t_n)$ – when $F(t_0)$ is reduced to, and retained within, $F(t_n)$. Explaining $F(t_0)$ is minimal when $F(t_0)$'s laws and data patterns are wholly distinct from $F(t_n)$ – when $F(t_0)$ is irreducible to $F(t_n)$ but is explained away by $F(t_n)$, eliminating $F(t_0)$. These retentive and eliminative cases form the opposite extremes of a continuum of intermediate cases. In physics, for example, the expression of the relevant inter-theoretic relation takes a particular mathematical form, the asymptotic relationship, for example (and roughly) $\mathrm{Lim}_{1/c\to 0}$ (special relativity) → Newtonian mechanics, $\mathrm{Lim}_{\lambda\to 0}$ (wave optics) → ray optics, $\mathrm{Lim}_{h\to 0}$ (quantum mechanics) → Newtonian mechanics, and so on. The exact retrieval of a predecessor theory in the asymptotic limit rarely happens, there are typically formulae on both sides that don't 'match up' asymptotically; so almost all cases, certainly all the central cases, strictly lie in an intermediate position between the extremes.

Our conceptions, old and new, of our own mental capacities reveal a similar situation. Mind/body reduction poses (but in an older mental-substance language) the issue of where the relation between older and newer agency theories falls on the retention-elimination continuum. Here Churchland contends that good new science leads us far away from the older commonsense belief-desire-action conception (RPM, Chapters 4, 5). The primary modelling tool, he argues, is that of connectionist neural networks, and he has devoted two books to exploring this framework (Churchland 1989, 1995). The older agency framework is to be replaced (or displaced, [RPM, 44, 81–2]) root and branch, with this best scientific version.[3]

And now our earlier puzzle is also solved and we can see how eliminative materialism is related to reductive materialism: both are possible versions of the mental cross-frame relations and, while as extreme versions of that relation they oppose one another in formal detail and consequent ontological retentiveness, they share the same overall strategic relationship to nonmaterialism, for both agree (as will intermediate alternatives) that commonsense agency theory is to be superseded by a scientific version (currently proving to be materialist).

With this general orientation in hand, we can profitably continue exploring Churchland's exposition of reduction in RPM, in particular Section 11

[RPM, 80–8]. Caveat: Despite earlier using the term 'reduction' in the traditional retentive sense, for example at [RPM, 5, 43], as Section 11 progresses its usage is slowly widened to eventually mean interframe displacement generally and hence to cover all intertheoretic relations along the retention-elimination continuum. The section begins using the term "ideal reduction" for retentive reduction and moves to consider those that "fall short" of the ideal [RPM, 83]. Thereafter to signal the retentive form Churchland uses expressions such as "maximally smooth reduction," but this makes an implicit claim about a mathematical asymptotics basis to intertheoretic relations that he never discusses, let alone defends (for this see e.g., Batterman 2002, Hooker 2004), whence I shall retain here the ontologically clearer terms "retentive," "eliminative." I shall use Churchland's term "displacement" to refer to an intertheoretic relation anywhere on the entire spectrum, with "replacement" as its action synonym.

The opening paragraph concludes with the claim that (retentive) reduction largely fails and that how it does and why it matters are the concern of Section 11. Then Churchland launches immediately into the view that reduction, like translation, effects a mapping between two sets of sentences, but with something weaker than meaning required to be preserved since intertheoretic synonymy between terms largely fails [RPM, 81]. The weaker requirement is never explicitly spelled out, but Churchland's discussion strongly suggests that it is something like explanatory capacity. For instance, speaking about the Classical Mechanics-to-Relativity [hereafter CM/ R] reduction, he raises the question of how we can validly decide whether we are experiencing, for example actually measuring, in the terms of the one theory or the other and says [RPM, 87]: "The resolution of such ambiguities does not proceed by appeal to any simpler or more neutral set of data, observational or otherwise. It proceeds by determining which of the two conceptual alternatives allows us to construct [in its terms] a consistent and coherent account of our experience as it is pressed farther and farther into unfamiliar domains." And on this score it is Relativity that has the larger explanatory capacity, for it can both more finely explain all CM phenomena, and CM's lesser accuracy, and accurately explain high velocity phenomena where CM explanations fail [RPM, 87–8].

Churchland's purpose in what follows is to develop this explanatory point; commencing from the retentive extreme he slowly pushes the envelope wider until he has established the explanatory strategy for the entire spectrum. The discussion begins by identifying what a successful retentive reduction provides: It provides a set of theorems of T_n as the mirroring objects of T_0's central principles, thereby insuring that the displacement

is orderly in that "T_n will cohere with the background [the "larger network of background beliefs"] in the same ways that T_0[4] did, and will perform all the same predictive and explanatory functions that T_0 performed" [RPM, 82]. Whence it "locates the newer theory [T_n] within the conceptual space currently occupied by the older theory" [T_0] [RPM, 81] through the constructed mirroring of old descriptions in the new terms.[5] It follows that "T_n can lay claim to confirmatory considerations systematically analogous to those that have already accrued to T_0" [RPM, 82]. Note that the claims are only said to be analogous; this is because the T_n mirroring expressions may not mean the same as their T_0 counterparts and because exact mirroring may only occur under 'idealized,' strictly nonexistent, conditions – both as illustrated, for example, in the CM-R case previously mentioned.

But then, Churchland concludes, "what a successful reduction shows us is that one way of conceiving things can be safely, smoothly, and – if the excess empirical content of T_n . . . is corroborated – profitably displaced by another way of conceiving things. And this, I submit, is the function of reduction" [RPM, 82]. Note the ambiguity in the last claim, it can be read to apply to just the retentive reductions being discussed, or to include more eliminative intertheoretic relationships, since these latter can equally displace their predecessors, and in a systematic way. The ground for this extension has already been partially prepared with Churchland's noting that the T_n mirroring sentences for T_0's principles need not be "semantically or systematically" important, and by calling T_n's confirmation in the T_0 domain only analogous to, rather than identical to, that of T_0 (the usual claim). Once it is accepted that the mirroring counterpart requirements can be weakened in this way it is a short step to weakening them to mirroring of principles only under some conditions, or mirroring only select empirical generalizations under their obtaining conditions, or only to within experimental errors, and in these ways encompass the eliminative cases as well. As retentiveness decreases the T_n image of T_0 becomes less and less like T_0 and merely analogous in certain respects. It will be the closeness of the analogy that determines the character of the reduction (the analogy being closest when it is an isomorphism).

In thus focusing on T_n imaging of T_0, rather than on deducing T_0 from T_n, Churchland takes a position that unorthodoxly plays down the role of the bridge laws that establish the correlations or identities between the two theoretical domains. In the orthodox discussion these connecting principles are taken as crucial, but in Churchland's view they are derivative. True, he speaks easily enough of bridge laws (e.g., [RPM, 81]), but there it is a successful reduction that yields *them*, not vice versa. How can this be?

It is certainly undeniable that the deduction of T_0 laws from T_n logically requires bridging laws or identities, as Nagel made clear when developing the deductive model of reduction (Nagel 1961). Since then the standard position has been that the 'bridge laws' (/identities) lie at the heart of a reduction.

But they do not lie at the heart of real reductive practice in science. By considering how scientists actually treat the reduction of scientific laws, for example, the reduction of the Boyle-Charles law to statistical mechanics, Nagel himself shows how scientists arrive at reduction of a law L_0 or property P_0 of theory T_0 respectively to a law L_n or property P_n of theory T_n by first showing how to choose conditions (real or idealized) under which it is possible to construct the imaging (here mirroring) relation and from that *deducing* that the reduction is shown achievable through the postulation of the relevant laws or identities. Churchland declares that a "successful [retentive] reduction . . . provides an excellent reason for asserting the relevant cross-theoretical identities, the best reason one can have" [RPM, 83]. In effect the bridging identities are deduced from the assumption of the reduction, rather than vice versa (Churchland 1985b, 11, cf. Ager and Aronson 1974), being asserted on the basis of achieving the reduction, supported in that light by appeal to Occam's razor and, where available, by claims of spatiotemporal coincidence. This is the position also taken in Hooker (1981, see explicitly Part I, 49), both commitments emerging from our mutual discussions over the preceding years (though I am sure I learned more from him than vice versa about reduction as a cognitive strategy).[6]

But Churchland's displacement perspective takes him still farther from orthodoxy here:

> . . . cross-theoretical identity claims, even if they are justly made, are not a part of the reduction proper, and they are not essential to the function it performs. The correspondence-rule pairings need not be construed as identity claims, not even as material equivalences [correlations], in order to show that T_n contains an equi-potent image of T_0. In fact, we can treat each correspondence rule as a mere ordered pair of expressions . . . and we will then need only the minimal assumption that the second element of each pair truly applies where and whenever the first element of each is normally thought to apply. Such an assumption, note, is strictly consistent with the idea that the first elements (the expressions of T_0) do not apply to reality at all [RPM, 83].

Whence a true theory can displace a false one. Put thus it sounds trivial, indeed a paradigm of progress; put as "a true theory can reduce a false one"

[RPM, 84] it sounds mad unless it is remembered that we are no longer reading reduction purely retentively and elimination/explaining-away are equally part of the spectrum of reductions in this wider, weaker sense.

Now Churchland begins exploiting the latitude in this wider, weaker sense, expanding the compass of reductive relationships so as to move beyond the purely retentive cases.

> First, the image, such as it is, of T_0 within T_n may not be a complete or wholly faithful image of T_0. It may be, for example, that one or two of the important principles of T_0 are mapped on to sentences in T_n that are just false . . . [or] . . . several of the principles of T_0 must be modified or "corrected" in some way . . . [Then] . . . it is not . . . T_0 that finds an appropriate image within T_n but rather some theory T_0' closely similar to T_0.
>
> Second, it may be that the image S_n of T_0 within . . . T_n is not an unaided consequence of the basic principles of T_n [but] . . . may be derivable within T_n only if we include some limiting condition or counterfactual assumption. . . . If the assumption in question relevantly approximates the domain in which T_0 has hitherto been successfully applied, the result will still be counted as a reduction. Here it is . . . not T_n that provides an equipotent image of T_0, but rather an augmented theory T_n' closely similar to T_n. And finally, some cases [of reduction] may instance both of these complications, T_0 being related to T_n only through some T_0' and T_n'" [RPM, 83–4].

These last cases will still count as reductions, claims Churchland, if they show how T_n could "more or less smoothly and painlessly" displace T_0 [RPM, 84].

But then if "we broaden our concept of reduction to include [these] . . . important cases . . . we must be prepared to count reducibility as a matter of degree." And the less smooth reductions may be as interesting "for what they fail to preserve [what they explain away], and more valuable for the revisions they dictate" [RPM, 84]. If we care about understanding through explaining (including explaining errors = explaining away) this is exactly how we will often judge. Whence "a reduction is *not* essentially reaffirmative or vindictive with respect to the categories and principles of the reduced theory" [RPM, 84, Churchland's italics]. This is because there is so much conceptual transformation now inserted between them: "For the falsity, even the radical falsity, of T_0 need not preclude its having some large fragment or modified version T_0' that finds an appropriate image S_n within some stoppered-down version T_n' of T_n" [RPM, 84–5].

In effect Churchland has shifted discussion from the specific *explanatory/ontological* requirements of strict retentive absorption to the broader

functional role requirements of an adequate explain/ explain-away successor, retentive or eliminative. The requirements for successfully explaining something are much stronger, place more coherent constraints, than do those for explaining something away. As Churchland notes, if T_n fully explains T_0 then T_n subsumes T_0's ontology and principles and the roles that these thereby continue to play has to be explained, or explained away, by all subsequent theories. But if T_n explains T_0 away the conditions for this may be as context-specific and idiosyncratic as need be, varying from case to case, since what were laws within T_0 no longer have to be treated as laws within T_n explanations, they might instead be treated as accidental or artifactual correlations, or as just illusory. To be sure all these conditions continue to have to be explained or explained away by all subsequent theories (just how did the accidental correlation arise?, etc.), but this constraint no longer has the bite that the retentive requirement does. As Churchland's discussion makes clear, displacement is more concerned with the formal mapping relationship involved than with ontological identification.

This shift has a dual importance for the approach to reduction: it marks a distinctive withdrawal from some traditional logical concerns in the conception of reduction and a correlative re-focusing on inter-theory mapping relations generally. Both shifts bring distinctive benefits with them.

These claims are nicely illustrated in recent work by Bickle (1998, 2003). The story begins with a subtle, but significant, difference between the displacement mapping approach and Schaffner's (1967) modification of Nagel's (1961) schema (cf. Schaffner 1992). Schaffner, equally aware of the many imperfect reductions in science, allows that, not T_0, but an analogous theory T_0^*, couched in T_0's terms, is what can actually be deduced from T_n and bridge laws and that it is the closeness (or otherwise) of the analogy that determines the retentiveness of the reduction. The two approaches share some obvious features: an intertheoretic comparison made through a deduction and a consequent analogy whose closeness (or otherwise) determines the retentiveness of the reduction, with adequate displacement as the general goal and explanatory adequacy as a primary criterion. The difference between these positions lies in what is deduced: an analog in T_n terms (Churchland-Hooker), with bridge laws derivative, versus an analog in T_0 terms (Schaffner) relying on use of strict bridge laws.

Bickle has made much of the significance of this difference in approach, citing a variety of advantages for the Churchland-Hooker version.[7] One important advantage has already been canvassed: release from having to find connecting bridge principles (laws or identities) to obtain any intertheoretic reduction relation, thus by-passing their numerous difficulties.[8] There

follows the advantage, noted above, that where the reducing theory corrects the reduced theory, including radically so, intertheoretic relations are dealt with in the same general way as the clearly retentive cases, while the Schaffner model will have difficulty providing coherent bridge principles since the reduced and reducing theories are logically incompatible. A further advantage is its amenability to a semantic model theoretic formulation, inherently more powerful than the Schaffner-style syntactic formulation. This provides a richer structure to the analog relation and thus can provide a finer set of distinctions with which to theorize intertheoretic relations. This extension of Hooker (1981) was developed by Bickle (1998) and labelled "new wave" reduction theory.

However, we need to keep abstract theorizing in perspective. Bickle's model theoretic apparatus does formalize some useful distinctions, ones that help to flesh out the account in Hooker (1981) both in respect of the general analog relation and its application to function-to-dynamics relations in particular. Bickle rightly says that in Hooker (1981) I make no attempt to theorize the analog relation and have no "deep yet simple insight" to offer into how to distinguish in a principled way genuine reduction from mere historical succession.[9] I took this stance because of the rich and subtle variety of inter-theoretic relations that science throws up (for just physics, see Hooker 2004). I still think they defy understanding within our present knowledge and that Bickle's apparatus does not much help here. Instead, in physics it is asymptotic analysis that has provided the insight and it will be a corresponding analysis of dynamical input/output maps in relation to neural assembly self-organization that is likely to provide neuro/ psychological insight. I note that Bickle in his valuable (2003) has turned in a similar direction for further neuroscientific understanding.

The important point for the present context is that the advantages of the T_n analog approach ultimately derive from the way it allows the treatment of reduction to expand from focusing on the specific requirements of strict retentive absorption to focusing on the broader requirements of an adequate displacing successor, retentive or eliminative. This is where, for example, asymptotic analysis enters, since it covers all the major cases in physics. This is just the dual face of Churchland's shift of focus, and it brings its own important benefits. For instance, now "a reduction may even pinpoint for us the way or ways in which the reduced theory is cockeyed. We need only examine the details of its deviation from an ideal reduction: the assumptions that had to be made to squeeze a suitable [T_0 image] S_n out of T_n, and the points in which T_0 differs from the closest legitimate S_n we could find for it" [RPM, 85]. The point applies well to the relationships in physics.

For instance, in the 'smoothest' case, CM/ R, the relativistic expressions can all be expanded in terms of the parameter 1/c (so $(1 + v^2/c^2)^{1/2} \approx 1 + v^2/2c^2 + O(v^4/c^4)$, and so on) in a way that reveals how the relativistic and Newtonian expressions differ in structure for $v \neq 0$, strictly coincide only in the dynamically degenerate limit $v = 0$, but nonetheless then yield $S_n = CM$, something that does not happen in the other prominent cases (Batterman 2002, Hooker 2004).

Scientists focus on displacement because they have to be concerned with continuing practical reliance on T_0 and hence with the errors involved in doing so, which asymptotic analysis reveals, while philosophers are concerned with accurately specifying ontological commitment in the light of T_n's explanatory superiority. Putting it this way shows how both physicists and philosophers are concerned with how things "fit together" and how both can pose this in terms of relations between successor and predecessor theories. But the difference is important. A formal morphing relation always holds between two successor dynamical theories, wherever their relationship lies on the retention/ replacement spectrum. But retentive reduction does not always hold. Thus we have a universal explanatory role for the formal morph relations, coupled with an asymmetric failure of the ontological schema – of retentive reduction. Asymptotic regularity specifies smoothness of error increase while retentive reduction specifies ontological identification.[10] Churchland is well aware of this but, lulled by his smoothly developed transition, less attentive readers might not be.

III

There is another way in which asymptotics is crucial to reduction: it grounds a dynamical account of self-organization proper and this in turns grounds principled dynamical accounts (I) of emergence, (II) of levels proper within complex systems and, on that basis, (III) of cross-level causal (more generally: dynamical) relations, leading to a principled dynamical account of causal multiple realisation, and (IV) of cross-level functional-to-dynamical relations, leading to a principled dynamical account of functional multiple realization, and (V) of context-dependent laws, and hence of the so-called 'special sciences,' such as biology and social theory.

A detailed explanation belongs elsewhere (see Hooker 2004) but the essential idea is that self-organization proper is the asymptotic amplification of a system fluctuation so as to dominate all components, forming a new top-down, dynamically stabilized constraint on them and altering

not only their behaviour but also their very dynamical form (the form of the dynamical equations which govern them), see Collier/Hooker (1999). (I) This creates a new dynamical existent (though not necessarily any new components), for it brings about the only thermodynamic condition under which new work can be performed, and so defines emergence proper (as opposed to mere new pattern). (II) The new dynamically stabilized constraint defines a level proper (all else is mensural, metaphorical, or muddled talk). (III) Each such level must be characterized by causal multiple realization since, being stabilized, a level dynamically filters out sublevel dynamical fluctuations. For instance, a metastable ("triggerable") switch, like a Geiger counter, involves a stabilized system constraint (e.g., counter construction) that insures that otherwise heterogeneous dynamical subsystem conditions of a certain class (e.g., counter chamber ionizing events) will produce the same effect (e.g., a counter pulse) with all other fluctuations filtered out. (IV) The dynamical imaging of functional capacities goes á la Hooker (1981), but with it now clearer what the roles of constraints are (cf. Geiger counter constraints for 'detects ionizing radiation') and how the application of determinate/determinable treatment confines the unavoidable multiple realization to within naturalisable determinate/determinable hierarchies where they are neutral to reduction issues.[11]

(V) Emergent constraints give rise to the context-dependent laws that mark the so-called special sciences – except that they are not special in this regard at all, the triggering law for Geiger counters, for example, only applies to those contexts where a Geiger counter constraint obtains and emerges with the emergence of that constraint during Geiger counter construction. So it is with all the manifold dynamic constraints characterizing living systems. Here we find very complex patterns of constraint-dependent dynamics generating complex histories of emergent phenomena marked by strong dynamical fixation of historical constraints where dynamical form may change as a system – including even just its initial conditions – changes. This means that it is impossible to form simple first-order laws about such domains, as is the norm in basic physics and chemistry; rather all such laws become strongly constraint-, and so historically-, dependent. In addition, because of their internal dynamical diversity these systems will often display diverse responses to the same situations, increasing the complexity. These twin heterogeneities provide the core reason why special sciences occur, are to that extent irreducible, distinctive and analytically complex. Yet these conditions too are ultimately explained by fundamental dynamics, one that grounds multiple reductions interwoven with them (cf. again the Geiger counter). The upshot is that dynamics grounds a unity

to scientific laws but this unity is internally very complex (see also Hooker 1997).

Applying these features reveals the dynamical organization of the richly layered suite of metastably context-dependent laws that we know as human individual and social behavior. At present, of course, we are incapable of more than hugely superficial sketches of fragments of it. Here Churchland has done us a large favor by exploring in a pioneering way something of the consequences of taking a crudely neural assembly approach seriously, in the form of simple connectionist nets.[12] In this domain, which has been his major preoccupation, he has been content to apply the general approach to reduction rather than develop any new theory for it. In RPM, Sections 15 and 16, for instance, he uses the general displacement framework to provide an illuminating sketch of the landscape of potential theories of mind as so many ways of displacing (including retentively) our present 'folk psychology' and/or future scientific psychology.[13] And there and throughout his subsequent work he uses that approach to argue forcefully for eliminative materialism, for example, against non-naturalist construals of the propositional attitudes, sensory experience, introspection, consciousness and so on. In all this he maintains, as any naturalist should, a fallibilist, empirical stance, allowing future research to determine the fate of his position.[14]

On this score, the consideration of asymptotics, and self-organized constraints in particular, begins to provide a scaffolding for filling out Churchland's approach. It has seemed to many that one serious problem with eliminativist reduction is its 'brute force' approach, the succeeded theory is simply discarded and no real explanatory appreciation of it is required. This has seemed particularly galling in the case of folk psychology. Of course, strictly, this is not fair criticism. In principle Churchland acknowledges the need to 'explain away' the erstwhile practical acceptance of such theories and even anticipates the use of asymptotics to guide conceptual analysis (see the [RPM, 85] quote). You can see in the simpler cases physics provides how this goes; for example, analysis of relativistic concepts shows precisely how the concepts modify (see Section II). But in practice little explaining away has been done in the case of living systems. It is easy to say why in general: the multiple interacting levels of living organisms has scarcely begun to be understood. But this is solely negative apology. We can now also at least sketch (here very briefly!) how a future respectful explanatory analysis might run.

Organisms are basically characterised by a complex global dynamical constraint that can be called autonomy (see Christensen & Hooker 2000). This constraint is powerful, requiring a metabolitic basis to life (Moreno

and Ruiz-Mirazo, 1999), but also grounding the basic features of all agency. (Churchland lacks the use of self-organized constraints in reduction as outlined above, and specifically the use of autonomy-style models of agency, so he has not had access to this approach; but he is not opposed to it – see Christensen and Hooker 1998 and Churchland's response.) It is the autonomy constraint that ultimately allows us to treat organisms as stable individuals and underlies the internal coherencies of character they display. In the interactivist-constructivist conception of organisms (see Bickhard 1993 and references, Christensen and Hooker 1999) management of interaction with the environment is the fundamental basis of understanding the evolution of intelligent capacities. The social environment is fundamental to human development and capacities.

Combining these themes, we understand agency concepts as implicitly relying on autonomous organization to allow us to ignore most of the internal complexity of organisms and develop a simplified practical set of procedures for managing social interaction, namely the commonsense belief-desire-action conception. To understand just how such concepts are simplifications of reality would require following them back through the autonomy-level across myriad asymptotic transitions to the full richness of neurophysiological and metabolitic dynamics involved, something still well beyond our ability, but no longer beyond our conception (cf. Christensen and Hooker 2001 on intentionality). Language is also at bottom a behavioral tool for modifying social interaction (see Bickhard 1993 and references). It leads to a very different conception of conceptual semantics (see also Christensen and Hooker 1999, Section 5), a cohering way to understand conceptual modification, and provides deep reasons not to rush to philosophical conclusions from its surface analysis. Thus can eliminativism begin to provide a more satisfying treatment of folk psychology.

IV

In conclusion, the examinations of the previous sections all support the overall contention that, with respect to reduction, while eliminativism can be enriched, extended, and deepened, this is largely left to others and Churchland's primary legacy lies in the rich sense of cognitive strategy he has been able to convey. Churchland's vision is of an unfolding scientific understanding that penetrates every area of life, from our social morality to our personal consciousness. Science itself emerges under our creation

from the primordial species capacities for exploration and adaptation and is itself unfolding as at once part of, and generator of, that understanding. In this grand but scientifically based view of a dynamic history Churchland belongs, in this respect, in the recent tradition of Jared Diamond's *Guns, Germs and Steel* and Tim Flannery's *The Future Eaters*. The primary subject domain for expressing Churchland's vision has been the understanding of mind. In this, reduction as a central part of cognitive strategy finds its proper and well utilized place.

Notes

1. See especially Feyerabend 1962 for reduction morphing into replacement as intertheoretic disparities increase, particularly for mind/body reduction; cf. Bickle 1998, Chapter 2 and 2003, Chapter 1.
2. To Churchland's 1975 compare, e.g., Hooker 1974 and subsequently our joint Churchland and Hooker 1985. Churchland's work culminated in his first book, 1979 (whose Preface makes kind reference to Hooker 1975). My complementary investigations resulted in my quatrology on reduction, Hooker 1979, 1981, a first serious (if incomplete) attempt to replace purely empiricist/analytical formal-logical criteria with more dynamical criteria in philosophy. In this critique of empiricism we had superb, if minority, support; both of us were broadly influenced by Feyerabend, e.g. 1961 [copy available on request], 1970, and by Sellars, especially his 1962, 1965 (indeed Sellars taught Churchland and was my 1970 Philosophy thesis examiner) and abetted by insightfully rebellious contemporaries, e.g., Brown 1979.
3. At this point in their developmental trajectories Churchland and Feyerabend start to diverge. Although continuing to emphasise its cognitively transformative nature, the later Feyerabend increasingly sought to de-throne science as cognitively distinctive or epistemically privileged, seeing it instead as one tradition among many (e.g., Feyerabend 1987). At one point Churchland flirted with a related project of abandoning the privileged role of truth in cognitive enterprise (Churchland 1985a), but seems not to have pursued it, continuing instead to explore a scientific naturalist account of mind. It was precisely at that abandon-truth point that I counted myself a restrained naturalist, since I argued that our scientific image of ourselves, especially the biological picture of agency, has underwritten a substantive role for truth, not undermined it. (See Hooker 1987, Section 8.4.3.9, 1995, Section 6.II.6 for some preliminary remarks on this issue.) Interestingly, it is also possible to read Feyerabend in a similarly more conservative way (Hooker 1991, Farrell 2003), and this is perhaps to be preferred (pun intended).
4. The text mistakenly has "T_n" here.
5. One might suppose instead that the reverse would be a truer rendering since T_n is *ex hypothesi* a more explanatory successor theory to T_0. But while this

claim is also true, Churchland is here interested in making the point that, if T_0 is integrated into our background beliefs (is perhaps 'commonsense'), then a retentive reduction does not challenge that background. Even so, he should have added, "within the domain of T_0." For T_n may yet challenge our background beliefs outside the T_0 domain. This is precisely what happens with the CM/R case; R reduces to CM in the limit $1/c \rightarrow 0$ and so behaves 'Newtonianly' in our normal low-velocity domain, but *outside* that domain R supports behaviour that is bizarre to commonsense (time dilation, etc.). The point is not that this makes it impossible to extend our background beliefs (including commonsense) to the larger R domain, that is understood, the point is that the bizarre occurs to erstwhile commonsense entities that can enter the wider domain (twins, for instance) and this must either destabilize to some extent our conviction in the complete adequacy of our background (commonsense) conceptions or force us to set up an antinaturalist, entirely human-imposed dichotomy between background and nonbackground belief 'worlds.'

6. More recently, Beckermann (1992) and Marras (2002) restate the imaging approach. Marras shows how this understanding both invalidates the earlier criticisms made by Kim of Nagelian reduction (Kim 1998) and demonstrates that Kim's most recent functionalising approach (Kim 1998, 1999) is after all essentially equivalent to Nagel's. Marras' most pointed criticism of Kim's functionalising position is its inability to coherently resolve the problem of multiple realisability. In Hooker 2004 I show how a dynamical analysis centred on self-organization as constraint formation and applied in the context of the mirroring condition will allow us to do decisively better.
7. As to which approach is more truly Nagelian, the point is not only irrelevant to best belief formation, it remains moot as well. Both share a Nagelian core and both modify Nagel's original schema. More importantly, on the crucial difference between them Nagel is ambiguous: he speaks theoretically of bridge principles in Schaffner's manner but, as noted, presents the only real scientific case he discusses (the reduction of the Boyle-Charles law within the thermodynamics/ statistical mechanics relation) essentially in the Churchland-Hooker manner.
8. However, some problems with intertheoretic correspondences must be met head-on, the most important concerning reduction of properties. Currently there is the widespread assumption that the proper identity criterion for properties is synonymy and that property identity is thus necessary or nonexistent. If this assumption is adopted then it will be impossible to find identities even for the most retentive reductions (trivial ones aside) because even when the T_n image mirrors T_0 isomorphically the 1:1 matched property pairs will typically still not be synonymous, cf. CM/R. But there are good cognitive reasons not to hold this position and instead to adopt a same-dynamical-role criterion of property identity (see Hooker 1981, Part II), which can be contingent, and this dissolves the problem. Moreover, a biologically grounded approach to cognition provides a supporting conception of semantics (see Christensen and Hooker 1999, 2001) quite different from any that might encourage the ignoring of dynamical grounding of property semantics evidenced in the synonymy criterion.

9. Bickle also takes me to task for not offering a detailed specific scientific reduction formulated in the T_n analog form. I did not do this because examples from science were readily available in scientific textbooks and always took this form – even Nagel's own case summary did so. Nor, as Bickle notes, did I tackle offering a detailed example from neurophysiology/psychology. That was because it would have required acquiring several specialties and writing a book – as Bickle's own discussions, especially his 2003, show. (Credit to him for not just complaining, but writing the book his complaint required.)
10. The two issues should not be conflated: that equation Ψ deforms into equation Φ under some parametric constraint, e.g., $Lim_{x\to 0}$, an *asymptotic* relation, has in itself no bearing on the retentive/eliminative character of the *reductive* relation between the theories expressed by Ψ and Φ. This is illustrated by considering that, given the relation $\Psi = \Phi + x\varphi$ (where Ψ, Φ and φ are functions, e.g. differential equation models of dynamics, and x a parameter), Ψ morphs into Φ in $Lim_{x\to 0}$ divorcing Ψ dynamics from φ dynamics irrespective of the nature of φ (so long only as it remains finite) and thus irrespective of whether φ provides for, or prevents, a larger reduction relation between Ψ and $\Phi + x\varphi$. This point, made long ago (Hooker 1979) against Yoshida (1977), bears repeating, for Batterman also falls foul of it (Hooker 2004).
11. Always assuming that syncategorimicity and context-dependence are fully applied as per Hooker (1981), cf. Hooker (2004). And while noting that no dynamical analysis can build a bridge to etiologically characterized functions because, not being groundable in current dynamical states, they are also neither accessible to, nor dynamically efficacious for, their possessors; on that score alone they should be dropped – cf. Christensen and Hooker (1999, 2001). That leaves for consideration only a multiple realisability so radically heterogeneous that it violates confinement to a dynamically matchable determinate/ determinable hierarchy even when dynamically emergent constraints are included – the sort of anomaly often attributed to mental, social and other agency properties. But science has evidently not needed to recognise any cases of this sort within its most mature disciplines (physics, chemistry) and, given (V) subsequently, I submit we have no reason to suppose it will need to do so in future.
12. See especially Churchland 1989, 1995 and elsewhere herein. In my view there are important limitations to present connectionist models that run deeper than their obvious disanalogies to neural structure (of which Churchland is, of course, well aware). Chief among these is the limitation on self-organization – on changing dynamical form (not just state) – that their invariant structure imposes. It corresponds to the problem of understanding fluid dynamical stabilization within neural nets. Consequently, I believe we must ultimately regard connectionist models as 'half-way houses' to fully dynamical models (cf. Hooker 1996).
13. Here Churchland put his general retention/elimination framework to work to generate a spectrum of possible mind/body positions as a function of what kind of future scientific psychology might displace our present intuitive agency theory (aka 'folk psychology'). He considers three dualisms and three materialisms. Interestingly, the list is not well structured. The six alternatives are neither mutually exclusive nor exhaustive. The alternatives are not mutually exclusive

because eliminative materialism is crucially nonspecific about the nature of what does and does not match up. The list is relevantly incomplete because missing from the materialist alternatives is quasifunctionalist, quasiidentity materialism, reductive (Armstrong 1968) and semireductive (Hooker 1981, Part III). This is a nontrivial omission, since it is the position delivered by syncategorematic construal of mental (and most social and biological) attributions. But Armstrong unfortunately made schematic use of the alleged gene-DNA identity as a model for mind/brain identities, treating the reduction as a uniform micro reduction governed by entity identities (Causey 1977) as does Schaffner (1967). This is a fundamental mistake, for both genes and minds, because it ignores the syncategorematic, context dependent and self-organizational nature of the relationships it obliterates, and hence the complex systems nature of the features involved. Kim's (1998, 1999) later version is also crucially less dynamically based and so also fails to adequately comprehend multiple realizability and like aspects of complex systems; see Hooker 2004, cf. Marras 2002. Note that a functional theory is not directly asymptotic to a dynamical one, but to a dynamically faithful functional one.

14. He even allows for the possibility that thinkers or intelligent agents may not form a substantial natural kind if the heterogeneity of their physical realisations is sufficiently radical. In this case, he suggests, we may be left with an abstract kind alone, as 'arithmetic calculator' is. This is, I think, a far too generous concession to abstract functionalism of the Putnam sort. A deeper biological understanding of the roots of agency welds metabolitic organisation to agency in much deeper ways than this, see especially Moreno et al. 1999 and Christensen and Hooker 2000, cf. Christensen and Hooker 1998 on Churchland.

References

Armstrong, D. M. (1968). *A Materialist Theory of Mind*. New York, Humanities Press.

Ager, T. A. and Aronson, J. L. (1974). "Are bridge laws really necessary." *Noûs* **8**: 119–34.

Batterman, R. W. (2002). *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction and Emergence*, Boston, MIT Press.

Beckermann, A. (1992). "Supervenience, emergence and reduction." In A. Beckermann (ed.), *Emergence or Reduction?* Berlin, De Gruyter.

Bickhard, M. H. (1993). "Representational Content in Humans and Machines." *Journal of Experimental and Theoretical Artificial Intelligence* **5**: 285–333.

Bickle, J. (1998). *Psychoneural reduction: The New Wave*. Boston, MIT/Bradford.

Bickle, J. (2003). *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Boston, Kluwer.

Brown, H. I. (1979). *Perception, theory and commitment: the new philosophy of science*. Chicago, University of Chicago Press.

Brown, H. I. (1988). *Rationality*. London, Routledge.

Causey, R. L. (1977). *Unity of Science*. Dordrecht, Reidel.

Christensen, W. D. and C. A. Hooker (1998). 'The dynamics of reason, critical symposium on Paul Churchland: *The Engine of Reason, the Seat of the Soul*.' *Philosophy and Phenomenological Research* **LVIII** (4): 871–8.

Christensen, W. D. and C. A. Hooker (1999). "An interactivist-constructivist approach to intelligence: self-directed anticipative learning." *Philosophical Psychology* **13**: 5–45.

Christensen, W. D. and C. A. Hooker (2000). "Organised interactive construction: the nature of autonomy and the emergence of intelligence." In A. Etxeberria, A. Moreno, and J. Umerez (eds.), *Communication & Cognition* **17**, Special Edition: The Contribution of Artificial Life and the Sciences of Complexity to the Understanding of Autonomous Systems, 133–58.

Christensen, W. D. and Hooker, C. A. (2001). "Self-directed agents." In J. S. MacIntosh (ed.) Naturalism, Evolution and Intentionality, *Ottawa: Canadian Journal of Philosophy* **27**: 19–52.

Churchland, P. M. (1975). "Two grades of evidential bias." *Philosophy of Science* **42**: 250–9.

Churchland, P. M. (1979). *Scientific realism and the plasticity of mind*. Cambridge, Cambridge University Press.

Churchland, P. M. (1985a). "Conceptual progress and word/world relations: in search of the essence of natural kinds." *Canadian Journal of Philosophy* **15**: 1–18. Reprinted in Churchland 1989.

Churchland, P. M. (1985b). "Reduction, qualia, and the direct introspection of brain states." *Journal of Philosophy* **82**: 1–22. Reprinted in Churchland 1989.

Churchland, P. M. (1989). *A neurocomputational perspective*. Cambridge, MA, Bradford/MIT Press.

Churchland, P. M. (1995). *The engine of reason, the seat of the soul: A philosophical journey into the brain*. Cambridge, MA, Bradford/MIT Press.

Churchland, P. M. and Hooker, C. A. (eds.) (1985). *Images of Science: Essays on Realism and Empiricism*, Chicago, University of Chicago Press.

Collier, J. D. and Hooker, C. A. (1999). "Complexly Organised Dynamical Systems." *Open Systems & Information Dynamics* **36**: 1–62.

Farrell, R. (2003). *Feyerabend and Scientific Values: Tightrope–Walking Rationality*. Boston Studies in the Philosophy of Science. Boston: Kluwer, **235**.

Feyerabend, P. K. (1961). "Knowledge without foundations." Oberlin College (mimeographed).

Feyerabend, P. K. (1962). "Explanation, reduction and empiricism." In Feigl, H. and Maxwell, G. (eds.) *Minnesota Studies in the Philosophy of Science*. **III**: Minneapolis, University of Minnesota Press.

Feyerabend, P. K. (1970). "Against method." In Radner, M. and Winokur, S. (eds.), *Minnesota Studies in the Philosophy of Science*, **IV**: Minneapolis, University of Minnesota Press.

Feyerabend, P. K. (1987). *Farewell to reason*. London, Verso.

Hooker, C. A. (1974). "Systematic realism." *Synthese* **26**: 409–97, reprinted in Hooker 1987.

Hooker, C. A. (1975). "The philosophical ramifications of the information-processing approach to the brain-mind." *Philosophy and Phenomenological Research* **36**: 1–15.

Hooker, C. A. (1979). "Critical notice of R. Yoshida, Reduction in the physical sciences." *Dialogue* **XVIII**: 81–99.

Hooker, C. A. (1981). "Towards a general theory of reduction." *Dialogue* **XX** (Part I, Historical framework, pp. 38–59, Part II, Identity and reduction, pp. 201–36, Part III, Cross-categorial reduction, pp. 496–529).

Hooker, C. A. (1987). *A realistic theory of science*. Albany, State University of New York Press.

Hooker, C. A. (1991). "Between formalism and anarchism: a reasonable middle way." In G. Munevar (ed.) *Beyond reason: essays on the philosophy of Paul Feyerabend*. Boston, Kluwer.

Hooker, C. A. (1995). *Reason, regulation and realism*, Albany: SUNY Press.

Hooker, C. A. (1996). "Toward a Naturalised Cognitive Science." In R. Kitchener and W. O'Donohue (eds.), *Psychology and Philosophy*. London, Allyn and Bacon.

Hooker, C. A. (1997). "Unity of Science." In W. H. Newton-Smith (ed.) *A Companion to the Philosophy of Science*. Oxford, Blackwell.

Hooker, C. A. (2004). "Asymptotics, reduction and emergence." *British Journal for the Philosophy of Science* **55**: 435–79.

Kim, J. (1998). *Mind in a physical world: an essay on the mind-body problem and mental causation*. Boston, MIT Press.

Kim, J. (1999). "Making sense of emergence." *Philosophical Studies* **95**: 3–36.

Marras, A. (2002). "Kim on reduction." *Erkenntnis* **57**: 231–57.

Moreno, A. and Ruiz-Mirazo, K. (1999). "Metabolism and the problem of its universalisation." *Biosystems* **49**: 45–61.

Nagel, E. (1961). *The Structure of Science*. New York, Harcourt, Brace and World.

Schaffner, K. F. (1967). "Approaches to reduction." *Philosophy of Science* **34**: 137–47.

Schaffner, K. F. (1992). "Philosophy of medicine." In M. Salmon et al. (1992). *Introduction to the Philosophy of Science*, Englewoods Cliffs, NJ, Prentice-Hall.

Sellars, W. (1962). "Philosophy and the scientific image of man." In R. Colodny (ed.) *Frontiers of Science and Philosophy*. Pittsburgh, University of Pittsburgh Press.

Sellars, W. (1965). "The identity approach to the mind body problem." In Cohen, R. and Wartofsky, M., *Boston Studies in the Philosophy of Science* **II**, New York, Humanities Press.

Sklar, L. (1967). "Types of inter-theoretic reduction." *The British Journal for the Philosophy of Science* **18**: 109–24.

Sklar, L. (1993). *Physics and Chance*. Cambridge, Cambridge University Press.

Yoshida, R. (1977). *Reduction in the physical sciences*. Halifax, N.S. Dalhousie University Press.

7 The Unexpected Realist

WILLIAM H. KRIEGER AND BRIAN L. KEELEY

INTRODUCTION

There are two ways to do the unexpected. The banal way – let's call it the expectedly unexpected – is simply to chart the waters of what is and is not *done*, and then set out to do something different. For a philosopher, this can be done by embracing a method of *non sequitor* or by perhaps inverting some strongly held assumption of the field. The more interesting way – the unexpectedly unexpected – is to transform the expectations themselves; to do something new and contextualize it in such a way that it not only makes perfect sense, but has the audience scratching their heads and saying, "Of course!" To do the unexpectedly unexpected on a regular basis is the true mark of genius. It recalls Kant's characterization of the genius as the one who not merely follows or breaks the rules of art but that, "Genius is the natural endowment that gives the rule to art."

We would not like to make the bold claim that Paul M. Churchland (PMC) is a philosophical genius of Kantian standards, but he sometimes achieves the unexpectedly unexpected and his position on the issue of scientific realism is a fine example of this. Given other views he holds and the philosophical forebears he holds dear, one might expect him to embrace an antirealism with respect to the posits of scientific theories. But, quite to the contrary, Churchland is one of the strongest contemporary philosophical voices on behalf of scientific realism. And, as we will discuss in this chapter, a closer look at this reasoning reveals that his realism is not perverse, it is exactly the sort of position he should be expected to hold, if only we understand the philosophical issues correctly.

THE DEBATE

Churchland does not describe his personal view on realism as "unexpected." Instead, PMC is, in his own words, "...a scientific realist, *of unorthodox*

persuasion" (1985: 35, our emphasis). In his 1979 book, *Scientific Realism and the Plasticity of Mind*, Churchland frames the realist/antirealist discussion in terms of the apparent difference between the knowledge claims we can make when dealing with the visible world as opposed to the world of science:

> The common opinion concerning scientific knowledge and theoretical understanding – of molecules, of stars, of nuclei and electromagnetic waves – is that it is of a kind very different from our knowledge of apples, and tables, and kitchen pots and sand. Whereas theoretical knowledge can be gained only by an act of creative genius, or by diligent study of the genius of another, knowledge of the latter kind can be gained by anyone, by casual observation. Theoretical understanding, it will be said, is artificial where the latter is natural, speculative where the latter is manifest, fluid where the latter is essentially stable, and parasitic where the latter is autonomous. (1)

In the next paragraph, he sums up: "That these specious contrasts are wholesale nonsense has not prevented them finding expression and approval in the bulk of this century's philosophical literature" (1). In this, PMC echoes the sentiments of another scientific realist, Grover Maxwell, who said this while introducing his own defense of the position in the early 1960s:

> That anyone today should seriously contend that the entities referred to by scientific theories are only convenient fictions, or that talk about such entities is translatable without remainder into talk about sense contents or everyday physical objects, or that such talk should be regarded as belonging to a mere calculating device and, thus, without cognitive content – such contentions strike me as so incongruous with the scientific and rational attitude and practice that I feel this paper *should* turn out to be a demolition of straw men. (1962: 3, emphasis in original)

However, be that as it may, both Maxwell then and PMC several decades later found realism in need of defense against other, very different notions of how to understand the posits of scientific theories.

What is at issue here? The stand one takes on the variety of philosophical realism at issue here is supposed to constitute a principled answer to this sort of philosophical question: *What metaphysical sense are we supposed to make of the sorts of entities to which scientific theories make reference all the time, but which, for a variety of reasons, no human being has directly experienced with his or her own senses?* No molecular biologist has ever seen a gene with her own two eyes. No subatomic physicist has ever felt the pull of the strong force

acting on a proton with his own fingertips. These being the case, should we consider genes, protons and a myriad of other so-called "theoretical entities" as metaphysically real as the sorts of objects with which we are far more familiar: rocks, mines, best-selling novels, and so on?

Realists, of which there are many types, have in common a positive answer to this sort of question. "Yes," they collectively reply, "the proton of the physicist's theory is *just as real as* the shoes on your feet. These theoretical entities are just unusual because their existence is of a kind that isn't compatible with the sensory systems with which Mother Nature has endowed us."

Antirealists, of which there are equally many (if not *more*) types, have in common a negative answer to this sort of question. "No," they collectively reply, "while talk of such entities is incredibly useful, perhaps even indispensable, to the practice of science, they aren't real in the same sense as everyday objects. Instead, we should think of them as 'useful fictions' – *façons de parler* – such as when demographers speak of the 'average American family' and its 2.3 children. We realize that the demographer doesn't propose that such posited ".3-children" exist in the same way as our *real* children do. The genes of molecular biology and the protons of the subatomic physics are of that different metaphysical category."[1]

An important element of this debate is a distinction traditionally drawn between theory and observation. According to this distinction, which is epistemic, there are some things that we observe directly with our own senses, say that this lead sphere is heavier in our hands than this equal-diameter aluminum one. Furthermore, there are other things that are theoretical claims, such as that only one of the two spheres has a molecular structure allowing the free flow of electrons between its constituent molecules because it is capable of conducting electricity. Any child or other person ignorant of science and its theories can know the former; the latter requires at least some acquaintance with a particular theory about the invisible molecular structure of metals and its relationship to electrical phenomena. The difference in what needs to be known in order to know the truth of a claim is the basis of this distinction between theory and observation.

Related to this epistemic distinction is a metaphysical one: namely that the metaphysical standing of that which we can directly observe with our senses is different from that which can only be known by virtue of its being embedded in a theory that posits it. We know of chairs because we observe them; we only know of electrons thanks to holding certain theories to be true. The difference in epistemic standing here is reflected in the different metaphysical standing of the entities in the two kinds of cases.

ENTER BAS VAN FRAASSEN

Realism has, until recently, been regarded as the dominant position in the realist/antirealist discussion. Scientists, when they can be coaxed to venture an opinion on the matter, tend to claim that they are realists, and that they couldn't remain scientists and still be antirealists. Ian Hacking, like Maxwell, agrees, asserting that at the level of scientific practice, "scientific realism [is] unavoidable" (1982: 71). According to these philosophers, realism, at least as regards theoretical entities, is necessary for any working scientist.

One traditional critique of antirealist (and pragmatic) positions is that there has been a negative platform against realism, but had no positive platform, no way forward, given the denial of things such as electrons. In response to traditional antirealism, and in opposition to traditional realism, Bas van Fraassen introduced a new position in 1980, called *constructive empiricism*. According to van Fraassen, the object of epistemological enquiry is to 'save the phenomena,' as opposed to any further commitment to truth (or to any other superempirical values). The worry van Fraassen was addressing is this: realism is often read as implying that the theoretical entities said to lay behind our experiences of things are *more real* than the appearances we experience. For example, the physicist, Sir Arthur Eddington famously writes in 1929: "I have settled down to the task of writing these lectures and have drawn up my chairs to my two tables. Two tables! Yes; there are duplicates of every object about me – two tables, two chairs, two pens" (ix). The first table is the table of our everyday experience: solid, with a smooth texture and color. The second table is the table of physics and it is mostly empty space and invisible atomic particles. Further, Eddington goes on to claim that only the scientific table exists: "modern physics has by delicate test and remorseless logic assured me that my second table is the only one which is really there – whatever 'there' may be" (xii). Such a realist account fails to save the phenomena.

In *The Scientific Image*, Van Fraassen asserts a real distinction between the visible and the theoretical, and argues that empiricism can lay claim to many of the virtues once touted by realists to dismiss antirealist positions. This is especially so for the claim that realism is simpler than antirealism; Van Fraassen counters that empirical adequacy is just as simple as realism, without the added complication or ontological baggage that accompanies words like 'true' or 'real.' In the end, van Fraassen's skepticism toward realism forced the full spectrum of realists to reevaluate their positions, bringing them back to the discussion.

CHURCHLANDISH REALISM

Churchland, like Maxwell and others, is explicitly a realist, but his grounds for realism are significantly different. Churchland's position, both before and after van Fraassen's critique is based on a rather universal skepticism. Throughout his works, Churchland regularly attacks orthodox realist virtues: that the historical process of theory development, testing, failure, and replacement is the best strategy, that the referents of the mature sciences necessarily map on to real things, or that some sort of cognitive evolution is guiding us in the right direction.

Given his dismissal of many of the cornerstones of realism, why accept Churchland's assertion that he is a realist? Because his skepticism does not recognize observability as relevant to the problem. The problem, for Churchland, is with cognition in general: "Since our observational concepts are just as theory-laden as any others, and since the integrity of those concepts is just as contingent on the integrity of the theories that embed them, our observational ontology is rendered *exactly as dubious* as our nonobservational ontology" (Churchland 1985, 36).[2] Churchland's unexpected realism is built upon a couple of positions, that empirical claims to knowledge can be shown to be just as (in) adequate as are those made by orthodox realists, and that the skepticism now focused on the status of unobservables should be refocused onto the status of theories in general, be they about the visible or the invisible.

Churchland's rejection of empiricism begins with the common supposition (supported by some realists) that there is a distinction between what is visible (and therefore true, if anything is) and what is invisible (whose truth or falsity is somehow to be seen as parasitic on visible facts). Van Fraassen believes that this supposition is warranted because, to him, the only measure of the truth of a theory is its empirical adequacy, its ability to "save the phenomena." All other measures or standards of truth, such as simplicity, explanatory coherence, and so on are inadequate (either subjective or arbitrary). Churchland explains how this "*selective* skepticism" works:

> The central element in this argument is the claim that, in the case of a theory whose ontology includes unobservables, its empirical adequacy underdetermines its truth. (We should notice that, in the case of a theory whose ontology is completely free of unobservables, its empirical adequacy does not underdetermine its truth: in that case, truth and empirical adequacy are obviously identical. Thus van Fraassen's *selective* skepticism with respect to unobservables.) That is, for any theory *T* inflated with unobservables,

> there will always be many other such theories incompatible with T but empirically equivalent to it. (1985: 38)

Leaving aside for the moment van Fraassen's dismissal of what Churchland calls "superempirical values," van Fraassen's position is called into question by Churchland on two grounds: (1) by attacking van Fraassen's assertions about underdeterminism and (2) by questioning the very concept of observability. Let's consider each of these individually.

First, underdeterminism is a thesis that can be traced back to the deductive-nomological (D-N) explanations preferred by logical positivists. To see what is at issue here, it will help to turn to the work of Carl Hempel for a short digression. According to Hempel (1965) these sorts of explanations are generally of the form:

$C, C_2, \ldots, C_k$ Explanans S
$L_1, L_2, \ldots, L_r$

E Explanandum-sentence

The D-N model states that an event (E)[3] is explained, given a set of explanatory sentences (S)[4], formed by a description of the initial conditions (C) and appropriate general laws (L). The relationship between the explanans and the explanandum is logical; that is, in a proper explanation, one can *deduce* the explanandum claim from the explanans. (This logical relationship is the "deductive" part; the necessary role for laws in the explanans is the "nomological" part.) In theory, so long as the appropriate laws and circumstances could be understood, the hope is that events can be fully explained. In cases where this works properly, the deductive form of the argument would guarantee the certainty (or truth) of the explanation. However, outside of Hempel's carefully chosen examples, this form of explanation rarely, if ever, works in practice.

In his earlier writings, Hempel agreed with his mentors – the logical positivists – that the mature sciences were naturally predisposed to scientific (D-N) analysis. However, Hempel came to realize that complete explanations, predictions, and the like were rarely, if ever, forthcoming. In many situations, there may be no way to determine all of the necessary conditions or to know the nature of the applicable laws. In these cases, the best that can be offered is an incomplete explanation, or *explanation sketch* of the event in question: "What the explanatory analysis of historical events offer is, then, in most cases not an explanation in one of the senses indicated above, but

something that might be called an *explanation sketch*" (Hempel 1965: 238, emphasis in the original).

These sketches were supposed to show what a full explanation would be, were the details available. The presumption was that, as the sciences evolved, general laws would be discovered, and explanation sketches would evolve into true explanations: "A scientifically acceptable explanation sketch needs to be filled out by more specific statements; but it points into the direction where these statements are to be found . . ." (Ibid.). Explanation sketches would point the sciences in the right direction. So, armed with hypothetical sketches of what explanations should look like, Hempel believed that scientists would then be able to wait for enough data to be collected for laws to fall out.

While there are problems with the very idea of general laws and with the relationship of those laws to scientific data (problems that Hempel and those who followed him have spent a lot of time defining and attempting to solve), there are more fundamental issues here. These have to do with the incompleteness of those data. Since scientists cannot fully control (or even collect) all of the relevant data, explanations – or more truthfully, explanation sketches – are necessarily underdetermined. For any imaginable set of data, there will be different theories that will be consistent with that data. So, the resultant explanations are best-guess approximations of truth. This acknowledgement, that many deductive scientific explanations suffer from underdeterminism, is hardly new, but the assertion that van Fraassen seems to make is that this problem only applies to phenomena that are unobservable. This is puzzling, as many scientists will readily admit that the laws they work with are approximations, that separating relevant initial conditions from irrelevant coincidences is difficult to do in the best circumstances. Outside of the strictly controlled, artificial conditions of a laboratory, phenomena (visible or not) rarely if ever behave *exactly* as they should. Whether this is given the current state of the sciences, or according to philosophers such as Cartwright (1983), given the nature of laws themselves, there is a strong relationship between laws and data. For Churchland, underdeterminism is as much a problem for cases involving observables as it is for cases involving unobservables.

Let's now turn to the second issue: the problem with the notion of observability. Although Hempel understood explanations to be underdetermined, he, like Churchland, is an unexpected realist. The logical positivists, as empiricists, focused on visible phenomena and considered exploration into empirically unverifiable entities as serving no epistemological purpose. For this reason, the logical positivists considered themselves to

be antirealists (at least on an epistemic level). For Hempel, explanations of phenomena (visible or not) involving theoretical entities only make sense if those entities exist, and if we can have enough knowledge about those entities to understand their role in the explanation. This is what philosophers of science would call a realist position. In his *Philosophy of Natural Science*, Hempel defends his realist conception of science, arguing against a number of rival (antirealist) theories concerning the existence of theoretical entities. For instance, while antirealists believe that there is a line that can be drawn between visible objects (considered real and directly knowable) and invisible entities (considered fictitious or knowable only via theory), Hempel believed that any such distinction would be arbitrary, and in the end, indefensible:

> Presumably [the class of observables] should include all things, properties, and processes whose presence or occurrence can be ascertained by normal human observers "immediately", without the mediation of special instruments or of interpretive hypotheses or theories . . . Wires . . . might count as observables. But we would surely not want to say that a rather fine wire becomes a fictitious entity when weakening eyesight compels us to use glasses to see it. But then, it would be arbitrary to disqualify objects . . . that no human observer can see without a magnifying glass. By the same token, we will have to admit objects that can be observed only with the aid of a microscope, and so on down to objects that can be observed only by means of Geiger counters, bubble chambers, electron microscopes, and other such devices.
>
> Thus, there is a gradual transition from the macroscopic objects of our everyday experience to bacteria, viruses, molecules, atoms, and subatomic particles; and any line drawn to divide them into actual physical objects and fictitious entities would be quite arbitrary. (1966: 81–2)

Certainly, in order for Hempel's models to be called explanatory, not only must observable entities be seen as real, so must the unobservable entities that play a role in those explanations.

Putting aside questions of relative completeness of explanations involving observables versus unobservables, Churchland also questions what van Fraassen considers observable, or what meets the criterion of empirical adequacy. Empirical adequacy refers to a particular (in this case, human) epistemic community, and empirical adequacy (as opposed to truth) is the final arbiter of theories.

To van Fraassen, the point of scientific theorizing is to explain our common sense world, perhaps echoing Sellars' contention that, "The aim

of philosophy, abstractly formulated, is to understand how things in the broadest possible sense of the term hang together in the broadest possible sense of the term. Under 'things in the broadest possible sense' I include such radically different items as not only 'cabbages and kings', but numbers and duties, possibilities and finger snaps, aesthetic experience and death" (1960/1963: 1). If we want to talk about electrons as the creators of the vapor trails visible in a cloud chamber, then that is fine; but we should not assume that those electrons are real in any meaningful sense of the world, and we should not expend energy trying to prove (or disprove) the existence of these scientific placeholders.

This proposed differentiation in ontological status between the visible and invisible leads Churchland to ask what van Fraassen means by *unobservable*. In "The ontological status of observables," Churchland points out that a given entity can be unobservable for a number reasons. For instance, I might not see something because I was in the wrong spatial relationship to it (say, if it were located on a small moon in the *51 Pegasi* system and I am on Earth), because I was not there at the right time (because the entity ceased to exist 2000 years before I was born), because I do not have the sensory acuity (due to poor my vision, say), or because I lack the proper apparatus to perceive it (such as attempting to observe an electron). These sorts of cases can be classed into two types. To discuss this (in Churchland's opinion) false difference, Churchland uses the following terminology to classify a difference that van Fraassen asserts is significant. Churchland calls items that are not observed due to spatiotemporal problems "observ*able*" while items that are imperceptible due to human physical limitations are "*un*observable." Van Fraassen distinguishes these two cases by a principle:

> X is observable if there are circumstances which are such that, if X is present to us under those circumstances, then we observe it. (van Fraassen 1980: 16)

The question is whether it is possible for van Fraassen to defend this principle, or what Churchland calls the observ*able* / *un*observable distinction. If there is a real line to be drawn between what is empirically available and what is not, then we must ask questions about the criteria offered by van Fraassen; criteria that van Fraassen uses to name members of the human epistemic community. In order to make sense of the two sorts of cases mentioned, let us consider two thought experiments. The first involves a hypothetical person who overcomes a happenstance spatiotemporal barrier for modern humans (such as the fact that I was born 2000 years too late to see a particular feature). The second case is a hypothetical person who

overcomes a happenstance physical barrier for modern humans (such as the fact that my range of hearing overlaps with, but is not as full as that of a dog). The point of these two cases will be to examine the differences between observ*ables* and *un*observables to see whether van Fraassen is correct in his assertion that these two cases are different in any meaningful way, and whether, as he claims, the former should be allowed into scientific discourse while the latter should be excluded.

OLD MEN AND NEW DOG TRICKS

In the case of the *2000-year-old man*, an archaeologist comes upon a certain feature, such as a line of collapsed stones. After careful study, that archaeologist explains this feature as an ancient city wall, torn down during an attack. Another archeologist, seeing the same feature and studying the same data, comes to the conclusion that the stones were once parts of a wall that fell as a result of a natural process (erosion, earthquake, whatever). In the first case, van Fraassen claims to have a way around my inability to see the past. He believes that an explanation is only truly able to save the phenomena when all of the phenomena (including those temporally removed from the investigator) are accounted for:

> A theory is empirically adequate exactly if what it says about the observable things and events in this world, is true – exactly if it 'saves the phenomena'. A little more precisely: such a theory has at least one model that all the actual phenomena fit inside. I must emphasize that this refers to *all* the phenomena; these are not exhausted by those actually observed, nor even by those observed at some time, whether past, present, or future. (van Fraassen 1980: 12, emphasis in original)

This distinction would solve the problem given above. According to van Fraassen, were the archaeologist sent back in time, (or, even better, were he 2000 years old) the original state of the wall would be visible, and the phenomenon saved.

In the *dog-eared boy* scenario, a man hears a series of very loud tones of increasing pitch. He is with his dog, who proceeds to howl after every tone. At some point, the man is no longer able to hear the noises, but the dog continues to periodically howl, falling silent at some time afterward. An auditory scientist would posit the existence of supersonic tones that the dog, but not its human companion, can hear. In this case, changes in location or life span would not solve the problem, rendering the unobservable, observable.

The only solution would be a physiological change (called canine-o-plasty) whereby the hearing range of the human ear is increased to include the range of the dog.[5] The question is why the *2000-year-old man* is acceptable for van Fraassen while the *dog-eared boy* is not.

If we took these situations to be similar, then, for van Fraassen, the proper epistemic attitude to science is focused on the observable. For an archaeologist, however, the proper focus may be on something altogether different. If a wall fell in battle, that will say something very different about a given city than if it fell naturally. In the former case, the archaeologist might try to determine why the city was significant militarily, while in the latter case, the investigator might seek to answer why the city was abandoned or in decline.

One response to this characterization might be to cite the open-endedness of science as vindication of van Fraassen's position. It is in fact the case that scientific methods are open ended, in the clear sense that conclusions are always tentative, more or less well confirmed, and subject to test. If this is true, then if the presently available evidence equally confirms two hypotheses, why not claim that van Fraassen's scientist would look for more evidence to separate them? The answer to this has to do with the goals of science. For an entity realist, the goal of scientific enquiry is to solve problems like 'do electrons exist?'. For a constructive empiricist, the goal of science is different. Van Fraassen can be ontologically agnostic as to the existence of theoretical entities, but epistemologically, the goals of science are clear:

> Thus acceptance [of a theory] involves not only belief but a certain commitment. Even for those of us who are not working scientists, the acceptance involves a commitment to confront any future phenomena by means of the conceptual resources of this theory. It determines the terms in which we shall seek explanations. (van Fraassen 1980: 12)

For van Fraassen, empirical adequacy is the goal of scientific explanation. Once that is achieved, there is no reason to continue. To do so would be to claim that there was something more (or real) to the entities van Fraassen claims are epistemically vacuous. As a result, if van Fraassen were deemed the victor in the realist/antirealist debate, then, the rigor of scientific experiments might have to drop to a level where a certain amount of sloppiness is deemed acceptable. For instance, if telescopes are considered to be beyond the boundaries of the human visual apparatus, then Ptolemaic epicycles might provide astronomers sufficiently accurate explanations, but

that possible world is probably not a place where working scientists would want to live.

Although explaining how the *2000-year-old man* sense of observ*able* (as temporally or spatially defined) would seem to be problematic for a person interested in empirical availability, it is the *dog-eared boy* sense of *un*observable (as physiologically defined) that van Fraassen has chosen to deny as being epistemically available to those who he defines as a specific (here human) *epistemic community*. In response to those who would see the two aforementioned scenarios as similar, van Fraassen asserts that this seeming similarity is merely a trick. As we read above, both Maxwell and Hempel assert that setting limits on scalar perspective is arbitrary. However, van Fraassen's critique of this position makes his distinction hard to take seriously. Van Fraassen charges Maxwell and Hempel of engaging in a number of flights of fancy. When they claim that the theoretical/observational distinction is a matter of perspective (using examples such as those discussed above), van Fraassen replies:

> This strikes me as a trick, a change in the subject of discussion. I have a mortar and pestle made of copper and weighing about a kilo. Should I call it breakable because a giant could break it? Should I call the Empire State Building portable? Is there no distinction between a portable and a console record player? The human organism is, from the point of view of physics, a certain kind of measuring apparatus. As such it has certain inherent limitations – which will be described in detail in the final physics and biology. It is these limitations to which the 'able' in 'observable' refers-our limitations, *qua* human beings. (1980: 17, emphasis in original)

If we are to grant that humans qua humans have certain limitations (a subject that we will revisit), the question of why one set of limitations (raised by the 2000-year-old man) is seen by van Fraassen as bridgeable, while another set of limitations (raised by dog-eared boys) is seen as being beyond his ability to understand. If there is such a thing as a human epistemic community, then one inherent limitation to those members would be an inability to travel to the past (or if it is not a limitation, time travel will surely be the death of archaeology as it is known now, making the point moot), it makes little sense for van Fraassen to deny Maxwell his argument on the grounds that only scalar sorts of human limitations mark the distinction between what is and is not real. If van Fraassen wants to show that some human limitations (such as being a giant or having X-ray eyes) are less likely than others (such as time travel), he will have to make that point more explicit.

Churchland uses a number of thought experiments to question exactly what van Fraassen has in mind when he defines the members of a particular epistemic community (1985: 43–4). For the *dog-eared boy* example, we know that each person has a different aural acuity. If one person were born with a larger than normal range of hearing (say, approaching that of a dog), would the sounds he hears somehow either distance him from the human epistemic community? Or would it cause the need for an 'on-the-fly' change to the definition of empirical adequacy? What about a person who is deaf? Would we consider that person inhuman, even if we are perfectly able to communicate with that person using means other than sound? Instead of canine ears, Churchland refers to humans with an electron microscope eye. In his article, "Empiricism in the Philosophy of Science," Van Fraassen claims that Churchland's mistake is to assume that scope-equipped humans are indeed human, and that there is some divine (objective) perspective that can be used to judge what is or is not human:

> But, on the one hand, it is not warranted to speak of their *experience* unless they are already assumed to be part of our epistemic community. On the other hand, while I do not know what, or indeed whether, factual conditions concerning causal connections suffice to make something a member of that community, I do know that what an empirically adequate theory implies about causal connections may not be true. We do not, in addition to the science we accept as empirically adequate, have a divine spectator who can tell us what is really going on. And, if we supposed that we had, there would still be two cases: does the supposition include that He is a member of the epistemic community or . . .? (Churchland 1985: 257, emphasis in original)

For van Fraassen, the mistake Churchland makes is to assert that a scope-equipped or dog-eared boy is human. In doing so, according to van Fraassen, Churchland is appealing to an outside objective observer who could somehow decide that these beings are human. On the other hand, van Fraassen believes that our natural limitations are what bind us together as a single epistemic community. An enhanced human would perceive the world differently, and would therefore not be human. To van Fraassen, the very idea of a being "exactly like us, except . . ." is mistaken on its face. In response, when Churchland proposes the possible existence of beings that are exactly like us, except for a particular modification in sensory acuity, he asks van Fraassen where the line should be drawn. Does the fact that a large portion of the world's populace need glasses in order to see mean that we should redraw the boundaries of observability? As humans grow taller (and can thereby see farther) with every generation, need this be taken into

account? Van Fraassen sidesteps the issue of where this boundary needs be drawn, but he asserts that the line is there and that it is commonly understood: "I think they [my critics] all agree also on the vagueness of observability and the irrelevance of exactly where the line is drawn. An electron is so unimaginably different from a little piece of stone . . . that minor adjustments would make no difference to the issues" (van Fraassen 1985: 254).

If there is a commonly agreed upon boundary somewhere between corrective lenses and electron microscopes, how is that boundary (the fact that we have certain limitations on our visual apparatus) any more significant than the boundaries mentioned in the first case above? Van Fraassen makes the assumption that the epistemic community is full of humans " . . . and that no one of us is really a person from Krypton . . ." (van Fraassen 1985: 254), but although seeing through Lois Lane's dress is epistemologically (not to mention ethically) inappropriate, time travel is apparently acceptable, so long as it involves a human using a machine (a la H. G. Wells) instead of a Kryptonian using his cape and tights. The fact that I cannot travel from the Earth to *51 Pegasi* is a matter of happenstance (a limitation in my lifespan that is a matter of biology), just as is the fact that I cannot hear everything that a dog hears. As such, the differential discrimination given against *un*observables as opposed to observ*ables* is difficult to accept.

Given our human limitations, Churchland sees no difference between observ*ables* and *un*observables. In each case, the evidence at hand will underdetermine explanations. However, as we have established above, the same underdeterminism applies to cases involving actually *observed* phenomena. This parity has led Churchland to argue against empirical adequacy, and instead for what he calls "Superempirical Virtues":

> Since there is no way of conceiving or representing 'the empirical facts' that is completely independent of speculative assumptions, and since we will occasionally confront theoretical alternatives on a scale so comprehensive that we must also choose between competing modes of conceiving what the empirical facts before us *are*, then the epistemic choice between these global alternatives cannot be made by comparing the extent to which they are adequate to some common touchstone, 'the empirical facts'. In such a case, the choice must be made on the comparative global virtues of the two global alternatives, T_1–plus-the-observational-evidence-therein-construed, versus T_2–plus-the-observational-evidence-therein-(differently)-construed. That is, it must be made on *superempirical*

> grounds such as relative coherence, simplicity, and explanatory unity. (Churchland 1985: 41–2, emphasis in original)

Empirical adequacy could only works as an arbiter of value there were such a thing as theoretically neutral empirical facts, a concept that Churchland dismisses. Regardless of the source of our epistemological underdetermination, the ideal account of empirical data being somehow free of theoretical bias is clearly false. Examples abound in the sciences of visible objects being invisible, being visibly different, or serving very different purposes in the past. Consider the following example from archaeology, taken from the work of Alain Schnapp (1997) commenting on the history of theoretical archaeology posits in that field:

> While some individuals enquired rigorously into the origins of objects and monuments, most of their contemporaries preferred to see these same objects as the product of the magical powers of mysterious beings, or of strange natural phenomena (34) [. . .] The "giants' footsteps", the "sorcerers' beds", corresponding to the scattered presence across the European landscape of megaliths and tumuli. . . . (148).

In these, and countless other examples, people's beliefs dictate what they actually see. This is not just a matter of interpretation. Collectors had curio cabinets full of fallen angels centuries before the science of archaeology was born. These angels became fossils only after the theory of evolution made its way into parlor circles (before the advent of scientific archaeology, archaeologists were generally just wealthy collectors).

If empirical values cannot serve as our universal, objective, standard, and as there is no other single standard able to do the job, Churchland understands the superempirical virtues as working together to guide us toward our presently underdetermined, imperfect view of the world, both on the empirical and the theoretical level. It is this position that labels Churchland as an unorthodox, or as we have labeled him, an unexpected, realist.

CHURCHLAND'S UNEXPECTED REALISM

"If observation cannot provide a theory-neutral access to at least some aspects of reality, then our overall epistemic adventure contains both greater

peril, and greater promise, than we might have thought" (Churchland 1988: 167). Given that realism about the visible is no different than (and is just as problematic as) realism about the theoretical, one might ask (and Churchland has) whether we can do better on both fronts. Churchland spends a great deal of time asking questions about the *reality* of our common sense understanding of the world.

Churchland's solution to his issues with realism is for humans to force themselves to evolve epistemologically. According to Churchland, our perceptual framework is governed by a conceptual framework, a learned, shared foundation within which our perceptions make sense:

> Our current modes of conceptual exploitation are rooted, in substantial measure, not in the nature of our perceptual environment, nor in the innate features of our psychology, but rather in the structure and content of our common language, an in the process by which each child acquired the normal use of that language. By that process each of us grows into a conformity with the current conceptual template. In large measure we *learn*, from others, to perceive the world as everyone else perceives it. (Churchland 1979: 7, emphasis in original) Churchland 1979: 7, emphasis in original)[6]

If this is so, Churchland asks, why not change this shared conceptual framework from its current form to a new, better one?

> After all, our current conceptual framework is just the latest stage in the long evolutionary process that produced it, and we may examine with profit the possibility that perception might take place within the matrix of a different and more powerful conceptual framework. (Churchland 1979: 7)

In order to answer questions about the future of human epistemology, Churchland delves into a discussion of human psychology and neurology. This conversation (primarily with Jerry Fodor) lies outside the scope of this chapter (an introduction to this discussion can be found in (Churchland 1988)). However, it is important to note that, for Churchland, if we want to understand the observable, observ*able* or *un*observable world, we need approach that world from a radically different perspective.

Another interesting effect of Churchland's unexpected realism is that his position allows for other so-called radicals to be brought back into the realist fold. Two philosophers of science who are famous for being outsiders are Kuhn and Feyerabend. Churchland sees both of their programs as holding

promise for his future as a realist. In fact, he has trouble choosing between the two:

> I confess that my political impulse inclines to Kuhn: it is a chaotic world out there and we must prize the monumental achievements of our institutions of scientific research and education. But my epistemological impulses, and my heart, incline to Feyerabend. Modern science is easily tough enough to tolerate, and even to encourage, the permanent proliferation of competing theories. We need only keep our standards high for evaluation their success. Maximizing the severity of those standards, it is arguable, was Feyerabend's ultimate end. (Churchland 1997: S419–S420)

While all of this is fine, we should note that PMC's realism comes at a cost. He is essentially saying that there is no *special* problem of the metaphysical status of theoretical entities; given his universal skepticism, they are no more (or less) problematic than the everyday objects of our experience. He sees this as nothing more than the unavoidable consequence of the theory-ladenness of observation. Of course, for some, this is a Pyrrhic victory at best and a self-inflicted *reductio ad absurdum* on Churchland's part, at worst.

Which brings us back to Churchland's unexpectedly unexpected stance on realism. He is a realist and it follows from the views he takes from those such as Feyerabend (such that, he attempts to show us why they *too* should be expected to embrace good old-fashioned scientific realism, too). This may be fine for Paul Churchland, whose occasional response to an accusation of a *reductio* is to embrace the absurd conclusion on the principle that one person's *modus ponens* is another's *modus tollens*. But while he may choose to do this, why ought the rest of us?

Notes

1. Of course, the actual philosophical positions are far more complicated and diverse than these cartoon sketches suggest. We beg the reader to allow us some poetic license here for the sake of exposition.
2. For more on Churchland and the concept of theory-ladenness, see Chapter 1.
3. Referred to as the "explanandum."
4. Collectively known as the "explanans."
5. Such a case may not be science fiction for long. Children (and adults) with hearing deficits are already being implanted with artificial cochlea. To date, they tend to use the human range of audition, but there is nothing in principle to prevent the range of audible frequencies being increased.
6. For an interesting critique of this work, see van Fraassen (1981).

Works Cited

Cartwright, N. (1983). *How the laws of physics lie*. Oxford, Clarendon Press.

Churchland, P. M. (1979). *Scientific realism and the plasticity of mind*. Cambridge, MA, Cambridge University Press.

_____. (1985). The ontological status of observables: In praise of the superempirical virtues. *Images of science: Essays on realism and empiricism, with a reply from Bas C. Van Fraassen*. P. M. Churchland and C. A. Hooker, (Eds.) Chicago, The University of Chicago Press, 35–47.

_____. (1988). "Perceptual plasticity and theoretical neutrality: A reply to Jerry Fodor." *Philosophy of Science* **55**(2): 167–87.

_____. (1997). "To transform the phenomena: Feyerabend, proliferation, and recurrent neural networks." *Philosophy of Science* **64**, **Supplement**: S408–20.

Eddington, A. (1929). *The nature of the physical world*. New York, Cambridge University Press.

Hacking, I. (1982). "Experimentation and scientific realism." *Philosophical Topics* **13**: 71–87.

Hempel, C. G. (1966). *Philosophy of natural science*. Englewood Cliffs, NJ, Prentice Hall.

Maxwell, G. (1962). The ontological status of theoretical entities. *Minnesota studies in the philosophy of science*. H. Feigl and G. Maxwell, (Eds.) Minneapolis, University of Minnesota Press. **III**: 3–14.

Schnapp, A. (1997). *The discovery of the past*. New York, Harry N. Abrams, Inc.

Sellars, W. (1960/1963). Philosophy and the scientific image of man. *Science, perception and reality*. W. Sellars, (Ed.) Atascadero, CA, Ridgeview Publishing Company: 1–40.

van Fraassen, B. C. (1980). *The scientific image*. Oxford, Clarendon Press.

_____. (1981). "Critical study of: Paul Churchland: *Scientific realism and the plasticity of mind*. Cambridge: Cambridge university press 1979. X + 157 pages." *Canadian Journal of Philosophy* **11**(3): 555–67.

_____. (1985). Empiricism in the philosophy of science. *Images of science: Essays on realism and empiricism, with a reply from Bas C. Van Fraassen*. P. M. Churchland and C. A. Hooker, (Eds.) Chicago, The University of Chicago Press, 245–308.

8 Two Steps Closer on Consciousness

DANIEL C. DENNETT

For a solid quarter century Paul Churchland and I have been wheeling around in the space of work on consciousness, and though from up close it may appear that we've been rather vehemently opposed to each other's position, from the bird's eye view, we are moving in a rather tight spiral within the universe of contested views, both staunch materialists, interested in the same phenomena and the same empirical theories of those phenomena, but differing only over where the main chance lies for progress. Our purely philosophical disagreements are arguably just matters of emphasis: we agree that folk psychological assertions limn real patterns in the world (to put it my way) and that these are (only?) useful approximations. Are they truths-with-a-grain-of-salt (my glass of mild realism is half full) or intermittently useful falsehoods (Paul's glass of eliminativism is half empty). We agree that there is no good motivation for shoehorning these folk categories into neuroscientific pigeonholes via a strict type identity theory, and even a strict functionalism would require some Procrustean labors that might better be postponed indefinitely, since the domain on the left hand side of the equation – the folk categories – is composed of items that are just not up to the task. This is true of folk categories more generally, not just the familiar terms of folk psychology. We also don't need counter-example-proof functionalistic definitions of charisma, moxie, or bizazz, though these are real qualities I have always admired in Paul.

To some observers, such as those of various mysterian persuasions, Paul and I are scarcely distinguishable, both happily wallowing in one 'scientistic' or 'reductionistic' swamp or another, taking our cues from cognitive scientists and unwilling or unable to begrudge even a respectful hearing to their efforts to throw shadows on the proceedings. For those who can see no significant difference between us, this essay will try to sharpen a few remaining disagreements, while at the same time acknowledging that in fact we are approaching harmony on a number of heretofore contested topics. I will try to close the gap further, much as I have always enjoyed his loyal opposition.

I met Paul in 1977, at the University of Manitoba, where I gave the first version of a talk that was subsequently published in *Brainstorms* (Dennett, 1978): "Two Approaches to Mental Images." The published version owed a lot to Paul's discussions, even though he is not referenced therein (I now note, with a smidgin of chagrin). We were both enthusiastic about enlarging the imaginations of philosophers of mind by getting them to dwell on actual scientific models and explanations, and since we tended to know different fragments of the relevant sciences, we had a lot to teach each other. His first book, *Scientific realism and the plasticity of mind* (Churchland 1979) was driven more by epistemological concerns than issues in the philosophy of mind but his discussion of perception and the plasticity of introspection was a major source of insights for me, especially informing my thinking about how important it was to ask what I later called the Hard Question: And Then What Happens? (*Consciousness Explained* 1991: 255). To capture the contents of consciousness, you need to see what a person can do with that state. Paul recently put it this way:

> Specifically, we both seek an explanation of consciousness in the dynamical signature of a conscious creature's cognitive activities, rather than in the peculiar character or subject matter of the contents of that creature's cognitive states. Dennett may seek it in the dynamical features of a 'virtual' von-Neumann machine, and I may seek it in the dynamical features of a massively recurrent neural network, but we are both working the 'dynamical profile' side of the street, in substantial isolation from the rest of the profession. (Churchland, 2002: 65)

Theories that stop when they reach some scarcely imagined "presentation" process (in the Cartesian Theater) are self-disabling, since they leave an unanalyzed (but knowledgeable, appreciative) witness to confront a now bafflingly contentful state with apparently miraculous powers of self-intimation, self-interpretation, and so on. You have to break up the given, and the taking of the given, into more modest parts whose operation we can actually begin to understand. Since people have no privileged access into this machinery, we'll just have to set aside the traditional philosophical method of introspection and turn to third-person models of processes that might arguably have the necessary competences. As he notes in this passage, Paul's favorite hammer and anvil for this breaking job is connectionism and vectors in a multidimensional space of content. I am impressed by some of

the work these tools can do, but don't view them as obviating the need for other perspectives, other levels of modeling.

Virtual machines versus recurrent neural networks – the opposition almost dissolves on closer inspection. I daresay that the virtual machines I like to talk about are implemented by the massively recurrent neural networks Paul likes to talk about. What else could do it? I entirely agree that there are massively recurrent neural networks churning away in our brains, and they are, as he has insisted for years, the key to understanding the neurocomputational perspective. So far, then, we differ only about whether in order to make sense of the powers of these networks we also need to describe their activity at a somewhat higher level, a virtual machine level. Should we take virtual machines in the brain seriously? That is the first question that divides us, and it leads directly to two others: Should we take memes seriously as what these virtual machines are 'made of'? and Is human consciousness a virtual machine (made of memes)? Note that the first two questions could get positive answers, and yet my most startling and revolutionary claim – answering yes to the third question – could be rejected. I suspect that part of Paul's motivation in dragging his feet so strenuously on the first two is that he wants to give himself plenty of room to maneuver in denying the third. Be that as it may, let me take them in order, and try to close the gaps. The history of our disagreement on this topic has had five steps:

1. CE. My 1991 book, *Consciousness Explained*, puts forward the virtual machine idea.
2. ER. In *The Engine of Reason, the Seat of the Soul* (1995) Paul includes a brief critique of my idea along with some delightful diagrams to illustrate it (264–69).
3. VV. In "The Virtues of Virtual Machines," (1999) Shannon Densmore and I portray step 2 as marred by a caricature of the idea of a virtual machine in the brain, and propose the division of labor suggested above: two levels of explanation, compatible and complementary.
4. VMC. In "Densmore and Dennett on Virtual Machines and Consciousness," his reply in the same issue of Phil and Phenom Research, Paul rejects our proposal quite vigorously (calling it "self-deceptive and uncomprehending"!) and rejects as well the language of virtual machines: "Those metaphors do not need to be qualified, they need to be junked" (763).

5. CC. In yet another essay, "Catching Consciousness in a Recurrent Net" in Brook and Ross (2002) Paul adds a further reading of the points of disagreement. It is clearly my turn to take a step – indeed I am one move behind his pace – so count this essay as two steps, taking his two prior steps as my springboard.

1. THE BASIC CONCEPT OF VIRTUAL MACHINES

Here I think the main obstacle to agreement is that Paul is still fixated on the wrong stereotype of virtual machines. He should clasp them to his bosom, not dismiss them scornfully. He needs them; they are just the right gadgets to complete his tool box. He got off on the wrong foot with an unflattering portrait of a virtual machine in his original critique in ER: "according to Dennett, our underlying parallel neural architecture is realizing a 'virtual' computing machine, whose activities are now of the classical, discrete-state, rule-governed, serial kind" (264). But as was made clear in VV, of the four adjectives in this list, I endorse only one: serial – and even that one gets highly qualified in my Multiple Drafts Model. In VMC, Paul acknowledged that I had never said my virtual machines were classical or discrete-state or rule-governed and indeed had quite explicitly cautioned against those interpretations, but, he claimed, this made matters even worse: my altered and metaphorical use of the concept of virtual machines was Pickwickian at best: by disavowing the very features that explain the power of classical virtual machines, I was creating an illusion of explanation where none existed. By Paul's lights, a virtual machine only makes sense when you have explicit source code (I had somewhat jocularly characterized a virtual machine as a "machine made of rules" and these are the "rules") implemented on a digital (discrete-state), serial machine. He says this, but he offers no reasons, and for the life of me, I don't see why he thinks so. To me, this is like insisting that you can implement a virtual machine only on an Intel chip. Why should other architectures, even non-digital architectures, be ruled out? His explanation doesn't help much:

> The network isn't following fuzzy rules, or imperfectly marshaled rules, or virtual rules; it isn't following rules at all. What we should look for, in explanation of the network's behavior, is the acquired dynamical landscape of its global activation space. That, plus its current activation state, is what dictates its behavior. Rules play no causal role at all, and neither do 'rules'. To use that term in scare quotes, as Dennett does, is just to undermine the primary negative point here, and to set up explanatory hopes and expectations

> (concern 'virtual machines' and 'design level explanations') that are doomed to go unsatisfied (765).

Strong talk, but there is a problem of levels lurking here, which we can make vivid by imagining an electrical engineer or chip designer making the parallel claims about a von Neumann machine:

> . . . it isn't following rules at all. What we should look for, in explanation of [the von Neumann machine's] behavior, is the [temporary] dynamical landscape of its global activation space. That, plus its current activation state, is what dictates its behavior. . . .

Programs in memory are, after all, just large fields of varying voltages that determine the dynamic sequence of voltage changes racing through circuit boards. Rules play no causal role in them either! Once you compile the source code, the "rules" evaporate. When you get right down to it, all the causal work is done at the level of flip flops and logic gates, and when a logic gate responds to its input "It is no more following rules than is the water of the Mississippi following rules in order to meander its way down a literal landscape from the northwest plains to the Gulf" (VMC: 764).

There is no rule-following in the hardware – either neural or silicon – but it doesn't follow from this that there is no rule-following (or "rule"-following – the scare quotes are needed in both cases) at a higher level. You simply cannot make sense of the versatile powers of a von Neumann machine without ascending to the program level, the virtual machine level, and at that level the regularities to be discovered, while based on or implemented via fundamental physical microregularities (the province of the gate designer) cannot be accounted for at the level of physics. To take a vivid example, consider the visible regularities of click-and-drag in the desktop user interface virtual machine. The icon on the desktop changes color and becomes somewhat translucent when it's being moved, and reverts to its original color when the cursor lets go of it. These regularities are not curious reflections of the implementation physics ("Hmm, could it be heating up due to some friction in the medium?") but regularities imposed by patterns in the implementation, and these regularities can be concocted ad lib, depending on features of the world outside the hardware, features tracked, or honored, or represented. The "physics" of the virtual machine is whatever the designers want it to be; it is virtual physics.

The same sorts of regularities are ubiquitous in minds. Consider, for instance, the regularities of people's reactions to a Stroop test, in which color words such as "red" and "green" and "blue" are written in differently colored inks and people are asked to name the colors of the inks, not the words written. People who are illiterate have no difficulty following the instructions and naming the colors; people who can read find it very hard to follow the instructions. Why is it harder for some people? Is it physically more difficult? Well yes, of course, in a sense. All difficulties are physical difficulties in the end. But the shape or pattern of the difficulties may need another level to describe. Consider Paul's claim: "What we should look for, in explanation of the network's behavior, is the acquired dynamical landscape of its global activation space. That, plus its current activation state, is what dictates its behavior." Yes, but the only explanatory way to describe that acquired dynamical landscape is in terms of the virtual machine thereby implemented. In this instance, what matters is whether there is an English-reading machine installed. In another instance, a much shorter-lived virtual machine might be responsible for a predictable effect. (As in the old trap questions: What kind of music did Woodie Guthrie sing? Folk. Who was President during the California Gold Rush? Polk. What do you call the white of an egg? Yolk. No, you dummy; albumin!) These are tiny toy examples to illustrate the phenomenon; when they are compounded into much more complex and highly articulated structures, the utility of the virtual machine perspective is undeniable. Cognitive psychology abounds in confirmed hypotheses about these machines, the conditions under which they are invoked and the circumstances under which they can be provoked into malfunction.

Perhaps Paul's longstanding distaste for the terminology of virtual machines should be catered to here, and we should let him treat himself to an alternative vocabulary for talking about the highly structured dispositions imposable (with a little practice or training) on the underlying "global activation space," just so long as he recognized that many of the highly salient regularities at one level will be inscrutable at his favored lower level, and that these regularities are mostly physically arbitrary in just the way the changing color of the dragged icon is physically arbitrary (from the point of view of the underlying machinery). Then there would be only a terminological preference separating us: what I and others (e.g., Metzinger 2003) insist on calling virtual machines, he would insist on calling something else. But I continue to urge him to chill out and recognize the tremendous utility, the predictive fecundity, the practical necessity of speaking of these higher levels as virtual machines. As a parade case, I commend Ray

Jackendoff's recent book, *Foundations of Language* (2002) which is a tour de force of (speculative, but highly informed, and deeply constrained) modeling of the virtual machine levels of neural implementation of language. The details matter, and I challenge anybody to say how they might recast all the insights in, say, Chapter 6, "Lexical Storage versus Online Construction," and Chapter 7, "Implications for Processing," in terms of the underlying recurrent neural networks. (See also pp. 22–3 for Jackendoff's reflections on this issue of the level of modeling.)

In CC, Paul's most recent step, he perseveres in his campaign against virtual machines, in a most curious way. First he notes that I am "postulating that, at some point in the past, at least one human brain lucked/stumbled into a global configuration of synaptic connections that embodied an importantly new style of information processing, a style that involved, at least occasionally, the sequential, temporally structured, rule-respecting kinds of activities seen in a typical vN [von Neumann] machine" (70). Yes, that's one way of putting it, and Paul goes on to acknowledge that indeed this possibility has been demonstrated in artificial recurrent networks. For instance, Cottrell and Tsung have trained networks to add individual pairs of n-digit numbers and distinguish grammatical from ungrammatical sentences in simplified formal languages.

> But are these suitably trained networks 'virtual' adders and 'virtual' parsers? No. They are literal adders and parsers. The language of 'virtual machines' is not strictly appropriate here, because these are not cases of a special purpose 'software machine' running, qua program, on a vN-style universal Turing machine (71).

This leaves me gasping. Paul, having just acknowledged that I am claiming that there is a perfectly good counterpart to classical virtual machines in the world of parallel machines, and having offered just the sort of example I would have chosen to illustrate it, pulls the definitional plug on me. This is not "strictly" appropriate use of the term "virtual machine" he says, because it isn't running on a vN machine! This begs the question. The Cottrell and Tsung machine is a special purpose software machine running, qua program, on a parallel machine. That very same 'hardware' recurrent network could have been trained up to do something else, after all. It was trained up to be, at least for a while, an adder or a parser. That's what a virtual machine is. A virtual machine does the very thing ("literally") a hardware machine does; it doesn't just approximate the task.[1] You can't retrain a hardware adder. If Paul thinks these trained neural networks are literal adders and parsers, I wonder what on earth he would call a virtual adder or parser.

Pursuing this definitional curiosity further, Paul sees an irony:

> if we do look to recurrent neural networks – which brains most assuredly are – in order to purchase something like the functional properties of a vN machine, we no longer need to 'download' any epigenetically supplied meme or program, because the sheer hardware configuration of a recurrent network already delivers the desired capacity for recognizing, manipulating, and generating serial structures in time, right out of the box (71).

This remark baffled me for some time. The underlying and untrained potential for recognizing, manipulating and generating serial structures in time is – must be – there, but saying that that capacity gives recurrent neural networks the functional architecture of a vN machine is like selling somebody a laptop without even an operating system and calling it a word processor. A randomly weighted recurrent neural net "right out of the box" is no serial vN machine. Precisely what we do need is the installation from outside of some highly designed system of regularities.

Sometimes we do the design work ourselves, laboriously, and sometimes we get a relatively easy download of largely predesigned systems. A natural language, as Chomskians are famous for telling us, installs itself in jig time in just about everybody, while sound probabilistic thinking is an unnatural act indeed, seldom successfully implemented in neural tissue. Several decades ago, I mastered the Rubik's cube, and got quite deft at spinning it into order. The fad expired; twenty years of disuse, like the similar hiatus in my use of German and French, have taken their toll, and a few months ago it took me quite a few hours to reinvent and re-optimize my cubist competence. (I guess I just needed to waste some precious time! During the obsessional phase, I couldn't stop imagining the subroutines and problems. Thank goodness I soon got over it.) If I don't rehearse my Rubik routines often in the months ahead, they will soon slip away again. What is this thing that can be problematically preserved in the connection strengths in my recurrent neural networks? It has structure that would be practically invisible to anyone intent on studying my neural networks and their dynamic properties, but readily describable as a sort of program that I have installed in myself and can run almost as mindlessly now as I usually run my English parser.

I think the case has been made for the appropriateness of virtual machine talk in cognitive neuroscience, and not just by me, and I look forward to the day when Paul retires from this dubious battle. I also anticipate a bounty of insights to flow from Paul when he exorcizes another bee in his bonnet: his mistrust of memes.

2. MEMES

I'll be brief about this, since Paul is brief and I have had a lot to say in defense of memes elsewhere (Dennett 1995, 2001a–c, 2002, forthcoming). Part of his problem with memes stems from his decision to take theories, probably the largest, rarest, hardest-to-transmit, most unwieldy of all cultural objects, and use them as his examples of choice.

"An individual virus is an individual physical thing, locatable in space and time. An individual theory is no such thing" (Churchland 2002: 66). True, but an expression or representation of an individual theory is an individual physical thing, and if we take the gene/meme parallel seriously, we recognize that a gene, too, is the information, not the vehicle of the information, which is always an individual physical thing. To see this vividly: ask yourself the following question. What if people in the future decided to forego sex and reproduce thus: Al and Barb both have their genomes sequenced, whereupon a meiosis program randomly composes unique Al-gamete and Barb-gamete specifications from their respective genomes and joins them into a zygote specification – a computer file that specifies the genome of an offspring. This specification is sent to a lab that thereupon hand-assembles that very genome out of materials taken from other biological sources, and creates an implantable "fertilized" embryo, which (for good measure) is then implanted in a surrogate mother, not Barb. Are not Al and Barb the "biological" father and mother of the resulting child? It's the information that counts. So genes are like theories in this regard: "abstract patterns of some kind imposed on preexisting physical structures . . . " (66).

"Furthermore," Paul goes on, "a theory has no internal mechanism that effects a literal self-replication. . . . " Neither does a virus, of course. It travels light and is artfully designed (by Mother Nature) to mindlessly commandeer the copying machinery in the cell it invades. A virus can be considered a string of DNA with attitude, but once again, it is the information that counts. Prions bring this out even more clearly, as Szathmary (1999) shows. Similarly a meme invades a body and gets itself copied, again and again, in a brain, But the physical token doesn't enter the body literally. A written word, for instance, does not enter the body (unless you're in the habit of eating your words!); rather, it produces an offspring on your retina, which then gets replicated again and again and again in your brain. Not so, says Paul. "It is that there is no such mechanism for theory-tokens" (67). I beg to differ, not just about individual words, and other individual vehicle-copies that get perceived, but even about "theories" large and small. This is what we call rehearsal or review, and it happens all the time. I just gave the vivid example

of my involuntary rehearsal of Rubik's cube memes, betokening themselves thousands of times in my poor brain, building ever stronger, better traces. What was being held constant while all the connection-strengths were being adjusted? The information.

Whole theories are unwieldy memes. Consider a much better example: a word. A grade school teacher of mine used to admonish "Say a word three times and it's yours!" and while the advice was largely gratuitous, the principle was right on target. Repetition is close to being a necessary condition for memorization, especially when we acknowledge that involuntary repetition (and unconscious repetition, which probably is ubiquitous) may do most of the work. If my Rubik's cube memes don't have offspring in the weeks to come, the lineage may well go extinct. What needs to be resurrected in me is not so different from woolly mammoth DNA after all. It lies unusable and unreplicable in the Vast state-space of my brain's parallel recurrent networks unless it gets regular cycles of reproduction. Paul (2002: 67) notes that "the 'replication story' needed, on the Dawkinsean view, must be nothing short of an entire theory of how the brain learns. No simple 'cookie-cutter' story of replication will do for the dubious 'replicants' at this abstract level." Exactly. Now where's the problem? Nobody ever said that a meme had to replicate by invading a single neuron.

It is curious that Paul ignores this perspective, since he has written hymns glorifying the repetitive power of recurrent neural circuits and their role in any remotely plausible theory of learning. The habit of rehearsal is a potent habit indeed, and it is required – Paul says as much in his discussion of the difficulties of internalizing a theory – to drive a theory into the network. How do you get to Carnegie Hall? Practice practice practice. But of course a lot of the rehearsal is not only not difficult; a lot of it is well nigh impossible to shut down. Rehearsal is itself a habit that is ubiquitous in our phenomenology – and it's just the tip of the iceberg! So here I'll just help myself to Paul's hymns to recurrence, for right there is the hardware that underlies the software, the built-in proto-rehearsal machinery that makes copying one's memes such an irresistibly easy step to take. The differential replication of memes within an individual brain is the underlying competitive mechanism of learning. And here a well-known evolutionary trade-off confronting parasites – should they specialize in the within-host competition against other strains of resident parasites (the path to virulence) or should they specialize on the competition to get from one host to the next (which leads to a-virulence, so that hosts can be up and about and in position to infect others)? – finds a parallel in the evolution of memes: getting a mnemonically potent phenotype that will get obsessively

rehearsed in one brain is part of the battle: getting transmitted favorably to another brain is a quite different venture. (I'll never forget John Perry's amusing bumper sticker: Another Family for Situation Semantics. John and a few colleagues and students had replicated the novel memes of situation semantics in uncounted rehearsals, but was anybody else ever going to be infected? John was not above trying the Madison Avenue approach.)

3. THE JOYCEAN MACHINE

But even if the virtual machine idea is worth pursuing, and even if the meme idea has some attractions, is there any hope for the preposterous claim that consciousness – consciousness! – is the activity of a virtual machine that only human beings implement, a virtual machine that depends on culture in general and language in particular? Surely this is just crazy! Many think so. Some of the wisest (and least conservative) heads in cognitive science think so. Paul thinks so.

> Instead, I shall argue, the phenomenon of consciousness is the result of the brain's basic hardware structures, structures that are widely shared throughout the animal kingdom, structures that produce consciousness in meme-free and von-Neumann-innocent animals just as surely and just as vividly as they produce consciousness in us (CC: 65).

This is a factual disagreement, not necessarily a philosophical disagreement of any sort, and he may be right. Or he may not. The point I want to make here is that his grounds for his belief are not anywhere near as strong as he thinks. I grant that we share a large part of our neurocomputational architecture with other animals, and that this shared architecture is sufficient to explain a great deal of the integrated, coherent, subtle behavior that both we and other animals exhibit, but I want to resist the further supposition, popular though it undoubtedly is, that this shared architecture (at the 'hardware' level) gives animals the sort of subjectivity, the sort of stream of consciousness, the point of view that we human beings all know that we share.

Paul is willing to grant that an uncultured, untutored, languageless mind is a relatively barren mind, perhaps even drab and boring in comparison to a normal (noninfantile) human mind:

> I do not hesitate to concede to Dennett that cultural evolution – the Hegelian unfolding we both celebrate – has succeed in 'raising' human

> consciousness profoundly. It has raised it in the sense that the contents of human consciousness – especially in its intellectual, political, artistic, scientific and technological elites – have been changed dramatically. . . . Readers of my 1979 book (see especially Chapters 2 and 3) will not be surprised to hear me suggesting still that the great bulk and most dramatic increments of consciousness-raising lie in our future, not in our past.
>
> But raising the contents of our consciousness is one thing – and, so far, a purely cultural thing. Creating consciousness in the first place, by contrast, is as firmly neurobiological thing, and that must have happened a very long time ago. For the dynamical cognitive profile that constitutes consciousness has been the possession of terrestrial creatures since at least the early Jurassic. James Joyce and John von Neumann were simply not needed (CC: 79).

That could not be clearer. I particularly applaud his allusion to Chapters 2 and 3 of his 1979 book, which remain, for me, my favorite bits of Churchlandiana. And as I say, he may be right. But until I am proved wrong, I am going to defend a more abstemious and minimalist view, one that resists the easy and popular course of supposing, with tradition, that our furry friends (and, if Paul is right, our feathered friends and even many of our scaly friends) have streams of conscious much like our own. I consider it telling that when Paul disparages this outrageous view of mine in ER, he shows a diagram of a grumpy-faced chimp (contrasted with a smiling member of H. sapiens) and goes on to say that "Dennett's account of consciousness is . . . unfair to animals" (269). The moral dimension is thus lurking not far beneath the surface, and we should all recognize that part of what is repugnant (to many) in my view is that it seems destined to license a shocking callousness with regard to animals (who are not really conscious, just as Descartes said, that evil man!). Recognizing that this is, or ought to be, an irrelevant consideration insofar as we want to know the scientific truth, and recognizing moreover that it nevertheless plays a potent role in biasing people against any hint of such a position, we ought to go out of our way to consider whether or not it might be true. That is why I continue to push my shocking view: because I see no good reason has been offered for not counting it as a serious candidate.

It might well seem that the disagreement between Paul and me here is just a special case of our earlier disagreement about whether a recurrent neural network counts as a serial architecture. He says yes, and I say no: the settings of the connections make all the difference, since they are what fix the truly remarkable powers of some such recurrent networks – by

programming them, in effect. Similarly, he says that animals are conscious, and I say that they are not, since what they are conscious of, the settings, if you will, that flavor their consciousness, do not do enough good work to count. But if that were all that divided us, it wouldn't be much of a disagreement. I could lament the fact that you just can't teach a chimp to solve the Rubik's cube, and so, you see, the chimp has such a paltry stream of consciousness that it hardly counts as conscious at all, and Paul could insist, on the contrary, that dim though a chimp's stream of consciousness is, it still counts as a stream of consciousness. But I am envisaging a more radical difference between the chimp and us. I am supposing that nothing like a stream of consciousness occurs in a chimp brain precisely because what kindles and sustains such a stream of consciousness in us is a family of microhabits of self-stimulation that have to be installed by culture. Without the cultural inculcation, we would never get around to having a stream of consciousness, though, of course, we would be capable of some sort of animalian activity. I am not denying that there are crucial architectural differences between chimp brains and ours. If it weren't for these, chimps could be enculturated and given human languages of some kind – manual sign languages most likely. But the differences might be quite subtle (see Deacon 1997 for an insightful account of the possibilities.) Deaf human infants, for instance, are intensely curious about human communication in spite of the absence of auditory input, while chimps that can hear perfectly well have to be heavily bribed with rewards to pay attention to human efforts at communication. Our brains are in some regards genetically designed to download cultural software, and chimps' brains are apparently not so designed.

Interestingly, Paul himself draws attention to this in a passage that is meant to cast doubt on the meme/virus parallel: "A mature cell that is completely free of viruses is just a normal, functioning cell. A mature brain that is completely free of theories or conceptual frameworks is an utterly dysfunctional system, barely a brain at all" (CC: 67). There are several points to make about these claims. First, it is not the case in general that normal cells can function without any viruses or other endosymbionts. After all, the mitochondria that are the prerequisite for eukaryotic life started out as cellular parasites, and more and more of the standard intracellular machinery turns out to have begun its career as software downloads of a sort. This is still going on, and it is well known that many cells cannot perform their current functions without the aid of "foreign" visitors of one sort or another. As in computer science, software development often precedes hardware development. More important for the present point, I agree with Paul that a mature human brain free of culture is utterly dysfunctional.

That's my point. But a chimp brain free of theories or conceptual frameworks – depending on what we mean by that – is not so obviously abnormal or dysfunctinoal. Animals learn from their own experience, by trial and error and general exploratory behavior, and Avital and Jablonka (2000) draw attention to the evidence that much of what has been standardly deemed to be "instinctual" knowhow transmitted through the genes is better considered animal "tradition" and can in fact be imparted by parent-offspring interactions and other social learning situations. (See also my review in *Journal of Evolutionary Biology*, Dennett 2002b). But no nonhuman animal species has a brain that is as adapted for massive cultural downloading as ours is, and hence no nonhuman animal is as handicapped by being denied its conspecific culture as we would be. Given these undeniably huge differences in both potential and dependence, the assumption that animal brains are architecturally enough like ours to sustain something properly called a stream of consciousness owes more to cultural habit than scientific insight.

It is worth noting that as primatologists and animal psychologists learn more and more about the minds of chimpanzees and bonobos (and dolphins and orangs and other favored species), they discover more and more surprisingly blank walls of incomprehension. The idea that these creatures are, in some regards, sleepwalking through life, to put it crudely and misleadingly, is not so easy to shake. I have been monitoring and occasionally contributing to the experimental literature on animal intelligence – especially higher-order "theory of mind" intelligence – for several decades, and to me the striking fact is that for every gratifying instance of (apparent) comprehension in one species or another, there are more instances of frustrating stupidity, unmasked tropism, and hard-to-delineate density that is hard to reconcile with the standard presumption that these creatures are confronting a world of experience pretty much the same way we are. Yes, they can be seen to be ignoring some things and attending to others, but the attention they can pay doesn't seem to enlighten them in many of the ways ours does. In short, they don't show much sign of thinking at all.

"But still, they are conscious!" Oh yes, of course, if all you mean is that they are awake, and taking in perceptual information, and coordinating their behavior on its basis in relatively felicitous fashion. But if that is all that you mean by asserting that they are conscious, you shouldn't stop at mammals, or vertebrates. Insects are conscious in that sense. Molluscs are too, especially the cephalopods. That is not what I am skeptical about. I am skeptical about what I have called the Beatrix Potter syndrome: the imaginative furnishing of animal minds with any sort of subjective appreciation, of fearful

anticipation and grateful relief, of any capacity to dwell on an item of interest, or recall an episodic memory, or foresee an eventuality. Animals can "learn from experience," but this kind of learning doesn't require episodic memory, for instance. When we see a dog digging up a buried bone it is quite natural for us to imagine that the dog is happily recalling the burying, eagerly anticipating the treasure to be recovered just as he remembered it, thinking just what we would if we were digging up something we had earlier buried, but in fact there is not yet any good evidence in favor of this delightful presumption. The dog may not have a clue why he is so eagerly digging in that spot. (For the current state of the evidence of "episodic-like" memory in food-caching birds and other animals, see Clayton and Griffiths 2002). And animals can benefit from forming a "forward model" of action that doesn't require the ability to foresee "consciously"; we ourselves are seldom conscious of our forward models until they trip up on an anomaly. Once we have stripped the animal stream of consciousness of these familiar human features, it is, I claim, no longer importantly different from a stream of unconsciousness! That is, it is a temporal flow of control processing, with interrupts (pains, etc.) and plenty of biasing factors, but it otherwise shows few if any of the sorts of contentful events that we associate with our own streams of consciousness. I think we need to set aside the urge to err on the side of morality when we imagine animals' minds; this attitude has its role in making policy decisions about how to treat animals, but should not be hardened into an unchallengeable "intuition" when we ask what is special about consciousness.

Paul's firm insistence that of course animals are conscious, and that human consciousness is just richer, is to me like the claim that five-year-olds write novels. Here's Billy's: Tom hit Sam. The End. It's a novel if you say so. But why are you so eager to say it's a novel? I am as eager as Paul is to support the humane treatment of animals, but I don't believe that the right way to do it is to saddle myself uncritically with the folk concept of (animal) consciousness that makes us happy to imagine that there is some (nice, warm, familiar, unified) sort of inner "show" in animal's brains, "the way there is in ours." When you look hard at the conditions for the "show" going on in ours, you begin to see that it is probably heavily dependent on a great deal of activity that is specific to our species. If you doubt this, ask yourself the following weird questions: What is it like to be an ant colony? What is it like to be a brace of oxen? The immediate, "intuitive" answer is that it is not like anything to be either one of these things, because these things, impressively coordinated though they may be in many regards, are not sufficiently unified, somehow, to support a single

(conscious) point of view. But just putting that ant colony inside a skull wouldn't automatically unify their activities the extra amount needed, would it? Tremendous feats of coordination are possible in control structures that nevertheless do not conspire to create the sort of user-illusion that we human beings call consciousness. If animal brains could do what ant colonies can do, why would animal brains bother doing all this further work? I have offered a sketch of an evolutionary explanation about why our brains, the brains of a linguistically communicating social species, would go to this extra work. I grant that there may well be an explanation for why the architectural features Paul finds shared in most if not all animal brains should be seen to yield enough of the human features to persuade us to call the processes that run so effectively therein conscious processes. But that is work still to be done, not presupposed. This is an empirical issue, not a philosophical one, except insofar as there are residual unclarities or misapprehensions about the meanings of the claims being advanced. And to echo the note from Paul with which I began this essay, for all our disagreements, Paul and I are united in thinking that these are scientific questions that cannot be solved, or even much advanced, by the intuition-mongering of armchair philosophy.

Note

1. It is worth remembering that today almost all hardware machines are designed and exhaustively tested as virtual machines long before the first hardware version is built. It is also possible, of course, for a virtual machine to be designed to approximate the task of a hardware machine. Virtual machines are the ultimate modeling clay – you can make just about anything out of rules.

Works Cited

Avital, E. and Jablonka, E. (2000). *Animal Traditions: Behavioural Inheritance in Evolution*. Cambridge, Cambridge University Press.

Brook, A. and Ross, D. eds. (2002). *Daniel Dennett*. Cambridge, Cambridge University Press.

Churchland, P. (1995). *The Engine of Reason, the Seat of the Soul*. Cambridge, MIT Press.

Churchland, P. (2002). "Catching Consciousness in a Recurrent Net." *Daniel Dennett*. Brook and Ross, 64–81.

Churchland, P. (1999). "Densmore and Dennett on Virtual Machines and Consciousness" *Philosophy and Phenomenological Research* **59** (3): 763–7.

Clayton, N. S. & Griffiths, D. P. (2002). Testing episodic-like memory in animals. In Squire, L. and Schacter, D. (eds.) *The Neuropsychology of Memory*, Third Edition. Chapter 38, New York, Guilford, 492–507.

Deacon, T. W. (1997). *The Symbolic Species*. New York, Norton.

Densmore, S. and Dennett, D. (1999). "The Virtues of Virtual Machines" in *Philosophy and Phenomenological Research* **59** (3): 747–61.

Dennett, D. (1978). *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge, MIT Press.

Dennett, D. (2001a). "The Evolution of Culture," *The Monist* **84** (3): 305–24.

Dennett, D. (2001b). "Memes: Myths, Misgivings, Misunderstandings," Chapel Hill Colloquium, October 15, 1998, University Chapel Hill, North Carolina, translated into Portugese and published in *Revista de Pop*, No. 30, 2001.

Dennett, D. (2001c). "The evolution of evaluators." In *The Evolution of Economic Diversity*, Antonio Nicita and Ugo Pagano, eds., New York, Routledge, 66–81.

Dennett, D. (2002a). "The New Replicators," In *The Encyclopedia of Evolution*, volume 1, Mark Pagel, ed., Oxford, Oxford University Press, E83–E92.

Dennett, D. (2002b). "Tarbutniks rule. Review of Eytan Avital and Eva Jablonka, Animal Traditions: Behavioural Inheritance in Evolution, 2000." *Journal of Evolutionary Biology* **15**: 329–34

Dennett, D. forthcoming, "From Typo to Thinko: When Evolution Graduated to Semantic Norms," In S. Levinson & P. Jaisson (Eds.), *Culture and evolution*. Cambridge, MA, MIT Press.

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford, Oxford University Press.

Metzinger, T. (2003). *Being No One: The Self-Model Theory of Subjectivity*. Cambridge, MIT Press.

Szathmary, E. (1999). "Chemes, Genes, Memes: A Revised Classification of Replicators." *Lectures in Mathematics in the Life Sciences*, **26**: American Mathematical Society, 1–10.

Index

A Neurocomputational Perspective, 3, 84, 113
Austin, J. L., 12

Barsalou, L. W., 103, 137
Batterman, R. W., 159, 165
Bermúdez, J. L., 22–23, 32–63
Bickhard, M. H., 168
Bickle, J., 163–164
Boghossian, P., 23, 33–35
Brook, A., 196
Brown, H. I., 157

Cartwright, N., 181
Chater, N., 130
Christensen, W. D., 167, 168
Churchland, P. S., 2, 3, 155
Clark, A., 2
Clayton, N. S., 207
commonsense psychology. *See* folk psychology
concepts. *See* prototypes
connectionism, 2, 23–24, 107–111, 113–149, 158, 167, 194
consciousness, 25–26, 67–68, 73–80, 193–208
Cottrell, G. W., 24, 51, 88–100, 108, 110, 113–149, 199

Damasio, A., 109
Deacon, T., 205
Dennett, D. C., 2, 25–26, 193–208
deVries, W. A., 8
Diamond, J., 169
Dretske, F., 2, 74, 75, 80
Duhem, P., 7
dynamic touch, 58–59

Ebbinghaus illusion, 53
ecological psychology, 57–58
Eddington, A., 178
Eliminative materialism, 1, 2, 11–13, 14, 15, 18–22, 23, 25, 32–63, 154, 158, 167, 193

Elman, J., 116
experience. *See* consciousness

Feyerabend, P. K., 1, 3, 10–18, 19, 22, 25, 154, 157, 190, 191
Flannery, T., 169
Fodor, J., 2, 23, 24, 38, 51, 88, 90–100, 101, 103, 104–108, 110, 115, 149, 190
folk psychology, 9, 11, 14, 18–22, 25, 32, 33–34, 35–46, 49, 52, 54, 57, 154, 167, 168, 193

Garzón, C., 97
Gibson, J. J., 57–58
Glymour, C., 147
Goldstone, R. L., 130, 131
Goodale, M. A., 53
Griffiths, D. P., 207

Hacking, I., 178
Hahn, U., 130
Hanson, N. R., 1, 3, 4–7, 10, 12, 13–18, 19
Harman, G., 74
Hebb, D., 88
Hempel, C., 180–182, 186
Holons, 93, 97–99, 102, 108, 110–111
Hooker, C. A., 24, 154–169
Hume, D., 23, 24, 88, 100–104, 107–111

intentionality, 68–70, 72, 78, 82–84, 108

Jackendoff, R., 199
Jackson, F., 38

Keeley, B. L., 1–26, 175–191
Kekulé, F., 5–6
Kind, A., 68, 77
Krieger, W. H., 24–25, 175–191
Kuhn, T., 3, 4, 5, 123, 190, 191

Laakso, A., 24, 51, 94–100, 110, 113–149
Lakatos, I., 4, 20
language of thought, 2, 47, 95, 105, 115, 117, 120

Lepore, E., 23, 24, 51, 88, 90–100, 101, 103, 104–107, 110, 149
Lewis, C. I., 10
Lewis, D., 38
Llinas, R., 113
Locke, J., 108, 109
Lycan, W., 75

Mandik, P., 23, 66–86
Matter and Consciousness, 2
Maxwell, G., 176, 178, 179, 186
memes, 201–203
Mervis, C. V., 130
Metcalfe, J., 93
Metzinger, T., 198
Mill, J. S., 9
Milner, A. D., 53
Mishkin, M., 53
Moore, G. E., 73, 74
myth of the given, 10, 13

Nagel, E., 161, 163
NETtalk, 51, 99, 108, 119
neural nets. *See* connectionism
Noelle, D., 149
Nosofsky, R. M., 138

Palmeri, T. J., 138
paralellel distributed processing (PDP). *See* connectionism
Pecher, D., 104
Pellionisz, A., 113
Perry, J., 203
phenomenal experience. *See* consciousness
Place, U. T., 14
Pollack, J., 116
Preston, J., 11
Prinz, J. J., 23–24, 88–111, 118, 132, 138
prisoners' dilemma, 42–46
prototypes, 23, 88, 89, 92, 93, 96, 97, 98, 100–102, 119, 124, 126, 131–138, 148
Pylyshyn, Z., 51, 98

qualia, 67, 68, 80–84, 119
Quine, W. V. O., 3, 91, 92, 114, 149

Richardson, L. B., 130
Rorty, R., 10
Rosch, E., 130
Rosenberg, J., 51, 99, 119
Rosenthal, D., 75
Ross, D., 196

Schaffner, K., 163, 164
Schnapp, A., 189
Scientific Realism and the Plasticity of Mind, 2, 25, 82, 156, 176, 190, 194, 203–204
Sejnowski, T. J., 51, 99, 119, 142
Sellars, R. W., 7
Sellars, W., 1, 3, 7–10, 11, 13–18, 19, 158, 182
Shastri, L., 140, 145
Sklar, L., 156
Smart, J. J. C., 14
Smolensky, P., 2, 50–51, 98, 101, 116, 145
social psychology, 60–61
Son, J., 131
Sorell, T., 4
Stich, S., 2
Szathmary, E., 201

The Engine of Reason, the Seat of the Soul, 3, 195
Touretzky, D. S., 116
Triplett, T., 8
Turing, A., 89
Tversky, A., 128, 130
Tye, M., 74, 75, 79–80

Ungerleider, L., 53

van Fraassen, B. C., 178–188

Wittgenstein, L., 12, 13–15