A. Carsetti
Editor

Theory and Decision Library A

Causality, Meaningful Complexity and Embodied Cognition

Springer

Causality, Meaningful Complexity and Embodied Cognition

THEORY AND DECISION LIBRARY

General Editor: Julian Nida-Rümelin (*Universität München*)

Series A: Philosophy and Methodology of the Social Sciences

Series B: Mathematical and Statistical Methods

Series C: Game Theory, Mathematical Programming and Operations Research

SERIES A: PHILOSOPHY AND METHODOLOGY
OF THE SOCIAL SCIENCES

VOLUME 46

Assistant Editor: Martin Rechenauer (*Universität München*)

Scope: This series deals with the foundations, the general methodology and the criteria, goals and purpose of the social sciences. The emphasis in the Series A will be on well-argued, thoroughly analytical rather than advanced mathematical treatments. In this context, particular attention will be paid to game and decision theory and general philosophical topics from mathematics, psychology and economics, such as game theory, voting and welfare theory, with applications to political science, sociology, law and ethics.

For other titles published in this series, go to
www.springer.com/series/6616

A. Carsetti
Editor

Causality, Meaningful Complexity and Embodied Cognition

Springer

Editor
Prof. A. Carsetti
Università di Roma - Tor Vergata
Facoltà di Lettere e Filosofia
Via Columbia, 1
00133 Roma
Italy

ISBN 978-90-481-3528-8 e-ISBN 978-90-481-3529-5
DOI 10.1007/978-90-481-3529-5
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2010921000

Cover design: Boekhorst Design b.v.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Acknowledgements

This book too like others I have edited, owes its existence to many sources. First of all I would like to express my deep appreciation to Julian Nida Rümelin. He encouraged me to edit this book and I have benefited from some discussions with him on the occasion of the International Colloquium on "Causality, Meaningful Complexity and Knowledge Construction" (Rome, 2008). I am very grateful to Charles Erkelens who examined and approved the first project of the book. I am also grateful to Elisabeth Leinfellner for her invaluable help in introducing me in the art of "shepherd" the different chapters in a volume.

I am indebted to my collaborators Andrea Cataldi, Pia't Lam and Enrica Vizzinisi for their help at the editorial level. I would like, in particular, to thank Lucy Fleet of Springer for her editorial comments and suggestions which contributed to the quality of the presentation of the book.

My deep thanks to the authors for their co-operation and, once again, to my students at the University of Rome "Tor Vergata" for their stimulus and their patience. Many thanks also to the researchers working at the level of the National Project of Research "Measures of epistemic complexity and knowledge construction" (MIUR, 2006) and, in particular, to Maria Carla Galavotti, Sergio Galvan and Roberto Festa, for their technical support and invaluable friendship.

Finally, I would like to express my gratitude to Franz Wuketits, Werner Leinfellner, Stephen Grossberg, Henri Atlan, Giuseppe Longo, Johann Götschl and Jean Petitot: I have greatly benefited from discussions with them about some specific guidelines concerning the organization of those particular International Colloquia of La Nuova Critica that constitute the conceptual "skeleton" of the volume. I will always remember the late Gaetano Kanizsa, Vittorio Somenzi and Valerio Tonini for their help and their teaching. They have been enormously helpful to me in thinking about the issues concerning meaningful complexity and embodied cognition.

Introduction

Arturo Carsetti

According to molecular Biology, true invariance (life) can exist only within the framework of ongoing autonomous morphogenesis and vice versa. With respect to this secret dialectics, life and cognition appear as indissolubly interlinked. In this sense, for instance, the inner articulation of conceptual spaces appears to be linked to an inner functional development based on a continuous activity of selection and "anchorage" realised on semantic grounds. It is the work of "invention" and generation (in invariance), linked with the "rooting" of meaning, which determines the evolution, the leaps and punctuated equilibria, the conditions related to the unfolding of new modalities of invariance, an invariance which is never simple repetition and which springs on each occasion through deep-level processes of renewal and recovery. The selection perpetrated by meaning reveals its autonomy above all in its underpinning, in an objective way, the ongoing choice of these new modalities. As such it is not, then, concerned only with the game of "possibles", offering itself as a simple channel for pure chance, but with providing a channel for the articulation of the "file" in the humus of a semantic (and embodied) net in order to prepare the necessary conditions for a continuous renewal and recovery of original creativity. In effect, it is this autonomy in inventing new possible modules of incompressibility which determines the actual emergence of new (and true) creativity, which also takes place through the "narration" of the effected construction. *Pace* Kant, at the level of a biological cognitive system sensibility is not a simple interface between absolute chance and an invariant intellectual order. On the contrary, the reference procedures, if successful, are able to modulate canalization and create the basis for the appearance of ever-new frames of incompressibility through morphogenesis. This is not a question of discovering and exploring (according, for instance, to Putnam's conception) new "territories", but of offering ourselves as the matrix and arch through which they can spring autonomously in accordance with ever increasing levels of complexity. There is no casual autonomous process already in existence, and no possible selection and synthesis activity via a possible "remnant" through reference

A. Carsetti (✉)
University of Rome "Tor Vergata", V. Columbia n.1, 00199 Rome, Italy
e-mail: carsetti@uniroma2.it

procedures considered as a form of simple regimentation. These procedures are in actual fact functional to the construction and irruption of new incompressibility: meaning, as *Forma formans*, offers the possibility of creating a holistic anchorage, and is exactly what allows the categorial apparatus to emerge and act according to a coherent "arborization". The new invention, which is born then shapes and opens the (new) eyes of the mind: I see as a mind because new meaning is able to articulate and take root through me.

In this sense, at the human level vision extends within a coupled system characterised by the presence of two different selective forces: the selection linked to the full expression of the original incompressibility, on the one hand, and the selection performed within an ambient meaning, on the other hand (this is a point of fact we are now ready to examine in the light of current achievements in contemporary theoretical Biology). Within the process, meaning reveals itself (albeit partially) in (and through) the effected emergence. Only in this way can a real assimilation process articulate, on the basis first of all of a coherent construction of possible schemes, self-organising models, falsification acts, and so on. In self-organising emergence, then, we find, simultaneously, a process of assimilation, one of growth, one of "inscription" and one of stabilisation through fixed points. It is therefore not surprising that, as soon as the assimilation (and the unfolding by unification at the brain level) of meaning occurs correctly, vision appears veridical. What this particularly presupposes as an essential component of the process is also the articulated presence of definite capacities of self-reflection and precise replication-mechanisms at the level of vision by models. If it is, actually, obvious that no thought can exist which has not first filtered through the senses, it is equally clear that there can be no effective vision, at the level of the model, unless specific elaboration has taken place able to "coagulate" the activity of "internal" selection. The outline offered by the model serves first of all to propose possible integration schemes able to support and prime the nesting proper to the "internal" selection. At the moment of the complete realisation of the embodiment, new vision by models emerges, and the outline as independent instrument is abandoned because superseded. In this sense, it is true that at the level of the eyes of the mind we finally have visual (and veridical) cognition, and not intellectual reading. Function and meaning articulate together, but in accordance with the development of a process of *adequatio*, and not of autonomous and direct creation. I will be unable to think of vision during emergence, but will be able to use it, once realised, to construct further forms of embodied cognition. Growth, modulation, and successive integration thus exist 'within and among' the channels together with specific differentiation processes.

This process can then gradually recognise itself in the realised emergence as an act of vision concerning the emergence itself. In this way a time of invention can be assured, but not a time of repetition: a time characterised by a specific process of renewal and recovery which continuously reveals itself as possible in proportion to the effective realisation of the "work" performed at the level of teleonomical activities. What determines the ongoing selection each time (with respect to the primary informational fluxes) is the new incompressibility which arises. This requires that the reference procedures posit themselves as an arch between invariance, on the one

hand, and autonomous morphogenesis on the other. In other words, they are only able to nurture new incompressibility where there exists a process of nesting of pure virtuality's original space. The important aspect is not, then, the remnant *in se* but the successful "narration". It is the effective and embodied inscription giving rise to new incompressibility which necessarily bypasses me. I will, then, ultimately be able to think of a new incompressibility which reveals itself as the ongoing fusion of emergent nuclei of creativity within the unity of an operant meaning.

It is far from easy to determine mathematics for processes of the kind, since it is clearly impossible to restrict the processes of self-reflection and assimilation totally within the limits of a mechanistic reductionism. Actually, the two involved selective forces are based on principles and choices which are articulated on a deep, productive level. Insofar as these principles and choices enter the scene, for example, at the second-order level, they cannot be previously determined at the first-order level; they are produced by the ongoing dialectics, by the symbolic dynamics in action and are revealed in emergence, i.e. when they really constitute me as the subject which sees and thinks. As for self-reflection, the space occupied by these choices, too, cannot be reductively determined: yet the thread must be untangled and the space explored. The mind has to function as a bridge between the two selective forces. This is the *Via*-Method, relying on the continuous invention of new mathematics, new geometry, new formal axioms, etc. Hence the importance of the eye of the phenomenologist, and in particular of the perceptologist, s/he who listens to the channels, and hence, at the same time, the importance of the eye of the mathematician, s/he who explores the thread of simulation as well as the path of the pure mental constructions. Amodal completion, for instance, in this context emerges as a privileged window opened on a microcosm which is largely articulated according to the fibres proper to the architecture of the mind. Objects are identified through the qualities elaborated and calculated along and through the channels. The function thus constructed that self-organises together with its meaning (in Atlan's words) permits a more coherent integration and articulation of the channels, laying the foundation for the self-organised synthesis of ever-new neural circuits. Objects, in their quality of being immersed in the real world, then emerge as related to other objects possessing different features, and so on. Through and beyond these interrelations, holistic properties and dimensions gradually reveal themselves, which I must grasp in order to see the objects with their meaning, if I am to understand the meaning of things. Apples exist not in isolation, but as objects on a table, on a tree: they are, for instance, in Quine's words, 'immersed in red', a reality I can only grasp by means of a complicated second-order process of analysis, elaboration, and comparison which can thereafter be reduced, through concatenations of horizontal and vertical constraints, specific rules and the successive determination of precise fixed points, to the first-order level. I thus need constant integration of channels and formal instruments to grasp information of the kind, i.e. to assimilate structural and holistic relations and relative ties in an adequate way. In other words, I will understand the meaning of things only if I am able to give the correct coagulum recipes with a view to their being selected so as to grasp and capture not only the superficial aspects of objects in the world, but their mutual relations as they interact in depth,

in obedience, in particular, to a specific intensional dimension. Here we can realise, as we have just said, the importance of the eye of the mathematician, s/he who explores the thread of simulation and the path of neural constructions in the regions of pure abstraction. In actual fact, if I want to understand how the assimilation process of structural relations works, I have first of all to make essential reference, from a mathematical point of view, to a specific theory of general structures. In the light of this theory the relations among individuals appear, from a general point of view, as submitted to a bunch of constraints, specifications and rules having a relational character, a bunch that is relative to the model which we refer to and which acts "from the outside" on the successive configurations of the first-order relations. In other words, as M. Manzano correctly states, in the universes of any second-order frame Ψ there are only relations among individuals, but it is no longer true that all the n-ary first-order relations on Ψ are into Ψ.

These hidden relations, these particular "constraints" play a central role with respect to the genesis of our models. In particular, let us remark that as a consequence of the action performed by these constraints, the function played by the individuals living in the original universe becomes more and more complex. We are no longer faced with a form of mono-dimensional relational growth starting from a given set of individuals and successively exploring all the possible relations among individuals, according to a pre-established surface unfolding of the relational texture. Besides this kind of mono-dimensional growth, further growth dimensions reveal themselves at the second-order level; specific types of development that spring from the successive articulation of the original growth in accordance with a well defined dialectics. As a result of the action of the rules lying at the second-order level, new dimensions of growth, new dynamic relational textures appear. Contemporarily the original universe of individuals changes, new elements grow up and the role and nature of the ancient elements undergo a radical transformation. The aforesaid dialectics reveals itself as linked to the utilisation of specific conceptual tools: limitation procedures, identification of fixed points, processes of self-reflection and self-representation, invention of new frames by "fusion" of previously established structures, coagulum functions etc. The plot of limitation procedures and cancellations of relations progressively constitutes itself as the gridiron of an intellectual order capable of allowing for the successive "production" (through the arising-irruption of new incompressibility and the successive "inscription") of specific *gestalten*, *gestalten* which, according to Monod, home the life and which, if enlightened by the truth, really support the development of rational perception. If we are able to recognise and follow the secret path of this order, we can finally manage to illuminate the "good" structures and to "read" (and "play") the progressive embodiment of that *Sinn* that selectively determines the real constitution of the events. Meaningful forms will then come into play, find reflection in a work, and be seen by an "I" that can thus construct itself and re-emerge, an "I" that can finally reveal itself as autonomous: real cognition in action. I neither order nor regiment according to principles, nor even grasp principles, but posit myself as the instrument for their recovery and recreation, and reflect their sedimentation in my self-transformation and my self-proposing as *Cogito*. Actually, I posit my work as the mirror for the new canalisation, in such a

way that the new emergent work (the self-organising "mirror"), if successful, can claim to be the work of an "I" which posits itself as an "added" creative observer. We are faced with a particular form of mental "exploration" that, if successful, "embodies" in an effective construction constraining the paths of our cognitive activity. As we have said before, this type of cognitive exploration articulates at the second-order level: it can be reduced however (if successful) at the level of many-sorted first-order logic, by means of well known logical procedures: hence the possible realization of an embodiment process.

In a nutshell, the nucleus of this kind of reduction consists in explicitly showing in many-sorted structures what is implicitly given in second-order or in type theory. In other words, we establish, via Henkin semantics, a form of reduction of second-order semantics to first-order semantics. Second-order logic with Henkin semantics is, in general terms, a many-sorted logic. However, we immediately have to emphasise that this kind of reduction does not imply that the secret "reasons" that guide, from within, the mental activity, the progressive unfolding of the processes of exploration and invention can be completely reduced to a first-order mechanism or to a set of pre-established rules. As a matter of fact, the first result of this very unfolding is the birth of specific (and previously unknown) differentiation processes, as well as the successive appearance of new universes of individuals. In this sense, there must be proofs that are not fully formalisable at a given stage in our mental experience, but that are "evident" to us at that stage on the basis of particular arrangements of limitation procedures, of the successive identification of fixed points, of the utilisation of abstract concepts, of the exploration of new universes of individuals, and so on. At the mental level, there are, for instance, proofs of Con (PA) (primitive arithmetic) that require abstract concepts as well as the necessary construction of new elements; concepts, for instance, that are not immediately available to concrete intuition (Hilbert's concrete intuition as restricted to finite sign-configurations). We need, in general, not only rules, but also rules capable of changing the previously established rules. In Gödel's consistency proof, for example, we can directly see that the theory of primitive recursive functionals requires the abstract concept of a "computable function of type *t*". Thinking in mathematical terms cannot be completely constrained within the boundaries of the syntax of a specific language. In fact, we would also need to know that the rules of this particular syntactical system are consistent. But in order to realise this, we need, by the second incompleteness theorem, to make reference to mathematics that is not captured by the rules in question. We have, in general, to utilise more and more abstract concepts in order to solve lower-level problems. According to Feferman (2006)[1], from a logico-mathematical point of view, and Carsetti (2000)[2], from a logico-epistemological point of view, the realm of mathematics must be considered as an open-ended domain not generated with respect to rules fixed in advance: we have to invent ever-new rules even if we

[1] Feferman S. (2006) The impact of Gödel's incompleteness theorems on mathematics, Notices of AMS 53, n. 4, pp 434–439.

[2] Carsetti A. (2000) Randomness, Information and Meaningful Complexity: Some Remarks About the Emergence of Biological Structures, La Nuova Critica, 36, pp 47–128.

are obliged, once the schematic principles employed reveal themselves as complete in a suitable sense, to act in accordance with them. In other words, the mathematical brain is able to constrain and "capture" the possible unfolding of natural causality as this is biologically expressed at the level of the newly-arising invariance and its propagation. The measures and the mathematical structures into play (the grid of mathematical "constraints" at work) determine how the different "entities" progressively emerge.

The utilisation at the semantic level of abstract concepts, the possibility of referring to the sense of symbols and not only to their combinatorial properties, the possibility of picking up the deep information living in mathematical structures open up new horizons with respect to our understanding of the ultimate nature of cognitive processes. At the mind level, in particular, we are actually dealing with a kind of categorial intuition (or rational perception) that does not concern simple data (relative to the inspectable evidence), but complex conceptual constructions. And we know that, in Husserlian terms, meaning "shapes" the forms creatively. However, we must immediately remark that categorial intuition appears to be embodied in a realm that is far beyond the limits of Gödel's primitive suggestions, in particular of his primitive Platonist approach. At the level of the articulation of mental constructions, we are actually faced with the existence of precise forms of co-evolution. Meaning selection is creative because it determines ever-new symbolic functions, ever-new processing units which support the effective articulation of new coherence patterns as well as specific embodiment processes. And it is precisely by means of these new patterns that we shall be able to "narrate" our inner transformation, to become aware of our mental development and, at the same time, to ascertain the objective character of the transformation undergone.

At the brain level and at the level of intuitive categorisation, we can perceive in turn the objective existence of abstract concepts only insofar as we transform ourselves into a sort of arch or gridiron for the articulation, at the second-order or higher-order level and in accordance with specific selective procedures, of a coherent series of conceptual plots and fusions, a series that determines a radical transformation of our intellectual capacities. It is exactly by means of the actual reflection on the new-generated intuitive constructions that I shall finally be able to inspect the realisation of my autonomy, the progressive embodiment of my cognitive activities in a "new" unitary system. At the level of Skolem's conception, for instance, ideas such as countability and uncountability are inherently relative: our belief that the power set of the natural numbers, $P(\omega)$, is uncountable is correct but must be understood relative to our own current viewpoint; from the point of view of another "observer", this set may in fact be considered as countable. From a more general point of view, we well know that there are some powerful characterisations of the system of natural numbers within an ambient set theory: according to Skolem's point of view, these set-theoretic characterisations are all relative. An internal observer, for instance, can find that in his/her world there is just one "system of natural numbers" satisfying Peano's second-order postulates. An external observer, however, can easily realise that this particular system is in fact non standard, containing infinite unnatural numbers. What it is important to underline

in this context, is the role played by the different observers and by the successive identification of the different ontologies. Things are even more complicated if we postulate, for instance, the existence of a circular link between the different observers in a co-evolutive ambient: the ontologies will undergo continuous changes. Then, according to this line of thought, we can effectively realise the importance of the progressive constitution at the co-evolutive level of the mind's eyes and the role played, with respect to this genesis, by the successive conceptual exploration of non-standard models. Actually, the complete opening of these eyes coincides both with the constitution process of an "I" as the "I" of an observer able to operate the "transversal identification "(Hintikka 1969[3]; Carsetti 1999[4]) and with the enlightenment on behalf of the truth of the mind proper to this very "I".

From a general point of view, while Gödel's theorem shows that sufficiently powerful systems of arithmetic are incomplete, Löwenheim–Skolem theorem (LST) shows that the real numbers cannot be specified uniquely by any first-order theory: in this sense, first-order theories with models appear importantly ambiguous: there can be plural models, plural interpretations in which the theorems come out true. No first-order system can fully capture the real numbers because of the ambiguity. Skolem discovered the existence of non-standard models of arithmetic in the 1930s. At the end of the 1940s Henkin utilised non-standard structures in order to prove his famous weak completeness theorem for the theory of types and, at the same time, outlined a non-standard model of $\mathbf{N}^2$. When we are in second-order logic, but we make essential reference to non-standard interpretations and allow structures with non-full relational universes, quantification only applies for the sets and relations that are present in the structure. In the general structures of Henkin, for instance, we put into the universes all sets and relations that are parametrically definable in the structure by second-order formulas. In this sense, it is not surprising that the set of standard numbers is not definable by a second-order formula in a structure having non-standard numbers. If we indicate with P. Def. (Ψ, L') the set of all parametrically Ψ – definable relations on individuals using the language L', we can say directly that a given frame Ψ is a general structure iff $\mathrm{D}_n = \mathbf{P}\,\mathrm{D}^n \cap \mathrm{P.\,Def.}\,(\Psi, \mathrm{L}')$.

What it is important to stress once again is the fact that hidden in the structure some specific relations exist, some "rules" (second-order relations) that cannot be defined as relations among individuals, but that are utilised to define first-order relations (i.e., relations among individuals). As a result, we obtain a particular structure where the n-ary relation universe is a proper subset of the power set of the n-ary Cartesian product of the universe of individuals. So, whereas in the standard structures the notion of subset is fixed and an n-ary relation variable refers to any n-ary relation on the universe of individuals, in the non-standard structures, on the contrary, the notion of subset is explicitly given with respect to each model. Thus, in the case of general structures the concept of subset appears directly related to the definition of a particular kind of constructible universe, a universe that we can

[3] Hintikka J. (1969) Models for modalities, Dordrecht, Reidel.

[4] Carsetti A. (1999) Mental constructions and non-standard semantics, La Nuova Critica, 33–34 pp 101–126.

explore utilising, for instance, the suggestions offered by Skolem (cf. his attempt to introduce the notion of propositional function axiomatically) or by Gödel (cf. Gödel's notion of constructible universe). In this sense, in accordance for instance with Németi's opinion, standard semantics is not logically adequate because it does not include all logically possible worlds as models. On the contrary, in Henkin's general semantics many "hidden" possibilities are progressively taken into consideration as possible models. We can, for instance, have models with or without GCH (generalised continuum hypothesis). Things are really different in the case of standard semantics.

In order to take into account and to face such a complex reality: i.e. to go into the paths of the inner structure of non-standard models and of their ultimate connections it appears suitable to resort, for instance, to the introduction of non-classical logics, to "creative" logics in particular, capable of elaborating in a finer-grained way the problem concerning the logical equivalence as well as the relationships existing between inductive inference and rational inference. It is according to these theoretical tools that the "life" existing in possible worlds seems, finally, to have the possibility of entering the stage of our awareness. If we want, for instance, to give a consistent explication of the meaning of linguistic expressions, of the deep information canalised by them, we have to situate these expressions within a general theory of meaning capable of giving an adequate explanation of the actual and global flow of real information. For the theorists of Situation Semantics, for example, the information flow concerns real things, living (and cognitive) entities which interact with their environment. Meaning lies in the systematic relationships existing or developing between different kinds of real situations. These crossed relations or constraints permit a given situation (an emergent phenomenon, in particular) to contain information concerning other different situations. The emergent phenomenon, in other words, is "captured by that which is describable in terms of the basic causal structure" (cf. S. Barry Cooper 2009, this volume) but with necessary and continuous reference to the models at work and to the complete unfolding of the canalization process. In addition we must postulate that, at the meaning level, information is "distributed" in a holistic way. In this sense, at the morphogenetical level it is the grid of measures in action that determines the quality of the emergent phenomena.

In this theoretical context the logic proper to a given natural activity as, for instance, visual cognition (and the correlated observational language) finally reveals itself as anchored to the set of constraints and meaning postulates in action that govern this very activity. This kind of logic contains, however, much more constraints and postulates than those of which we are aware as human beings. Within the existing Reality a deep information exists that partially escapes us, an information that can express itself only within the dynamic and coupled frame of a universe of constraints and postulates and that contemporarily appears as linked to a series of specific and continuous observational acts. As we have just said, seeing is observing with the mind's eyes in the light of the "irradiation" of the emergent meaning. The surfacing of meaning posits itself as an essential support of the government in action, it expresses itself as (and through) the logic concerning this particular (and natural) self-organisation process, as the grid of constraints that it co-determines and as the continuous renewal of this very grid.

The logical and inferential inquiry is precisely that particular type of cognitive activity that aims to explore "facts" in order to extract additional information implicitly contained in them, i.e. to open, in the first instance, the deep content of the original informational flow. This certainly does not exclude the validity of the utilisation of the rules of classical extensional semantics. These rules, however, concern only a particular sort of constraints, only some of the modalities that are necessary to pick out deep information. So, in order to collect additional information we have to explore and introduce further constraints, through the "intelligent" utilization of ever new methods (in particular, we have to close our flesh with respect to the "wounds" determined by Nature by means of a guided and meaningful "enumeration"): in a nutshell, we have to feed meaning in an adequate way. In particular, we have to feed the genesis of the Form constituting ourselves as prototypes and joining the emerging and irradiating grid. Hence the importance of a guided "*adequatio*". This *adequatio* does not concern simple things or given structures but the specific development of a capacity: only through an adequate construction of prototypes will it be possible to realise a more coherent expression of the government in action, only if we are able to join the secret grid according to a specific replication code, will we be able to feed meaning in an adequate way. Then truth will possess our minds: we shall finally be able to open the mind's eyes but in accordance with the truth, to constitute ourselves as minds in action. In this sense, we have to feed meaning in an "intelligent" and guided way, hence the importance of a correct identification of the Method, of the construction of adequate tools at the logical level. In order to see more and more I have to support a better canalisation of the original flow and to feed a more coherent "circumscription" at the meaning level. Hence specific forms will reveal themselves as natural forms through the progressive realisation of my embodiment: in order to join meaning and canalise the *Sinn* I have to "fix" the emerging flow into the genesis relative to the Form, I have to give life to specific prototypes and I have to recognise myself by identifying previously my role in and through them.

But, we may wonder: which paths do we have to follow at the cognitive level for it to be possible to carry out a conservative extension of the logical and semantical analysis at the level of a coupled system? In which way can we enter the mysterious kingdom of non-standard models? What is the role, however, played by the observer at the level of this particular kingdom? What about, for instance, the link between the observer's activity and the unfolding of the "nesting" process? How is it possible to realise a complete expression of the original *telos* considered as a needle in action outlining the drawing relative to that engraved path of the secret wounds that identifies the labyrinth and which finally recognise himself in this very path? In a nutshell: what methods do we have to adopt and follow in order to see and think more deeply according to the truth?

Chapter 1 aims at a very clear exploration of the role played by the models at the level of Cognitive Science and in particular at the level of visual cognition. According to Grossberg, the brain is organised in parallel processing streams. These streams are not independent modules however: as a matter of fact strong interactions occur between perceptual qualities. Actually, we experience the world as a whole.

"Although myriad signals relentlessly bombard our senses, we somehow integrate them into unified moments of conscious experience that cohere together despite their diversity. Because of the apparent unity and coherence of our awareness, we can develop a sense of self that can gradually mature with our experiences of the world. This capacity lies at the heart of our ability to function as intelligent beings".[5]

Each stream can possess multiple processing stages, a fact which, according to Grossberg, suggests that these stages realize a process of hierarchical resolution of uncertainty. The computational unity is thus not a single processing stage but a minimal set of processing stages that interact within and between complementary processing streams. "The brain thus appears as a self-organising measuring device in the world and of the world".[6]

Starting from the ART Hypothesis: All Conscious States are Resonant States, Grossberg aims to suggest a possible outline of brain's global functioning. Such an analysis is not easy because it requires that one have knowledge of a multiplicity of disciplines. For example, at the level of brain organization it is the network that determines behavioral success. However, one needs to properly define the individual nerve cells and their interactions in order to correctly define the networks whose interactive, or emergent, properties map onto natural behavior. In order to realise this difficult program we need a sufficiently powerful theoretical language. The language of mathematics has proved to be the relevant tool, indeed a particular kind of mathematics. All of the self-adapting behavioral and brain systems that Grossberg introduces are nonlinear feedback systems with large numbers of components operating over multiple spatial and temporal scales. As Grossberg remarks "The nonlinearity just means that our minds are not the sum of their parts. The feedback means that interactions occur in both directions within the brain and between the brain and its environment. The multiple temporal scales are there because, for example, processes like STM are faster than the processes of learning and LTM. Multiple spatial scales are there because the brain needs to process parts as well as wholes".[7] With respect to this, a second important metatheoretical constraint derives from the fact that no single step of theoretical derivation can derive a whole brain. One needs to have a method that can evolve with the complexity of the environmental challenges that the model is forced to face. "This is accomplished as follows. After introducing a dynamic model of a prescribed set of data, one analyzes its behavioral and brain data implications as well as its formal properties. The cycle between intuitive derivation and computational analysis goes on until one finds the most parsimonious and most predictive realization of the organizational principles that one has already discovered".[8] Such a theoretical analysis also discloses the shape of the boundary, within the space of data, beyond which the model

[5] Chapter 1, p 3.

[6] Grossberg S (2000) Linking mind to brain: the mathematics of biological intelligence, Notices of AMS, p 1364.

[7] Chapter 1, p 7.

[8] Chapter 1, p 7.

no longer has explanatory power. The shape of this boundary between the known and the unknown can often clarify what design principles have been omitted from the previous analyses.

The chapter is full of ideas and new methodologies. Let us just remark that, from an epistemological point of view, simulation models no longer appear, in this theoretical context, as "neutral" or purely speculative. True cognition, on the contrary, appears to be necessarily connected with successful forms of reading, those forms that permit a specific coherent unfolding of the deep information content of the Source. Hence the importance of taking into consideration both the interplay between the observer and the real world and the role played by intentions at the level of this mysterious and continuous unfolding.

The following two chapters are precisely centered on an in-depth analysis of the emergence of intentional procedures and goal representations at the level of neural networks as well as at the level of the cerebral cortex, although according to different theoretical and modelistic perspectives.

In Chapter 2, Atlan and Y. Louzoun aim to "analyze under which conditions a positive answer could be given to the following question: can neural networks self-organize so that not only structures and functions not explicitly programmed emerge from their dynamics, but also goals for intentional actions, set up and achieved by themselves? Such mechanistic models of intentional self-organization are useful in that they allow to circumvent the usual circular explanation of intentionality by causal effects of assumed intentional mental states on bodily movements".[9]

The authors take into consideration intentionality in a pragmatic sense as it is observed in intentional actions to solve two problems of causality: the apparent time inversion involved in final causes and the "mind–body" causal relationship involved in the usual picture of a mental state being the case of bodily movements and actions.

The system that they develop is designed to devise new goals by itself and to reach these goals. According to the authors "The goals are determined by the capacity of a network to learn a relation between effects and the events that caused them. The model is a metaphor for the psychophysical goal learning process in cognitive beings. This process involves the ability to predict rapidly the result of a set of events, so that an initial event is reproduced knowing its expected result. In other words, prediction (which is knowledge) and intentional action are closely related. That is why this capacity is modeled using a non-supervised learning network associated to a recurrent neural network. However, while the prediction capacity is obviously based on memory of previous experience, this knowledge must be allowed some degrees of freedom, which produce new predictions of new events and the achievement of new goals".[10]

In accordance with the model, intention and action appear to be one and the same realization, simply represented in different ways. This implies that an intention to act is always normally associated with its execution. In other words, both the

[9] Chapter 2, p 47.

[10] Chapter 2, p 47.

action and the intention are represented by links between initial and final states. The difference between the action and the intention is actually the difference between an action actually performed and its initiation, as indicated by neurophysiological data. "This difference results in our capacity as human beings to stop an action once initiated. We would call an action interrupted after being initiated, an intention to do an action and invent a mental state to represent it".[11] This view, as the authors correctly remark, is opposed to the usual mentalist assumption that an intention exists first in the mind as a "pure" mental state, possibly, but not normally associated to its execution.

One feature of these views is the monist ontology involved in the approach to the mind–body problem. Spinozist philosophy is certainly the most radical monist attitude towards this problem. As is well known, Spinoza denies the possibility of causal relationships between the mind and the body, not because they would pertain to two different substances, as in Descartes, but precisely because they are "one and the same thing, though expressed in two ways" (*The Ethics*, II, 7, note).

In this sense, according to the authors, the cause of a voluntary bodily movement must always be some previous bodily (brain) event or set of events, and not a conscious decision viewed as a mental event as described by subjective reports about conscious experiences. The difference from a non-voluntary movement is the nature and degree of conscious experience accompanying it. But in any case, a conscious mental event in this context may accompany the brain event but not be its cause, being in fact identical with it, although not describable in the same language. At the end of this very incisive chapter, the authors finally remark that results from neurophysiology support this view (cf. Libet 1992): unconscious initiation of voluntary action precedes the conscious decision to trigger the movement.

In Chapter 3, L. Fogassi outlines how imitation is the first function that comes to mind when one thinks to the possible use of mirror neurons, because they possess the property enabling the observer to immediately translate the visual information on observed action into the motor parameters necessary for reproducing it. In his opinion, from a general point of view, imitation in humans appears to require the involvement of the mirror neuron circuit, with the additional activation of prefrontal areas when recombination of already existing motor representation in novel sequences is required.

"It remains to be explained how imitation in monkeys is minimal, in spite of the presence of a well-developed mirror neuron system. There are probably many reasons for this apparent contradiction. First, in monkeys a lower percent of mirror neurons show a strict congruence between observed and executed action, the majority coding the action goal. Second, as shown above, in humans a crucial role in imitation learning is played by the prefrontal cortex, a region that is much more developed in the human brain in respect to that of monkeys".[12]

The mirror neuron mechanism appears to be very close to the mechanism that, during inter-individual communication, enables the listener/observer to understand

[11] Chapter 2, p 50.

[12] Chapter 3, p 66.

the meaning of the message emitted by the sender. The central point is that both sender and receiver share the same motor programs necessary to produce a message and the pathway that allows to access these programs. As Fogassi remarks "The proposed homology between F5 and Broca's area is in favor of the idea that language can be derived from a system involved in action and, lately, in gesture understanding..... All these data corroborate the idea that an ancient observation/execution matching system, as that found in monkeys, may have paved the way to the evolution of human language. This process occurred through many steps, two of which, however, are assumed to be very important. The first is the transition from a motor system coding actions to one with the capacity to encode also intransitive actions, probably through a process of ritualization of goal-directed actions. ..... The second is represented by the association between a gesture and a sound. The possibility to use facial and brachiomanual gestures in association with utterances provides a higher combinatorial power, allowing to create a richer vocabulary. The presence in monkey area F5 of a large population of neurons coding both hand and mouth actions and its access to auditory input could have been important elements, in evolution, for facilitating the occurrence of the proposed association gesture/action-sound".[13]

According to Fogassi, when we observe somebody else performing goal-directed action, in most cases we are able to infer the intended goal, even though the action is not completely accomplished: we really have the capacity to understand the intention of other individuals. Since mirror neurons provide a mechanism to understand the goal of motor acts performed by others, it is natural to raise the issue of whether they can also play a role in intention detection. In a recent experiment, the visual response of parietal mirror neurons was studied in the same conditions that were used for studying motor properties of IPL (inferior parietal lobule) grasping neurons.

On the basis of these experimental results, the IPL mirror neurons, in addition to recognizing the observed motor act, appear also able to discriminate among identical motor acts according to the context in which they are executed. "Because the discriminated motor acts are part of chains, each of which leading to a specific final goal, this capacity allows the monkey to predict what is the goal of the observed action and, in this way, to "read" the intention of the acting individual. If grasping neurons belonging to a particular chain fire, the observed acting individual is going to bring the food to the mouth; if, in contrast, another set of grasping neurons belonging to another chain fire, the observed acting individual is going to put the food away".[14]

Fogassi lastly affirms, in accordance with his central thesis, that the mirror neuron system in monkeys provides the first neural substrate for a primitive understanding of other intentions, that probably paved the way for the evolution of the more sophisticated aspects of mind reading present in humans. Thus, once again intentions and actions appear indissolubly linked. How is it possible, however, to identify new mathematical languages able to enlighten this mysterious interplay?

[13] Chapter 3, p 67.

[14] Chapter 3, p 68.

The second part of the volume is precisely devoted to a thorough analysis of a number of conceptual and mathematical tools that in the last decades revealed themselves particularly useful in interpreting cognitive and mental phenomena. Deterministic chaos, incompleteness results, the genealogical analysis of the mathematical structures etc., are extensively utilised in the different chapters in order to clarify both the mysterious relationships between truth and randomness and the real interplay between the emergence of intentionality and the self-organisation processes involved in intuitive categorisation. Actually, in order to outline more sophisticated models of cognitive activities (and in particular of that inextricable plot constituted by the circular link between "rational perception" and "intuitive categorization") we have to examine and individuate specific theoretical methods also capable, for instance, of taking into account the intentional and semantic aspects of that particular biological (and neural) process linking together growth with differentiation which characterises human cognition.

D. van Dalen in Chapter 4 starts from the analysis of Brouwer's mathematical universe. As is well known, according to Brouwer the objects of mathematics come first in the process of human cognition, and description and systematization (in particular logic) follow later.

In the final presentation, *Consciousness, Philosophy and Mathematics* (CPM) (Brouwer 1949), the great mathematician expresses explicitly his thought in this way: " 'By a move of time a present sensation gives way to another present sensation in such a way that consciousness retains the former one as a past sensation and moreover, through this distinction between present and past, recedes from both and from stillness and becomes mind.' Thus the subject has created a 'twoity' of a past and present sensation. The process evidently can be iterated, and complexes and strings of sensation become the object of attention. The sensation complexes form a bewildering mixture, in which a certain order is introduced by the *causal attention*. This carries out a process of *identification*. One may think of the identification of 'similar' complexes, or of *abstraction*".[15]

In CPM the notion of causal sequence is further refined: " 'An iterative complex of sensations whose elements have an invariable order of succession in time, whilst if one of its elements occurs, all following elements are expected to occur likewise, in the right order of succession, is called a causal sequence'. It might be tempting to explain these, let us say 'strongly causal sequences', *scs*, by a causality, independent of the will of the subject. This, however, is rejected by Brouwer. On the contrary, causality is explained by the notion of strong causal sequence. A scs can be put to use by the subject in order to realize events that are not immediately obtainable. One only has to realize the first event of a scs, or an intermediate one, in order to obtain the final event. The procedure of realizing the final (and desirable) event by realizing a preceding event was called the '*jump from end to means*', and later the *mathematical* or *cunning act*".[16]

[15] Chapter 4, p 77.

[16] Chapter 4, p 78.

In this way, the basic material of "discrete mathematics" is at the disposition of the subject. This part of the process of creating is later called *the first act of intuitionism.* On the contrary, as van Dalen remarks, the continuum is given in the move-of-time act as the 'between'. "In his Rome lecture (1908) Brouwer explicitly points out that 'the first and the second are thus kept together, and the intuition of the continuum (*continere* = keeping together) consists of this keeping together'. And he adds: 'This mathematical ur-intuition is nothing but the contentless abstraction of the sensation (experience) of time'. Time is thus created by the subject through the 'move of time', together with the continuum and the natural numbers. *The second act of intuitionism* is the creation of 'more or less freely proceeding infinite sequences of mathematical entities previously acquired' and of 'species', i.e. 'properties supposable for mathematical entities previously acquired' ".[17]

On the basis of his deep knowledge of Brouwer's universe, van Dalen first of all points out that "In CPM the two acts are tacitly lumped together under the act of 'unlimited unfolding' ".[18] In any case, the process of creation of causal sequences and complexes does extend beyond the realm of mathematics; indeed the physical world, as well as the social one is made up of those objects. If we look at the physical phenomena within the boundaries of a Brouwer's universe, then we can individuate the role of mathematics as follows: the objects of the physical world are obtained by abstraction from sensation complexes, a further abstraction gets the subject to mathematical objects and structures. Hence a natural and priviliged connection between the physical universe and the mathematical one.

As is well known, Weyl adopted Brouwer's intuitionistic programme, adding his own interpretations to it. In particular Weyl did not give the same status to choice sequences Brouwer did. As the author clearly remarks for Weyl choice sequences did not belong to mathematics proper; all he accepted was the real status of initial segments. As a consequence arbitrary reals were replaced by generating intervals. ".....an interval, say (a, b) for rational a and b, represents for Weyl the open horizon of 'the reals that are potentially given by the interval'. Concrete real numbers are given by lawlike sequences of intervals, and arbitrary ones by choice sequences, in the representing interpretation. Hence there is on Weyl's approach a fundamental distinction between existential quantification (over lawlike reals), and universal quantification (over choice reals)".[19] Here Brouwer's and Weyl's roads separated. The words by van Dalen about the final difference between Brouwer's universe and Weyl's universe are particularly important in order to understand the successive development of the concept of extended Turing universe as this concept is presented in the following chapters. As Chaitin, for instance, points out in his chapter, there is a precise link between Weyl's conception and Popper's first analysis of the concept of simplicity. The first theoretical bases of AIT are envisaged by Popper by means of the outlining of a new kind of relationship between humans and Nature that maintains some original Weyl's suggestions.

[17] Chapter 4, p 78.
[18] Chapter 4, p 79.
[19] Chapter 4, p 80.

In Chapter 5, G. Longo starts from an in-depth analysis of the link between randomness, determinism and knowledge in order to discuss, from a logico-epistemological point of view, Turing's original contribute. Once again the problem is represented by the confrontation between the machine and the continuum. As Turing understands very well, 'the nervous system is surely not a DSM': on the contrary, in Longo's opinion, we can affirm that the brain rather is a dynamical system (and Longo correctly remarks that Turing calls these systems "continuous"). Then, how to compare a DSM with the brain? The comparison is functional and relative to the only possible access to the machine, during the imitation game: the finite sequences of a teleprinter's signs. Under these conditions, according to Turing, we would be unable to distinguish a continuous system, as the brain, from a DSM; if the continuous machine makes its response though a printer, it will be undistinguishable from a DSM's response, even if obtained by different means (continuous variations instead of discrete steps). Hence the Turing's central hypothesis: if the interface with the dynamical system is given by a "discrete access grid", then it will be undistinguishable from a DSM.

"In fact, today's physical DSM, our computers, simulate dynamical systems in a more than remarkable way. They develop finite approximations of the equations which model them with great efficiency: nowhere may we better see the "form" of an attractor than on the screen of a powerful enough machine. Their applications to aerodynamics (simulation of turbulence), for example, has considerably lowered the price of airplanes (almost no more need for wind tunnels). But ... what are the conceptual, mathematical, physical differences?".[20]

From a general point of view, Longo firmly states that a DSM is surely not a model of the brain, at least if we consider the latter, with Turing, a continuous system, as opposed to what is pleaded in the field of classical Artificial Intelligence and by many modern cognitivists. However, the real problem is the following: can a DSM imitate the brain? According to Longo "Turing is perfectly aware of the difference between imitation and mathematical modeling for a quite simple reason: he is already working upon a remarkable mathematical model of morphogenesis in a field of chemical diffusion .... In fact, the most interesting property the equations to be found in (Turing 1952), is that a very small variation of the boundary conditions, obviously in a continuous system, can radically change the evolution of the model. And this property is not the laplacian nondeterminism or randomness, but the sensitivity to the contour conditions and situates itself at the heart of the deterministic model of morphogenesis à la Turing. One thing is thus the "imitation game", another mathematical modeling of physical and physico-chemical or biological phenomena: the Turingian DSM does not claim to model the brain, in the physico-mathematical sense – the latter is a continuous system for Turing – it can only attempt to trick an observer".[21]

As a matter of fact, an abstract, mathematical DSM, such as Turing's machine, is not conceived as a physical machine, but as a logical machine, a human in "the

[20] Chapter 5, p 90.

[21] Chapter 5, p 98.

minimal act of thought" – of formal thought. Consequently, its expressivity is mechanical yet purely logico-formal: typically, its expressive power is independent of spatial dimensions, a property absolutely foreign to the physical processes, which all depend and strongly upon the dimensions of space. "Let's forget the comparison between formal DSMs and living machines, which are physical, obviously, but are moreover subject to phenomena of integration-regulation which keep them in an "extended critical state"; this state is unknown by the non-living and its mathematics; mathematics which must therefore be extended and adapted to the new job (dynamical systems are "only" one of the best approximations we have, for the moment). It is exactly this integration of the brain within a body, their reciprocal regulation and by such a rich environment that confers it a quite peculiar structural and functional stability; and when these regulative/integrative linkages by/of/in a body are weakened – in the course of a dream for example – the brain appears to be rather unstable (likewise in case of serious deprivation – artificial, for example – from sensation)".[22]

In spite of the difference between a DSM and the brain, the distinction hinted by Turing (a distinction that is at the heart of Longo's analysis) between modeling (as mathematical proposal of constitutive principles for a physical process) and imitation (functional imitation, with no commitment on the "nature" of phenomena) is a fundamental idea. It should be taken up today, both from a foundational and practical view point, as discrete-state machines are essential to modern science by their extraordinary' modeling/imitation abilities.

Let us underline, however, as Longo correctly points out, that beside the imitation, when we look at brain's functioning also simulation procedures and intentional decisions take on a decisive role not only with respect to brain's functional architecture but also with respect to the continuous growth of its inner complexity, and to the full expression of its real plasticity.

What about, however, the possibility to model in mathematical terms the role played by intentionality? It is precisely to this general problem that the successive chapter is devoted.

From a technical point of view, the main subject of Chapter 6 by S. Galvan is the ω-incompleteness of a formal theory which seeks to formalize finitist arithmetic. PRA (i.e. primitive recursive arithmetic) is normally considered to be the theory that formalizes finitist arithmetic. But the arguments which the author illustrates also hold if one assumes PA (i.e. Peano arithmetic) as the theory formalizing finitist arithmetic (in a broader sense, of course). Galvan adopts two points of view: one internal to the theory, and one relative to some suitable non-conservative extension of it. He seeks to show that (i) with respect to the first point of view, ω-incompleteness entails an irreducible distinction between truth in finitist arithmetic and provability through methods based on finitist (finitary and concrete) evidence; (ii) with respect to the second point of view, this irreducible distinction can be overcome, but only if one accepts a form of evidence (non-finitary with respect to content, finitary in form

[22] Chapter 5, p 99.

but abstract). Abstract evidence appears thus, in his opinion, as the finite expression of an intensional relationship between the subject and an infinite reality.

According to Galvan the main problems arising from the attempt to consider intuition as a way of access to mathematical reality and, therefore, as a modality of justification of the mathematical sentences themselves, concern, first of all, the possible examination of two different types of intuition: intuition of an arbitrary or abstract natural number and intuition that allows to introduce structures with countable support, i.e. omega-structures. The main questions regarding these different types of intuition are: (a) What formal theories are justificable by intuition? (b) What is the difference between justification by intuition and justification by proof in the context of more powerful formal theories? (c) How is the assertion to be understood that the hierarchy of induction principles measures the degree of complexity of the corresponding forms of numerical intuition? Another order of problems concerns intuition of the standard model of numbers. How such an intuition is possible? Does the use of a second-order language guarantee the possibility of representing linguistically such a model? But if neither a second-order linguistic dimension is sufficient to this aim, what characterizes the 'surplus' present in intuition? How can be consistently argued that intuition of the notion of a standard number is emergent on the syntactic and even on the semantic dimension of numerical theories?

What about, moreover, the connection between intuition and intentionality? In other words, what does the abstractness of non-finitist evidence have to do with intentionality? In the final part of his paper Galvan affirms that forms of non-finitist evidence have a distinctive intentional character in the classical sense. "But what does intentionality in the classical sense mean? In the contemporary theory of knowledge, by 'intentionality' is normally meant the relation, inherent in every activity by a subject, of being oriented to an objectual content. Of course there are very different opinions on whether some activity or other is oriented to an object and is therefore intentional. However, intentionality consists in directedness to an objectual content. ..... What matters in this relation is not so much the identity (which simply expresses the fact that the subject enters into 'contact' with the object) as the fact that the object is grasped (received) by the subject as something else (*aliquid aliud*)".[23]

Intentionality considered as simple directedness at the object can in fact be interpreted as a causal relation on behalf of the object which exerts a stimulus on the subject which is then processed by the subject himself/herself. In this case, directedness is determined by the fact that not all stimuli are processed, but only those which match the structures responsible for stimuli apprehension and processing. Non-finitist evidence therefore requires the activating of this capacity for intentioning the mathematical object which is realized in the multiple forms (visual, geometric, combinatorial, set-theoretic, etc.) of the being present, of the being seen, in a word, of the appearing. This capacity, in Galvan's opinion, is to be understood in terms of intentionality of consciousness, and intentionality – in as much as it is the place where the object is present to consciousness – is just what mechanical minds lack. In

[23] Chapter 6, p 123.

which way, however, can we find the possibility to hear from a Source which comes out to dictate at the level of biological structures the message of its wild autonomy? How can we fix the "code" of this mysterious transmission?

In the third part the central core of the analysis is represented by the definition of a multiplicity of concepts intersecting many different realms of contemporary scientific research: Algorithmic Information Theory, Computability Theory, Measurement Theory, Alternative Set Theory and so on. After introducing in the second part the Brouwer universe and the Turing universe as well as the multiple facets of intentionality the chapters now focus first of all on the extended Turing universe and on the link between incomputability and incompleteness. The utilisation of oracles as well as of specific non Cantorian tools allows for a real enlargement of the analysis and an exploration of the specific power proper to the "mathematical brain". For many aspects, the different chapters underline the necessity of introducing a more subtle analysis of natural causality by means of the tools offered by meaningful complexity.

In Chapter 7, G. Chaitin starts with a revisitation of some fundamental ideas as proposed by Weyl and Popper: "Weyl observes that this crucial idea of complexity, the fundamental role of which has been identified by Leibniz, is unfortunately very hard to pin down. How can we measure the complexity of an equation? Well, roughly speaking, by its size, but that is highly time-dependent, as mathematical notation changes over the years and it is highly arbitrary which mathematical functions one takes as given, as primitive operations. Should one accept Bessel functions, for instance, as part of standard mathematical notation? This train of thought is finally taken up by Karl Popper in his book *The Logic of Scientific Discovery* (1959), which was also originally published in German, and which has an entire chapter on simplicity, Chapter VII. In that chapter Popper reviews Weyl's remarks, and adds that if Weyl cannot provide a stable definition of complexity, then this must be very hard to do. At this point these ideas temporarily disappear from the scene, only to be taken up again, to reappear, metamorphised, in a field that I call algorithmic information theory. AIT provides, I believe, an answer to the question of how to give a precise definition of the complexity of a law. It does this by changing the context. Instead of considering the experimental data to be points, and a law to be an equation, AIT makes everything digital, everything becomes 0s and 1s. In AIT, a law of nature is a piece of software, a computer algorithm, and instead of trying to measure the complexity of a law via the size of an equation, we now consider the size of programs, the number of bits in the software that implements our theory: **Law**: Equation → Software, **Complexity**: Size of equation → Size of program, Bits of software".[24]

According to Chaitin's model, both the theory and the data are finite strings of bits. A theory is software for explaining the data, and in the AIT model this means the software produces or calculates the data exactly, without any mistakes. In other words, a scientific theory is a program whose output is the data, self-contained software, without any input.

[24] Chapter 7, p 129.

But "what becomes of Leibniz's fundamental observation about the meaning of "law?" Before there was always a complicated equation that passes through the data points. Now there is always a theory with the same number of bits as the data it explains, because the software can always contain the data it is trying to calculate as a constant, thus avoiding any calculation. Here we do not have a law; there is no real theory. Data follows a law, can be understood, only if the program for calculating it is much smaller than the data it explains".[25]

In this sense, understanding is compression, comprehension is compression, a scientific theory unifies many seemingly disparate phenomena and shows that they reflect a common underlying mechanism. The best theory, in Chaitin's opinion, is the smallest program that produces that data, that precise output. This can be considered as a variant of Occam's razor. As the author affirms, this approach enables us to proceed mathematically, to define complexity precisely and to prove things about it. There are, however some precise proviso: "once you start down this road, the first thing you discover is that most finite strings of bits are lawless, algorithmically irreducible, algorithmically random, because there is no theory substantially smaller than the data itself. In other words, the smallest program that produces that output has about the same size as the output. The second thing you discover is that you can never be sure you have the best theory".[26] As is well known, in Chaitin's opinion, Ω is a random real with lots of meaning but this information is stored in Ω in an "irreducible" way, with no redundancy.

What about, however, the ultimate role of meaning at the level of AIT? In which way can we model the link between incompressibility and irreducibility at the morphogenetical level? How can we explore, in accordance with Leibniz's original suggestions, another kind of model: the extended Turing universe?

Some of the fundamental ideas that are at the basis of AIT are revisited, for many aspects, by S. Barry Cooper exactly by means of a merging of these very ideas in an extended Turing universe. In particular, the view that Cooper wants to pursue in Chapter 8 is "that emergent phenomena not only yield up descriptions, using different language to that used in describing the underlying design; they are actually determined, constrained, *captured* by that which is describable in terms of the basic causal structure".[27] The intuition that entities exist because of, and according to, mathematical laws, is not new, of course, as Chaitin extensively shows in his chapter. One can detect it in the words of Leibniz from 1714 in the *The Monadology*, section 32: "there can be found no fact that is true or existent, or any true proposition, without there being a sufficient reason for its being so and not otherwise, although we cannot know these reasons in most cases".

According to Cooper, natural phenomena not only generate descriptions, but arise and derive form from them. So connecting with a useful abstraction, that of mathematical definability – or, more generally, invariance (under the automorphisms

[25] Chapter 7, p 130.

[26] Chapter 7, p 130.

[27] Chapter 8, p 142.

of the appropriate structure). "This gives precision to our experience of emergence as a potentially non-algorithmic determinant of events. On the one hand one can attempt to frame criteria for emergence in terms of the complexity of the language used to describe it, and one can also use the known associations between informational and computational complexity to constrain the computability-theoretic character of physical phenomena".[28]

What one would expect from this very clear connection between the underlying basic causal structure (the 'design' in Cooper's terms) and the emergent phenomenon would be a certain level of robustness of the emergence. "What one is suggesting, via the association with mathematical definability, is a direct causal relationship between 'design' and emergent phenomenon – and one which is unlike the usual fundamental laws of nature, in that it is more global in respect of the causes it works with – and potentially, with respect to the effects".[29]

As Cooper remarks, Turing's approach is largely proof-theoretic, growing out of his interest in Gödel's incompleteness theorem, and what it tells us about the extent of the boundaries of the computable world. Turing shows that despite Gödel's proof that no consistent first-order theory captures arithmetic, we can hierarchically transcend this barrier, in a quite constructive way – one just iterates the Gödel argument, computably generating new unprovable theorems which are then used to enlarge the theory. One uses computable ordinal notations to iterate this process into the transfinite in a constructive way, thus giving the appearance of computably transcending Gödel's theorem. "But a little thought reveals the snag – identifying the route to a new theorem involves using an incomputable oracle, so we avoid the reductionist paradox".[30]

Having tried unsuccessfully to 'compute the incomputable', Turing introduced a model of natural causality between real data, which could be incomputable. The model – now called an oracle Turing machine – was essentially just a Turing machine which could ask questions of an external 'oracle' (usually a set of natural numbers). The number of questions during a particular computation was finite, of course. "The result was that instead of getting computable real numbers via the collating of computational outputs of a machine, one now got real numbers computable relative to an oracle. Considering the oracles to be inputs, a given machine might capture a particular computable function over the reals, notated as a Turing functional from reals to reals..... This is not surprising, since such simple basic transformations are routinely captured via functions over the reals which can be computed up to any practicable level of approximation by a real-world computer. Here we have again basic computability leading very quickly via descriptions to a situation with computational content, but not necessarily computable".[31]

With respect to this context, Turing's oracle machines precisely provide a model of computable content of structures, based on partial computable functionals over

[28] Chapter 8, p 143.

[29] Chapter 8, p 143.

[30] Chapter 8, p 146.

[31] Chapter 8, p 147.

the reals. As Cooper remarks, this model – the Turing universe – is really capable of capturing basic computable causal structure in the real world, with the expectation, based on experience, that any incomputable causality would be definable in some natural way from this basic structure.

The general (and thoughtful) intuition underlying Cooper's considerations is that the Turing invariant relations are key to pinning down how basic laws and entities emerge as mathematical constraints on causal structure. "At one time, it was thought that the structural pathology exhibited by the Turing universe, and the disproportionate technical difficulty of proofs in the area, was evidence of mathematical ugliness, disqualifying the field from serious attention of non-specialists. It is now understood that the richness of Turing structure discovered so far provides the raw material for non-trivially defining a multitude of relations. And that the complexity and pathology of the structure is only what one would expect of something aiming to model global aspects of the real world".[32]

In accordance with the afore mentioned general intuition, in Chapter 9 E. Beggs, J. F. Costa, and J. V. Tucker develop a mathematical theory about using physical experiments as oracles to Turing machines. They suppose first of all that an experiment makes measurements according to a physical theory and that the queries to the oracle allow the Turing machine to read the value being measured bit by bit. Using this theory of physical oracles, an experimenter performing an experiment can be modelled as a Turing machine governing an oracle that is the experiment. In particular, the authors consider this computational model of physical measurement in terms of the theory of measurement of Hempel and Carnap and finally note that once a physical quantity is given a real value, Hempel's axioms of measurement involve undecidabilities. To solve this problem, they introduce time into Hempel's axiomatization. Focussing on a dynamical experiment for measuring mass, they finally show that the outlined computational model of measurement satisfies their generalization of Hempel's axioms. This analysis also explains undecidability in measurement and that quantities are not always measurable.

From a general point of view, the authors develop a methodology and a mathematical theory to examine how data is represented and computations are performed by physical systems. In particular, as we have just said, they introduce a Principle which changes the perspective of the mathematical theory of Turing machines with physical oracles. Instead of viewing the experiment as an oracle boosting the power of Turing machines, they view the Turing machine as controlling and, indeed, performing the experiment. Specifically, this Principle leads to suppose that: The Turing machine models a human experimenter conducting the experiment.

So the relationship between experimenter and experiment is modelled by the protocols that apply to the oracle queries. A question, however suddenly arise: To what extent is this computational model of experimentation general? And, in general, what is measurement?

[32] Chapter 8, p 148.

In their important paper the authors begin to explore these questions with the help of the philosophy of physics. They relate their computational model to the desiderata of Geroch and Hartle for an investigation into computable aspects to measurement. Then they consider the axiomatic theory of measurement established by Carl G. Hempel and elaborated by Rudolf Carnap and apply it to their computational models of measurement. In particular they introduce the operational concept of computational resources, specifically time, into Hempel's axioms. "The idea of considering time as a cost in deciding the equality of measurements is suggested by our previous technical work on the model".[33]

But, how does the Turing machine communicate with Nature? The authors propose "that this interaction is captured by the concept of the continuing evolution of a physical experiment acting as an oracle.... The measurement apparatus is taken to be an oracle to a Turing machine. The interaction is achieved through a protocol which counts time. After each consultation, the oracle may provide one bit of the measurement. This bit also provides the necessary information to the machine to proceed with the experimental procedure".[34] These technical results can be used to show that the task of measuring quantities in physics can be classified by well known complexity classes. In this sense, will a TM model precisely be a human experimenter only if it is able to calculate the complexity classes in an adequate way. The chapter opens up new theoretical horizons: in actual fact, according to Calude a TM with an oracle of quantum random bits has hypercomputational power. But, how powerful is such a machine?

In Chapter 10 S. Livadas starts from an accurate revisitation of Husserlian doctrine. As is well known, Edmund Husserl held the early idea that pure mathematics belongs to the exact sciences dealing with idealities whereas phenomenology is a descriptive eidetic science of pure mental processes as viewed in the phenomenological attitude. They are fundamentally different in that they use both different cognitive tools and turn their view to essentially different objects. This is Husserl's prevalent attitude to which he makes references especially in *Ideen I*, where he supports that they can combine though they cannot take the place of one another.

Livadas aims to demonstrate how the phenomenological analysis of time consciousness can not only provide a model of the intuitive continuum, something that had already attracted, as we have just seen, the theoretical interest of prestigious mathematical names as that of H. Weyl and L. E. J. Brouwer in early twentieth century, but can also motivate a new approach to the ontological nature of intuitive continuum and its ad hoc axiomatization in the language of non-Cantorian theories. On a phenomenological level, Livadas starts from the analysis of the phenomenological constitution of time as it is developed in Husserl's *Phänomenologie des inneren Zeitbewußtseins* (Husserl 1996) and of the work of J. Patočka (1992) as well as of the more general Husserlian idea of genetic-kinetic constitution. As is well known, following this analysis Husserl confronts in *Phänomenologie des*

[33] Chapter 9, p 156.

[34] Chapter 9, p 168.

inneren Zeitbewußtseins the issue of a transcendental, non-temporal subjectivity objectivated in the self-constituting unity of the flux of consciousness which in a somehow circular turn is successively considered as constituted in accordance with a kind of transcendental "genesis" constantly generating temporality. Becoming convinced that the transcendental *ego* is given in temporal profiles – "time is the universal form of all egological genesis" he professed in the Fourth Cartesian Meditation – he was inducing an impredicativity in the phenomenology of time, a kind of radical transcendence.

In Livadas' opinion, the phenomenological constitution of time provides a model for the intuitive continuum and its impredicativity, a motive to reflect on its representation as a kind of essential "extension" within the realm of certain non-Cantorian mathematical theories that provide an alternative, phenomenologically oriented version of standard mathematics by negating conventional infinity and following the ever shifting horizon of our incorporating life-world (*Lebenswelt*) as is the case with Alternative Set Theory (AST) of the Prague School (Vopěnka 1979).

"We support in this paper that the adoption of ad hoc extension principles or "external" predicates in non-Cantorian theories with respect to vagueness or fuzziness (that is, uncountable infinity) reflects on a formal-axiomatical level the impredicativity of the transcendental ego of consciousness in its Husserlian sense meant as the constituting factor of the continuous unity of the flux of internal time. This is also the case with respect to the intuitionistic approach to continuum by a choice sequence modeling, where a strong extension principle is adopted for the elements of the universal spread C (Van Atten et al. 2002). It should be noted again that intuitionistically oriented H. Weyl had already developed in *Das Kontinuum* (1918), a view of the intuitive continuum based largely on the phenomenological description of the consciousness of internal time (Van Atten et al. 2002)".[35]

On the basis of these alternative approaches to continuum, the author lastly points out its inherent indescribability by means of a first-order formal language. "We hold that this indescribability manifests itself in the phenomenology of consciousness as the irreducibility of the continuous unity of the constituting flux of consciousness in-itself to the discrete mode of appearances of phenomena constituted as immanent unities in it".[36]

In the fourth part of the volume the analysis is centered on the link between epistemic complexity and causality. As we have just seen, the interface between causality, incomputability and meaningful complexity represent the secret thread of the third part. With respect to this, the words by Cooper were illuminating: the Turing invariant relations are key to pinning down how basic laws and entities emerge as mathematical constraints on causal structure. Hence the importance of a deep analysis of this very structure. What methods, however do we have to follow in order to understand the hidden aspects of natural causality? What is the role played by causality at the level of knowledge construction?

[35] Chapter 10, p 186.

[36] Chapter 10, p 186.

In Chapter 11, J. Nida-Rümelin remarks, first of all, that in order to present reasons against the possibility of naturalizing reasons it is necessary to present an account of naturalization. "Naturalism with respect to a domain D is the view that all entities or properties out of D can be naturalized. There are many different kinds of characterizing naturalism. The broadest account takes naturalism as being the view that nature is a coherent whole, and that human beings and all their properties are a part of nature. This account is not quite clear-cut and I would like to avoid answering the question "are you for or against naturalism?" understood in this sense..... There are many and competing accounts of what explanation in the natural sciences is but there exists an almost unanimous consensus that reference to *telē* cannot be a legitimate part of explanation in the natural sciences. Put differently: Teleological explanation is different from causal explanation and the natural sciences aim at causal explanations only. For example, take game theoretic models in evolutionary theory. Game theory has developed from the analysis of human agents. Utility and probability functions that one can attribute to these human agents constitute its conceptual frame. But the evolutionary story is exclusively causal. The talk of "selfish genes" (Dawkins)is merely metaphorical. The causal explanation contains no reference to intentions, aspirations, reasons, *telē*. Explanation in the natural sciences is causal – deterministic or probabilistic – it deduces *explananda* (natural events) from causes (antecedent natural events) together with natural laws. The *explananda* and the antecedent natural events do not contain intentional states and a fortiori do not contain reasons".[37]

According to the author "Naturalism" is the view that the methods of natural science suffice to describe and explain not only those events that are generally accepted as natural events in the sense of being adequate objects for scientific explanation, but also of events that are usually not objects of natural science. In this reading "Naturalism" is the meta-theoretical view that all events can in principle be explained by natural science. It is obvious that this meta-theoretical view makes sense only if it is based on a more general naturalistic world view regarding the ontological constitution of the entities and the range of the laws of Nature.

As Nida-Rümelin remarks, if Naturalism were to be true, epistemic reasons could be naturalized. On the contrary, the paper aims to introduce three reasons against the possibility of naturalizing epistemic reasons: the argument from normativity, the argument from objectivity, the argument from non-computability.

"The *non-standard view* I am arguing for, rejects this dichotomy between theoretical and practical reasons and it rejects the idea of desires as given, desires which cannot be criticised and modified. In giving up the idea of given desires we reject foundationalism regarding practical reasons. The non-standard view is coherentist. The practice of giving and taking reasons is not split into two separate parts with different rules of inference. A reason to act results in a belief that this act would be a good one..... Reasons speak for or against a propositional attitude. Some of these propositional attitudes have practical implications in the sense that a rational person having this propositional attitude acts accordingly.

[37] Chapter 11, p 203.

This description ..... is compatible with a close linkage between theory and practice, between propositional attitudes and actions. Propositional attitudes reveal themselves in acting. Preferences reveal themselves in choices. Wishes reveal themselves in motivations for action etc. A person may say that she believes that *p*, but if she acts as if *p* were not the case, we may doubt whether the person indeed has this belief. Reasons are epistemic. Reasons justify propositional attitudes. Propositional attitudes represent practices or, to put it more generally, whole *forms of live*".[38]

From a general point of view, all reasons can be transformed into epistemic ones. Moreover at least in an indirect and implicit way all reasons have some practical implications taken as a whole. In this sense, in author's opinion, it is not difficult to show that it is impossible to naturalize epistemic reasons.

Actually, life world reasoning is usually very complicated. The interplay of giving and taking reasons is essential for it. At this level epistemic reasoning is usually not algorithmic "epistemic reasoning cannot be identical with some causal-deterministic neurophysiological process, because causal deterministic processes in principle can be produced by Turing machines. This is obviously true for the classical deductive-nomological model of causal explanation, but it can be extended to more complex models of causal explanation including probabilistic ones. The validity of the argument from non-computability depends heavily on theories of causal relations. Whereas natural scientists stick to the classical model of causality as algorithmic, philosopher of science developed accounts of causality during the last decades that made causal relations part of epistemic reasoning..... But as far as causality is understood as a relation between natural, empirically accessible events, whereas this relation is lawful and this natural law allows producing the sequel of caused events by a Turing machine, non-computability is a strong argument against the possibility of naturalizing epistemic reasons".[39]

As is well known, simultaneously with the articulated investigation of the initial concept of information, since the 1960s and in the decades immediately afterwards, the concept of causality has come to revive through a fruitful and renewed link with the Theory of processes and the Theory of probability. In the 1980s, in particular (see W. Salmon's theoretical investigations), a new conception of causality emerged with success. It gives an important role to the notion of "invariance", insofar as it explicitly considers causal processes as instruments to propagate invariant structures. During the 1990s, new theories were hence put forward that directly connect causality with the procedures concerning the transmission and transformation of information. The link between invariance and information grounds also another fundamental contemporary approach to causation: the so-called "manipulative approach", which claims the content of causal assertions takes root in what we know, as cognitive agents, about how Reality can be modified and manipulated. The causal asymmetry which characterises such as conception of causality thus appears as a manifestation of the fact that causal notions originate in our experience as cognitive

[38] Chapter 11, p 205.

[39] Chapter 11, p 209.

agents. The "projectivism" which inspires this approach essentially appeals to the cognitive agent's capability to compress and manipulate information. Hence the very recovery of the crucial nexus between causality, compression, probability and scientific explanation, along the lines of what Chaitin originally postulated.

This recovery implies the necessity of an accurate and deep theoretical enquiry, and an innovative outlining of new instrument of investigation on a methodological level. Chapter 12 by R. Campaner and M.C. Galavotti aims to discuss a number of highly controversial issues, such as the problem of the relationship between causal models and intentional, goal-directed action, and the more general problem concerning the clarification of the interconnection between explanation, prediction and causation within a probabilistic framework. As the authors remark, at the beginning of the twentieth century the role of the category of causation in the building of scientific knowledge has been strongly challenged, mainly because of the progress of physics. Since the early 1970s, however, the notion of cause has been treated jointly with the notion of probability, and has been thus at the centre of a real revival. The chapter, in particular, intends to analyse, as we have just said, the theoretical bases of the two different conceptions that characterise the contemporary debate about the ultimate nature of causality. The first is represented by the mechanistic conception, which claims causal nexus, physical and objective, constitute a network which underlies phenomena and is responsible of their occurrence. This approach centres in the notions of causal process and causal mechanism, which are defined in different terms in the different mechanistic theories. The second successful approach taken into consideration is the manipulative approach, which traces causation back to our fundamental cognitive structures and to our capacity to manipulate reality: a "cause" is maintained to be something on which a free agent intentionally intervenes in order to obtain his target, the "effect". Therefore, causation is conceived of as a conceptual category we project on the world from our peculiar structure as cognitive agents. The authors want to consider the applicability of this conception to various scientific disciplines and to analyse its relationship with a general pragmatist perspective. Furthermore, they investigate the relationships between the mechanistic and the manipulative conception and their intersections: if mechanisms constitute the causal structure which underlies reality, their main interest for such a structure stems from the possibility to elaborate more and more effective manipulative strategies to control it. In the light of these considerations, in authors' opinion, among the "keys" to grasp causation the notions of stability and invariance seem to play a primary role. Actually, causal nexus, conceived of both in a mechanistic and in a manipulative sense, have to show a stable functioning, a behaviour that is invariant under intervention. In this sense, the notions of invariance and intervention allow to intertwine and integrate the main concepts the literature proposes as tools for the identification of the causal structures of Reality. The chapter exactly aims at deepen the features of such an intertwinement. In particular, revisiting Suppes' original suggestions, the authors underline the fact that the great American scientist "developed a pluralistic view of theories based on models, according to which theories are representable by means of a hierarchy of models characterized by different degrees of abstraction, which range

from empirical models, or "models of data" describing experimental evidence, to abstract mathematical models characterizing the theory".[40] The models linking a theory to empirical phenomena can be shown to preserve a certain structure under certain operations. In the authors' words: " 'the structure of a set of phenomena under certain empirical operations is the same as the structure of some set of numbers under arithmetical operations and relations' (Suppes 1967, p 59). Invariance, taken as the capacity to preserve structure, is therefore a pivotal feature of this view".[41] "Suppes does not impose particular requirements on causal chains, and claims that causality can be defined both in terms of random variables and events, without specifying in a univocal fashion what counts as an "event". Remarkably, no "ultimate genuine causes" are contemplated. By contrast, the notion of cause, genuine or spurious, is strictly linked to the specification of the set of concepts on which the set of events that can serve as causes in a given context is to be defined. In other words, both the notion of event and that of cause are linked to the specification of the set of concepts characterizing a given context".[42] Hence the necessity of a deep analysis of the Bayesian methods and a more precise definition of what counts as evidence.

In Chapter 13, J. Williamson precisely focuses on a particular kind of epistemic complexity, namely complexity of evidence. In particular it looks at the question of how complex evidence should impact on the strengths of an agent's beliefs.

As the author affirms: "It is a platitude to say that the strengths of our beliefs should depend on our available evidence, but it is notoriously hard to say exactly *how* evidence constrains appropriate degrees of belief. Bayesian epistemology begins to tackle this question, but typically considers only the simplest kinds of evidence, e.g., the case in which the evidence consists of a set of atomic propositions, or the case in which the evidence consists of a large database of good quality data. Reality, of course, is rarely if ever so simple. Evidence can be structured in a number of ways – causally, hierarchically, logically, for instance – and tends to be multifarious, a mixture of different kinds of structure from a mixture of different sources. In this paper I will show how *objective Bayesianism* – one particular version of Bayesian epistemology – can help shed light on the precise relation between complex evidence and belief".[43]

In particular, the author shows that evidence of empirical probability constrains degrees of belief in a rather straightforward way: the set of probability functions compatible with evidence is just the convex hull of the set of functions in which (according to the evidence) the empirical probability function lies. But evidence can contain information other than information about empirical probability, and the question arises as to what constraints $\mathcal{E}$ imposes on degrees of belief in such.

[40] Chapter 12, p 225.
[41] Chapter 12, p 225.
[42] Chapter 12, p 226.
[43] Chapter 13, p 231.

"Causality is an *influence relation* in the sense that learning just of new non-influences provides no grounds for changing degrees of belief. More precisely, if the language $\mathcal{L}$ is extended to $\mathcal{L}'$, which expresses a new proposition, and it is known that the corresponding variable is not a cause of any of the former variables, and other information in $\mathcal{E}$ does not indicate otherwise, then the agent's degrees of belief over the former language should not change: $P_{\mathcal{E}}^{\mathcal{L}'}(\theta) = P_{\mathcal{E}_{\mathcal{L}}}^{\mathcal{L}}(\theta)$ for each sentence θ of $\mathcal{L}$, where $\mathcal{E}_{\mathcal{L}}$ is the evidence in $\mathcal{E}$ that concerns $\mathcal{L}$. Hence causal evidence imposes equality constraints on degrees of belief".[44]

Causal structure provides one kind of evidential complexity, but according to the author there are others. For instance, hierarchical structure normally occurs in descriptions of mechanisms. In describing mechanisms in the human body we often need to talk simultaneously about processes that occur at the level of the body as a whole (e.g., the circulation of the blood), those at the level of the cell (e.g., oxygenation of haemoglobin), and those at the level of the genome (e.g., mutation of a single nucleotide of the β-globin gene). Hierarchical structure also occurs in describing causal relationships, because causal relations can themselves act as causes and effects. For example, smoking causing cancer causes governments to restrict tobacco advertising, which prevents smoking and thereby prevents cancer. This example shows that the same variable can occur at more than one level in the hierarchy.

In this sense, complexity of evidence is one kind of epistemic complexity. In his chapter the author aims to show how objective Bayesian epistemology can begin to take into consideration this kind of epistemic complexity. Objective Bayesianism offers, in his opinion, a unifying framework for integrating and interpreting not just evidence of empirical probability, but also evidence of causal, hierarchical and logical structure. Objective Bayesian probability can be defined over predicate languages as well as propositional languages, and the machinery of objective Bayesian nets can be used to represent and reason with objective Bayesian degrees of belief.

The fifth part of the volume is devoted to a variegated analysis concerning embodied cognition, the link between mind and brain, the role of creativity in cognitive activities and lastly the very possibility of doing metaphysics with robots. Actually, the extent to which the brain succeeds, albeit partially, in encapsulating the secret cipher of the cognitive abilities of other intelligent beings through a specific chain of programs determines the brain's capacity of grasping and reproducing these very abilities and prepares the possible successive irruption of new patterns of creativity.

In Chapter 14, W. Leinfellner remarks first of all that genetic algorithms demonstrate that a higher organism in its environment or society can modify its behavior (humans their societal decisions) by a selective and adaptive learning process which is regimented by ad-hoc game-theoretical and statistical societal default rules. These rules may change even genetically fixed rules; their use can generate new ones which our brain evaluates and the organism must store all of them in its memory system. Thus, evolutionary processing by learning, rule generation, and rules of innovations

[44] Chapter 13, p 235.

can totally describe the evolutionary and evolutive dynamics into play. "It is characteristic for mental evolutive processing after randomizations to progress gradually by using default rules, step by step, beyond the established knowledge. The use of default rules by humans can lead, as we will show, to mental innovations and the creation of entirely new solutions of conflicts between different mentifacts, sociofacts, artifacts, and technifacts".[45]

According to the author's view, the protosemantic function of the human brain, the representation of the external happenings of the world onto our brain's memory represents one of the fundamental pillars of human cognition. As is well known, P. Churchland rejects the traditional direct representations of the external world unto our language as a mere dogma of analytic "philosophy without brain" and calls it "sentence crunching". The protosemantics proposed by the author, on the contrary, may serve as the missing cognitive link which can fill the gap between the external world and its internal representation (mapping) onto our language. "From the societal, historical evolution of the human brain and from the most recent cognitive, brain-physiological, and linguistic research, we know that cognition, evaluations; memory storing, decision making, problem solving, and the realization of decisions and solutions of societal conflicts include a brain-based, evolutive, mental neuronal processing which involves the entire body as well (Damasio 1994; Basar 1988). The direct representation onto memory$_1$ presupposes a non-linguistic, brain-physiological, physical, cognitive protosemantics. There is no direct representation onto linguistic memory$_2$ (Churchland 1989)".[46]

In this sense, according to the author we have to go back to the physical grass roots of the cognitive and evaluative protosemantic functions of our neuronal brain. "Memory storing of happenings, of empirical causal networks begins in each case with the cognitive representation of the external, sensed, causal episodes, of the statistico-causal pairs of events .... and their statistico-causal concatenations in our memory system$_1$. These primitive, causally ordered tuples (basic causal pairs = CEP's) are represented and stored unconsciously into neuronal brain-wave patterns, they permit the recognition and afterwards the retrieval from memory$_1$ as internally sensitized episodes at our sense organs, without language. We become aware, but not fully conscious, of the neuronally stored and sensed images when the stored neuronal wave patterns, e.g., sound waves, are retrieved".[47]

In author's opinion, chance alone is the origin of every innovation, of all creation in the biosphere. This central concept of modern biology is no longer one among other conceivable hypotheses. It is today the sole conceivable hypothesis – the only one that squares with observed and tested facts. According to Leinfellner, nothing warrants the supposition or the hope that on this score our position is likely ever to be revised (Monod 1970).

But how does creativity function when societal conflicts have to be solved, for example by creating new culturefacts (mentifacts, sociofacts, artifacts, and

[45] Chapter 14, p 249.

[46] Chapter 14, p 251.

[47] Chapter 14, p 252.

technifacts)? The same holds for innovations, or partial creations and improvements, of culturefacts or methods. Here, like in all creative mental processes, mental randomizers and our simultaneous evaluations of the outcomes of mental lotteries play a leading role. They enable a new way of expected evaluations in case we don't know anything and have to search for a solution never used before; they also enable the realization of new solutions of social conflicts. According to Leinfellner (as well as to Penrose, Kauffman, Ruelle, Basar, and Freeman), internal neuronal randomizers are strange attractors, since they produce a vast number of expected and possible solutions, each of them with a certain value for us, in short: a lottery.

"These mental, neuronal randomizers are strange combinatorial or chaotic attractors (Ruelle 91, p 64; Kauffman 93, p 178); but only they can initiate the creation of new mentifacts in a way that is similar to, but more complex than, the biological creation of species. . .. There are no counterarguments to the explanation of self-organization as an evolutionary, and creativity as an evolutive, process; they differ just as to their empirical interpretation. Internal randomizers function often within immense populations, for example neurons, as Minsky has said. Here they are seen as the primordial, initial, and blind source, possible prestages of any mental creations".[48] In this sense, we have continuously to confront ourselves with chaos in order to construct by self-organisation our intellectual tools.

A. Corradini in Chapter 15 aims to show that emergentism in the philosophy of mind should be understood as a dualistic position. In order to achieve this goal she first of all revisits and analyses some of T. O'Connor's fundamental theses.

As is well known, in order to outline a strong ontological concept of emergence Timothy O'Connor characterizes emergent properties as "non structural" properties. In his opinion, an emergent property is defined as the property of a composite system that is wholly nonstructural, and emergentism is defined as the view according to which there are basic, non structural properties had by composite individuals. In this sense, we have to distinguish structural properties from the non structural ones.

"But, how to figure out the relationship between these two different sorts of properties? O'Connor complains that the relationship is often conceived as synchronic, static and formal, due to the contemporary tendency to assimilate emergentism to non-reductive physicalism and, as a consequence, emergence to the concept of synchronical supervenience. Rather, the relationship of micro-level structures and macro-level emergent properties should be viewed as dynamic and causal. In fact, the causal action of the underlying properties is needed to explain the occurrence of emergent properties at a given level of complexity. Yet, emergent properties have causal powers which are irreducible to those of the micro-level structure and which exert at their turn an influence on lower-level and/or same-level entities".[49]

According to Corradini, O'Connor's claims about the causal relationship between macro- and micro-level are the most critical aspects of his proposal. On

[48] Chapter 14, p 259.

[49] Chapter 15, p 269.

the one hand, he defends the typical emergentistic doctrine of the existence of a downward causation. On the other hand, however, O'Connor also maintains that emergent properties, as everything that occurs, depend on the causal dispositions of the fundamental physical properties. So, he emphasizes that an emergent system is not causally closed as regards its purely physical aspects and that emergent properties are thus not epiphenomenal. But, immediately after making this claim, he states that it is true in an emergentistic scenario that everything that occurs rests on the complete dispositional profile of the physical properties prior to the onset of emergent features. At the end of her analysis Corradini can lastly remark that an unambiguous reading of O'Connor's reasoning brings us to a more explicit form of dualism than that allowed by the author himself.

"Yet, however strong O'Connor's objection may be, it does not affect my own position. Substance dualism with ontological independence of the mind implies an impossible creatio ex nihilo only under the condition that the processes from which the mind emerges are merely material processes. Thus, this criticism can be countered if the development of the mental substance is traced back not only to material components, but also to a distinctive, non-material dimension of reality, endowed with ontological independence and existing from the very beginning of the emergent process. Such a dimension is the origin of the potentiality of development of the mental substance, which becomes actualized at the moment in which the biological structure reaches the necessary degree of complexity. Emergent dualism champions the idea of a co-evolution of mind and body, at the ontogenetic as well as at the phylogenetic level, on whose basis the realisation of non-biological potentialities is induced by the development of the biological structure, which, in its turn, is afterwards affected by the causal activity of the conscious mind (see on this Hasker 2008). Moreover, it is worth mentioning that the process of actualization of the mental substance also implies its particularization, its being the mind of a specific human individual. As we have just seen, the actualization of the mind is induced by a biological process of high complexity, but increasing complexity is also a sign of increasing individualization, so that my position does not face the problem of having to explain why a certain mental substance exerts its causal powers exclusively on its brain and not on somebody's else brain".[50]

D. Parisi in Chapter 16 remarks first of all that science and philosophy are both rational attempts at understanding reality but they are attempts of a different nature. A crucial difference is that scientific theories are supposed to generate specific predictions that match reality as we systematically, and possibly quantitatively, observe it with our naked senses or aided by instruments, whereas philosophical theories are normally supported only by arguments and are evaluated only through analysis and discussion. But he immediately underlines that the advent of the computer is likely to change this traditional conception. "Until now science has studied reality using two 'arms': the empirical observation of reality and the formulation of theories that try to explain what is observed. The computer makes it possible to

[50] Chapter 15, p 271.

use a third 'arm': the reproduction of reality in artefacts. The artefacts are simulations and robots or collections of robots. If an artefact behaves like some aspect or phenomenon of reality, we can claim that the principles we have followed in constructing the artefact are the same principles that govern that aspect or phenomenon of reality, and therefore we have understood that aspect or phenomenon of reality. Simulations and robots are a new way of expressing scientific theories. Traditionally, scientific theories are expressed either in words or using the symbols of mathematics. A computer simulation or the control system of a robot is a theory expressed as a computer program. This forces the researcher to formulate his or her theory in a precise and unambiguous way because, otherwise, the theory cannot be expressed in a computer program or in the control system of a functioning robot".[51]

According to Parisi, philosophers do metaphysics through conceptual analysis, reasoning, imagination, the proposition of ideas and theories, and discussion with colleagues. Their work, as always in philosophy, takes place entirely through the medium of language: all they do is speak and listen, write and read. The new cognitive and social scientists, on the contrary, will do metaphysics in a different way: by constructing robots. The metaphysics described by them will be the metaphysics of the robots that they will construct, reality as the robots know and understand it. "Robots are physical artefacts, whether they are simulated in a computer or physically realized, and this is very important because the knowledge that any organism has of reality depends on the organism's body, its external morphology of size and shape and its internal structure of organs and systems, and on the nature of its sensory and motor organs. A robot is a simulation of the body of an organism and of its sensory and motor organs. By constructing robots, and by comparing robots with different bodies and different sensory and motor organs, one can do "comparative metaphysics", trying to identify what general view of reality develops in each type of robot and comparing these different views..... since simulations and robots can be used to study not only real "reality" but also possible "reality", we can construct robots that do not resemble any animal that actually exists or has existed on Earth, or robots that live in an artificial environment which in different from the environment which exists on Earth, and determine what is their general view of reality. In other words, we can do not only "comparative metaphysics" but also "experimental metaphysics", determining how the metaphysics of an organism changes as we manipulate the organism's various properties".[52]

In this way, a new sort of evolution finally appears: by constructing robots we can more easily see how we put different objects in the same category not because they are similar from a sensory point of view but because we respond to them with the same action(s), how knowledge of where things are in space is knowledge on how to reach things with our eyes, hands, or feet, how counting is always counting only our actions, how time is counting our actions in time, etc. Unless we recognize the crucial role of our actions, and of the body that accomplishes these actions, in the definition of reality, we will describe an imaginary or superficial metaphysics.

[51] Chapter 16, p 276.

[52] Chapter 16, p 277.

"When we do metaphysics what we actually do is describe the particular adaptive pattern of a particular species of organisms, Homo sapiens. Doing metaphysics by constructing robots makes this entirely clear. A robotic metaphysics is a scientific metaphysics. It is metaphysics as done by science. And this is advantageous because it introduces a useful comparative approach that considers different species of organisms and different views of reality and because it creates a relativistic attitude towards our conception of reality. It shows us that what we call "metaphysics" is only one among many existing conceptions of reality, those possessed by other species of animals, while this is normally not recognized because the conception of reality that we try to describe when we do metaphysics is the conception of reality of the species that does the description".[53]

The view of reality possessed by the organism is entirely objective in that it is the only one that allows the organism to survive and reproduce and, therefore, it is "forced" on the organism, not chosen by the organism. Let us just remark that this sort of reality also depends (with respect first of all to its inner evolution) on the tools offered by the organisms in order to fix the path of their self-organisation.

In Chapter 17 A. Carsetti remarks first of all that, from an informational point of view, "the world which comes to "dance" at the level of the eyes of the mind is essentially impregnated with meaning. The "I" which perceives it realises itself as the fixed point of the interwoven "garland" with respect to the "capturing" of the thread inside the file and the genealogically-modulated articulation of the file itself which manages to express its invariance and become "vision" (visual thinking which is also able to inspect itself), anchoring its generativity at a deep semantic dimension. The model can shape itself as such and succeed in opening the eyes of the mind in proportion to its ability to permit the categorial to anchor itself to (and be filled by) intuition (which is not, however, static, but emerges as linked to a continuous process of metamorphosis). And it is exactly in relation to the adequate constitution of the channel that a sieve can effectively articulate itself and cogently realise its selective work at the informational level..... It is the (anchoring) rhythm-scanning of the labyrinth by the thread of meaning which allows for the opening of the eyes, and it is the truth, then, which determines and possesses them. Hence the construction of an "I" as a fixed point: the "I" of those eyes (an "I" which perceives and which exists in proportion to its ability to perceive according to the truth). What they see is a generativity in action, its surfacing rhythm being dictated intuitively. What this also produces, however, is a file that is incarnated in a body that posits itself as "my" body, or more precisely, as the body of "my" mind: hence the progressive outlining of a meaning, "my" meaning which is gradually pervaded by life".[54]

"Vision as emergence aims first of all to grasp (and "play") the paths and the modalities that determine the selective action, the modalities specifically relative to the revelation of the afore-mentioned semantic apparatus at the surface level according to different and successive phases of generality. These paths and modalities thus manage to "speak" through my own fibres. It is exactly through a

[53] Chapter 16, p 280.

[54] Chapter 17, p 284.

similar self-organising process, characterised by the presence of a double-selection mechanism, that the mind can partially manage to perceive (and assimilate) depth information in an objective way. The extent to which the network-model succeeds, albeit partially, in encapsulating the secret cipher of this articulation through a specific chain of programs determines the model's ability to see with the eyes of the mind as well as the successive irruption of new patterns of creativity. To assimilate and see, the system must first "think" internally (at the iterative level) the secret structures of the possible, and then posit itself as a channel (through the precise indication of forms of potential coagulum) for the process of opening and anchoring of depth information. This process then works itself gradually into the system's fibres, via possible selection, in accordance with the coagulum possibilities and the meaningful connections offered successively by the system itself".[55]

This "I" as incarnated, embodied mind, gradually becoming "occupied" by meaning while it articulates as life, ultimately reveals itself as the "I" of a body ("my body"), a body that articulates as an autonomous production of forms, the achieved extension of the meaning within the file, and as the world of virtual possibility in the guise and limits of necessity. "It acts as the "I" of a body-meaning which, in articulating as "my" body, can posit itself as the source of new creativity. In actual fact, it is this body, intended as an operant form-production allowing for the inscription of the file within itself, which finally articulates as a guide and support for the activity of ring-threading by conceptual schemata proper to the file itself, which determines the rising and the extended articulation of the neural connections at the level of the brain. This is the drawing which is ultimately donated: a drawing for the Other, however. The abstract frame in accordance with which the body progressively disincarnates itself, and which outlines the contours of cerebral connections, is related to the Other and is for the Other. While the body in which the mind is incarnated is my body, the brain through which the body is disincarnated (through simulation) is a brain which serves the intentionality of the Other, progressively inhabited by the meaning of the Other: indeed, it is the Other's brain in that I, as body, simulate it. Its constituting itself as autonomous unit marks and identifies my body-brain's constitution as an objective measuring device in the world and of the world".[56]

In conclusion of this short and incomplete presentation of the main guidelines of the book, let us now make just a few final remarks.

According to the suggestions presented by the authors in the different chapters (and in spite of the obvious difference in theoretical approaches), true cognition appears as constrained by the continuous reference to a number of specific analytical tools: computability and the Turing universe, incompressibility and the oracles in action, self-organising nets, deterministic chaos, non-linear mathematics, second-order structures and so on. With respect to this particular framework, the simulation activity, the construction, for instance, of an adequate semantics for natural language, presents itself as a form of interactive knowledge of the complex

[55] Chapter 17, p 284.

[56] Chapter 17, p 286.

chain of biological realisations through which Nature reveals itself to our brains in a consistent way (by means, for example, of the intelligent design of specific experiments at the level of the extended Turing universe). To simulate, in this sense, is not only a form of self-reflection or a kind of simple recovery performed by a complex cognitive net in order to represent itself at the surface level and "join" the government in action. The simulation work, in effect, offers the semantic net real instruments in order to perform a self-description process and to outline specific procedures of control as well as a possible map of an entire series of imagination paths. The progressive (and selective) exploration of these paths will allow, then, external information to canalise in an emergent way, and to exploit new and even more complex patterns of interactive expression and action. It is exactly the framing of this particular kind of laboratory of possible emergence that will assure the successive revelation of ever new portions of deep information: that particular "irruption" of the Other (the Source) which can express itself only within those particular fibres of the imagination and within that variant geometrical tissue of the forms which characterise, in an ultimate way, at the symbolic level, the cognitive activity of the subject. With respect to this frame of reference, we are no longer only faced with an observation activity that directly identifies itself as vision according to the truth but also with a simulation activity and a metamorphosis of meaning which express themselves by means of use and interaction, by the continuous surfacing of new forms of intentionality. When we pass from a world of objects to a world of constructions we are no longer exclusively faced, for instance, with boolean algebras, first-order structures and observational acts, we are really faced with a dynamic and functional universe characterised by inner circularity, by self-organisation and by the presence of specific categorisation processes as well as of evolutive differentiation patterns. Moreover, at the level of this particular world, as we have just said, the role played by meaning is different; meaning is now characterised in terms of a symbolic dynamics in action and with reference to a precise simulation language. As a consequence of this particular articulation, specific limitation facts can arise at the level of the progressive unfolding of this very language. New theoretical perspectives will reveal themselves with respect, in particular, to the inner self-organising aspects of the emerging structure and to the specific constitution of the individuals inhabiting this very structure considered as individuals essentially characterised not only in terms of their properties but also in terms of their relations (and their secret "affordances" at the symbolic level).

In a self-organising net the successive bifurcations, the recurrent delimitations imposed on the primitive predicate-inputs, actually appear as temporal and connected determinations of information fluxes. In this sense, such determinations (differently from Hintikka's appraisal of Kant's primitive intuitions), appear to concern not the (direct) successive presentation-construction of individuals, but the (previous) construction of patterns of constraints, of clusters of selective choices. Hence the essential link both with the traditional contemporary definitions of complexity at the propositional (and monadic) level, and with the revisitation of some Leibniz's (and Spinoza's) original intuitions as presented, for example, by Chaitin and other authors in their respective chapters.

In this sense, insofar as the aforesaid determinations of time articulate modulating, in a recurrent way, the action of the generators, the self-organizing nets present themselves progressively as frozen surface images of the originary informational Source and as a tool for the further construction-unfolding of its inner creativity, as a sort of arch and gridiron for the construction (and the recovery) of the "Other" through the constraints of an intended "sacrifice".

According to this frame of reference, the deep meaning appears first of all as relative to the action performed by precise semantic *fixed-points*, to a manifold, in particular, of subtended circumscription functions and to the progressive expression of specific postulates. The fixed-points of the resulting dynamics represent the "true" revelation of that specific tuning that characterises and identifies the predicates at work. Thus, at the monadic and polyadic level, we are obliged to outline a new, and specific kind of model: a self-organising (and coupled) structure not bound to sets and individuals, (with relative attributes) but to generators and fluxes of tuned information. In this new theoretical framework, the simple reference to possible worlds (as in Frege or Hintikka, for instance) in order to take into account the structure of intensionality is no longer sufficient, One has also to resort, in the first instance, to the dynamics of the constraints, to the identification of the indices and of the recurrent paths of the informational flow as well as of the role played by the observer, i.e. to the interplay existing between intervening and change.

Moreover, when we enter the polyadic realm and come to use, for instance, primitive binary relations, we must immediately make a series of choices (and assumptions) which are relative to the structural properties of such relations. Actually, in consequence of the structural properties that characterise, precisely, the dyadic predicates (i.e. which such predicates possess in a exclusively conceptual way), some specific conjunctions of these same predicates will be shown to be inconsistent. This means that what must be joined together will no longer consist of simple entities or sets of properties but of configurations and graphs. The conjunction, at the level of generators, should thus be realised respecting precise constraints of a "geometric" nature, connected, in particular, to the successive gain of configurations of "points-patches" which possess determined characteristics. The role of compatibility factors becomes particularly essential. From there both the birth of complex cancellation procedures and the introduction by construction of new individuals, in a potentially unlimited way, arise. Likewise, we would have, in a correlated way, the introduction of nested quantifiers. Thus, the role played by meaning really assumes a specific and deep relevance. As a matter of fact, at the level of this type of structure, we can individuate the existence of an essential plot between the successive "presentation" of the constraints and the action of the meaning postulates, on the one hand, and the articulated design of mutations, cancellations and contractions of the predicates-inputs that characterise the higher layers of formal constructions, on the other hand.

When, finally, we take into consideration the second-order structures and the general structures, things appear even more complex. As we have just said, what it is important to stress, in this particular case, is the fact that hidden in the structure some specific relations exist, some "rules" (second-order relations) that cannot be

defined as relations among individuals, but are utilised to define first-order relations (i.e., relations among individuals). It is, precisely, at the level of these tools that the action performed by meaning reveals all its subtleties.

Within the realm of general structures the original self-organising "glove" that imposes shape on itself acts contemporarily as a real support for the code inscription and, through the nesting process, for the complete unfolding of the limitation procedures: linkage operated by the *telos* allows an abstract design-frame to emerge, connected with an emergent nucleus of creativity through which other nuclei will manage to perceive and recognise themselves. What is presented, then, is a vision by principles, a process of concrete abstraction allowing for a new flame of invention which self-ignites. The file which inscribes itself as a code providing the support for the nesting process, permits a progressive and genealogical unification at the level of the activity of form-production. Hence a vision which can reflect itself as thought, and which can ultimately see by principles according to specific unification procedures. A new nucleus of individual creativity can emerge through which new postulates and axiomatic principles manage to find concrete self-expression: hence the unfolding of a production of forms which disincarnates itself in pure abstraction. In this sense, the embedding at work, in conjunction with the inscription, allows operative abstraction, and a meaning can finally to be embodied, a meaning which is able to posit itself as the source of new and pure vision by principles.

It is in the framework of this mysterious path, in itself already complicated enough, that we can individuate the progressive emergence, at the co-evolutive level, of the processes of rational perception proper to the human mind as well as of the categorisation processes that underlie the simulation language. It is with reference to this same framework that a precise dynamics of graphs will finally enter the stage with the subsequent introduction of cycles, attractors, fixed points etc. as well as the revelation of further constraints relative to problems of fitting, consistency etc. Precise forms of classification and therefore precise contexts of sense will appear. In this way, specific intensional structures will begin to emerge: in particular, intensional grammars defined with reference to orders and spaces of higher level. Thus, meaning can show its immense power at the selective level.

From here comes the necessity of outlining, in the case of dyadic structures (and, in general, of second-order structures), the sophisticated dynamism of a great book of Language that presents itself at the level of the conscious representation, like an effective reality in action. A reality which also emerges through our thinking and which, at the same time, determines, first of all at the genetic level, this same thinking. We no longer have before us a static book of Reality written in linguistic and mathematical characters. We have, on the contrary, a language in action which makes itself the Word of reality, the book in progress of linguistic constructions and which by reflecting the original pure generativity in a simulation space (of which, what is more, as human beings we are the support) assumes its primary forms and represents itself to itself by means of the tools of a precise symbolic dynamics. We are no longer faced, therefore, with concrete signs-symbols but rather with complex conceptual structures which are fitted into the effective articulation of a coupled process, a process into which, alongside the aforesaid dynamics relative to

configurations and graphs specific informational evaluations proper to the subject, to the structures of reflection and cognition that characterise his activity, will also be inserted.

We have seen how, for Putnam, the invention of new language represents the main tool to open up reality, to discover new horizons of meaning. The awareness that comes out from the intensional analysis of the semantic structures of natural language and of the cognitive functions that subtend these same structures, leads us to clearly understand that the problem is not only that of extracting the information living deeply within things. It is in addition that of building simulation models able to bring out the information contained in the fibres of reality in such a way that this same information irrupting into the neural circuits of elaboration proper to the subject can, finally, induce and determine the emergence of new forms of conceptual order and linguistic construction. The problem is, likewise, that of supplying coagulum functions which are capable of leading the Source to nest deeply, according to stronger and more powerful moduli. The emergence process and the same creativity that has been progressively realised, will present themselves as the "story" of the performed irruption and of the nesting carried out. They will articulate as forms of conceptual insight which spread out into a story, the story, in particular, of a biological realisation. In order to "open" reality, language must be embodied as an autonomous growth so that it will be possible, in perspective, to coagulate new linguistic constructions. Hence the importance of resorting to the outlining of recurrent processes and coupled processes in order to model the brain's functions. Likewise, this is the importance of that vertical (and intensional) dimension which grows upon itself, according to the exponential coefficients introduced and presented in the first part of the volume, and which appear indissolubly linked to the appearance and the definition of ever new forms of meaning. Forms which necessarily spring up through the successive discovery-construction of new *substrata* and of new dependency links according to Husserl's primitive intuitions.

Genealogical processes, recurrent processes, coupled structures, new measure spaces, new orders of acting imagination: such is the scenario within which the new information can, finally, emerge. This is Language in action. Here we may recognise the birth of new forms of seeing. Herein we can find the possibility to hear from a Source which comes forth to dictate from the interior of biological structures, like a new "*daimon*", the message of its self-representation, of its "wild" autonomy and of its renewed creativity. Cognitive activity, in this sense, is rooted in reality, but at the same time represents the necessary means whereby reality can embody itself in an objective way: i.e., in accordance with an in-depth nesting process and a continuous surface unfolding of operational causality.

Contents

Part I
Consciousness, Intentionality and Self-Organization

Chapter 1
The Link Between Brain Learning, Attention, and Consciousness

Stephen Grossberg

1.1 How Do We Continue to Learn Throughout Life?

We experience the world as a whole. Although myriad signals relentlessly bombard our senses, we somehow integrate them into unified moments of conscious experience that cohere together despite their diversity. Because of the apparent unity and coherence of our awareness, we can develop a sense of self that can gradually mature with our experiences of the world. This capacity lies at the heart of our ability to function as intelligent beings.

The apparent unity and coherence of our experiences is all the more remarkable when we consider several properties of how the brain copes with the environmental events that it processes. First and foremost, these events are highly context sensitive. When we look at a complex picture or scene as a whole, we can often recognize its objects and its meaning at a glance, as in the picture of a familiar face. However, if we process the face piece-by-piece, as through a small aperture, then its significance may be greatly degraded. To cope with this context sensitivity, the brain typically processes pictures and other sense data in parallel, as *patterns* of activation across a large number of feature-sensitive nerve cells, or neurons. The same is true for senses other than vision, such as audition. If the sound of the word GO is altered by clipping off the vowel O, then the consonant G may sound like a chirp, quite unlike its sound as part of GO.

During vision, all the signals from a scene typically reach the photosensitive retinas of the eyes at essentially the same time, so parallel processing of all the scene's parts begins at the retina itself. During audition, each successive sound reaches the ear at a later time. Before an entire pattern of sounds, such as the word GO, can be processed as a whole, it needs to be recoded, at a later processing stage, into a simultaneously available spatial pattern of activation. Such a processing stage is often called a working memory, and the activations that it stores are often

S. Grossberg (✉)
Director, Center for Adaptive Systems, Boston University, 677 Beacon Street, Boston, MA 02215
e-mail: steve@bu.edu

A. Carsetti (ed.), *Causality, Meaningful Complexity and Embodied Cognition*,
Theory and Decision Library A 46, DOI 10.1007/978-90-481-3529-5_1,

called short-term memory (STM) traces. For example, when you hear an unfamiliar telephone number, you can temporarily store it in working memory while you walk over to the telephone and dial the number.

In order to determine which of these patterns represents familiar events and which do not, the brain matches these patterns against stored representations of previous experiences that have been acquired through learning. Unlike the STM traces that are stored in a working memory, the learned experiences are stored in long-term memory (LTM) traces. One difference between STM and LTM traces concerns how they react to distractions. For example, if you are distracted by a loud noise before you dial a new telephone number, its STM representation can be rapidly reset so that you forget it. On the other hand, if you are distracted by a loud noise, you (hopefully) will not forget the LTM representation of your own name.

The problem of learning makes the unity of conscious experience particularly hard to understand, if only because we are able to rapidly learn such enormous amounts of new information, on our own, throughout life. For example, after seeing an exciting movie, we can tell our friends many details about it later on, even though the individual scenes flashed by very quickly. More generally, we can quickly learn about new environments, even if no one tells us how the rules of each environment differ. To a surprising degree, we can rapidly learn new facts without being forced to just as rapidly forget what we already know. As a result, we do not need to avoid going out into the world for fear that, in learning to recognize a new friend's face, we will suddenly forget our parents' faces.

Many contemporary learning algorithms would not be so lucky. Speaking technically, the brain solves a very hard problem that many current approaches to technology have not solved. It is a self-organizing system that is capable of rapid yet stable autonomous learning of huge amounts of data in a nonstationary environment. Discovering the brain's solution to this key problem is as important for understanding ourselves as it is for developing new pattern recognition and prediction applications in technology.

I have called the problem whereby the brain learns quickly and stably without catastrophically forgetting its past knowledge the *stability–plasticity dilemma*. The stability–plasticity dilemma must be solved by every brain system that needs to rapidly and adaptively respond to the flood of signals that subserves even the most ordinary experiences. If the brain's design is parsimonious, then we should expect to find similar design principles operating in all the brain systems that can stably learn an accumulating knowledge base in response to changing conditions throughout life. The discovery of such principles should clarify how the brain unifies diverse sources of information into coherent moments of conscious experience.

This article reviews evidence that the brain does operate in this way. It summarizes several recent brain modeling studies that illustrate, and further develop, a theory called Adaptive Resonance Theory, or ART, that I introduced in 1976 (Grossberg 1976a,b, 1978, 1980, 1982). In the present article, I briefly summarize results selected from four areas where ART principles have been used to explain challenging behavioral and brain data. These areas are visual perception, visual object recognition, auditory source identification, and variable-rate

speech recognition. On first inspection, the behavioral properties of these visual and auditory phenomena may seem to be entirely unrelated. On a deeper computational level, their governing neural circuits are proposed to incorporate a similar set of computational principles.

I should also say right away, however, that ART principles do not seem to be used in all brain learning systems. Whereas ART learning designs help to explain sensory and cognitive processes such as perception, recognition, attention, reinforcement, recall, working memory, and memory search, other types of learning seem to govern spatial and motor processes. In these latter task domains, it is adaptive to forget old coordinate transformations as the brain's control systems adjust to a growing body and to other changes in the body's sensory–motor endowment throughout life.

Sensory and cognitive processes are often associated with the What cortical processing stream that passes from the visual cortex through the inferotemporal cortex, whereas spatial and motor processes are associated with the Where (or How) cortical processing stream that passes from the visual cortex through the parietal cortex (Goodale and Milner 1992; Mishkin et al. 1983; Ungerleider and Mishkin 1982). Our research over the years has concluded that many processes in the two distinct streams, notably their matching and learning processes, obey different, and even complementary, laws. This fact bears heavily on questions of consciousness and helps to explain why procedural memories are not conscious (Cohen and Squire 1980; Mishkin 1982; Scoville and Milner 1957; Squire and Cohen 1984). Indeed, a central hypothesis of ART since its inception is:

ART Hypothesis: All Conscious States Are Resonant States

As noted in greater detail below, many spatial and motor processes involve a form of inhibitory matching and mismatchbased learning that does not support resonant states. Hence, by the ART Hypothesis, they cannot support a conscious state. Although ART predicts that all conscious states are resonant states, the converse statement, that all resonant states are conscious states, is not asserted.

It might be worthwhile to note immediately that various other models of cognitive learning and recognition, such as the popular backpropagation model (Parker 1982; Rumelhart et al. 1986; Werbos 1974), are based on a form of mismatch-based learning. They cannot, therefore, generate resonant states and, in fact, are well known to experience catastrophic forgetting under real-time learning conditions. A comparative survey of ART vs backpropagation computational properties is provided in Grossberg (1988).

1.2 The Theoretical Method

Another point worth noting is how one arrives at a psychophysiological theory such as ART which attempts to link behavioral properties to the brain mechanisms which generate them. Such a linkage between brain and behavior is, I believe, crucial in any mature theory of consciousness, since a theory of consciousness that cannot explain behavioral data has failed to deal with the contents of consciousness, and

a theory of consciousness that cannot link behaviors to the brain mechanisms from which they emerge must remain, at best, a metaphor.

A particular type of theoretical method has been elaborated over the past 40 years with which to approach such complex behavioral and brain phenomena. The key is to begin with behavioral data, typically scores or even hundreds of parametrically structured behavioral experiments in a particular problem domain. One begins with behavioral data because the brain has evolved in order to achieve behavioral success. Any theory that hopes to link brain to behavior thus needs to discover the computational level on which brain dynamics control behavioral success. One works with large amounts of data because otherwise too many seemingly plausible hypotheses cannot be ruled out. A crucial metatheoretical constraint is to insist upon understanding the behavioral data – which comes to us as static numbers or curves on a page – as the emergent properties of a dynamical process which is taking place moment-by-moment in an individual mind. One also needs to respect the fact that our minds can adapt on their own to changing environmental conditions without being told that these conditions have changed. One thus needs to frontally attack the problem of how an intelligent being *can autonomously adapt to a changing world*. Knowing how to do this is presently an art form. There are no known algorithms with which to point the way.

Whenever we have attempted this task in the past, we have resisted every temptation to use homunculi or else the crucial constraint on *autonomous* adaptation would be violated. The result has regularly been the discovery of new organizational principles and mechanisms, which we have then realized as a minimal model operating according to only locally defined laws that are capable of operating on their own in real time. The remarkable fact is that, when such a model has been written down, it has always been interpretable as a neural network. These neural networks have always included known brain mechanisms. The functional interpretation of these mechanisms has, however, often been novel because of the light thrown upon them by the behavioral analysis. The networks have also typically predicted the existence of unknown neural mechanisms, and many of these predictions have been supported by subsequent neurophysiological, anatomical, and even biochemical experiments over the years.

Once this neural connection has been established by a top-down analysis, one can work both top-down from behavior and bottom-up from brain to exert a tremendous amount of conceptual pressure with which to better characterize and refine the model. A fundamental empirical conclusion can be drawn from many experiences of this type; namely, the brain as we know it can be successfully understood as an organ that is designed to achieve successful autonomous adaptation to a changing world. I like to say that, although I am known as one of the founders of the field of neural networks, I have never tried to derive a neural network. They are there because they provide a natural computational framework with which to control autonomous behavioral adaptation to a changing world.

Such a real-time analysis is not easy because it requires that one have knowledge, and even mastery, of several disciplines. For example, it has always proved to be

the case that the level of brain organization that computes behavioral success is the network or system level. Does this mean that individual nerve cells, or even smaller components, are unimportant? Not at all. One needs to properly define the individual nerve cells and their interactions in order to correctly define the networks and systems whose interactive, or emergent, properties map onto behavior as we know it. Thus one must be able to freely move between (at least) the three levels of Neuron, Network, and Behavior in order to complete such a theoretical cycle.

Doing this requires that one has a sufficiently powerful theoretical language. The language of mathematics has proved to be the relevant tool, indeed a particular kind of mathematics. All of the self-adapting behavioral and brain systems that I have ever derived are nonlinear feedback systems with large numbers of components operating over multiple spatial and temporal scales. The nonlinearity just means that our minds are not the sum of their parts. The feedback means that interactions occur in both directions within the brain and between the brain and its environment. The multiple temporal scales are there because, for example, processes like STM are faster than the processes of learning and LTM. Multiple spatial scales are there because the brain needs to process parts as well as wholes. All of this is very easy to say intuitively. But when one needs to work within the tough honesty of mathematics, things are not so easy. Most of the difficulties that people seem to have in understanding what is already theoretically known about such systems derives from a literacy problem in which at least one, but often more than one, of the ingredients of neuron, network, behavior, and nonlinear feedback mathematics are not familiar to them.

A second important metatheoretical constraint derives from the fact that no single step of theoretical derivation can derive a whole brain. One needs to have a method that can evolve with the complexity of the environmental challenges that the model is forced to face. This is accomplished as follows. After introducing a dynamic model of a prescribed set of data, one analyzes its behavioral and brain data implications as well as its formal properties. The cycle between intuitive derivation and computational analysis goes on until one finds the most parsimonious and most predictive realization of the organizational principles that one has already discovered. Through this analysis, one can also identify various "species-specific variations" of such a prototypical model and apply them to different types of data. Such a theoretical analysis also discloses the *shape* of the boundary, within the space of data, beyond which the model no longer has explanatory power. The shape of this boundary between the known and the unknown then often clarifies what design principles have been omitted from the previous analyses. The next step is to show how these additional design principles can be incorporated into a more powerful model that can explain even more behavioral and neural data. In this way, the model undergoes a type of evolutionary development, as it tries to cope behaviorally with environmental constraints of ever increasing subtlety and complexity.

The metatheoretical constraint that comes into view here is an *embedding* constraint; in other words, one needs to be able to embed the previous model into the new model. Otherwise expressed, the previous model needs to be "unlumpable" as it evolves into an increasingly complex "brain." This is a type of *correspondence principle* that places a surprisingly severe test on the adequacy of the previ-

ously discovered theoretical principles. Many models regularly fail the embedding constraint. That is why they come and go with surprisingly rapidity and do not get integrated into burgeoning theories of ever greater predictive power.

The crucial importance of being able to derive behavioral mechanisms as emergent properties of real-time brain mechanisms, and being able to embed a previous model into a more mature model that is capable of adapting to more complex environments, led me to the name Embedding Fields for my earliest models of brain and behavior (Grossberg 1964). The word "fields" is a short-hand for the neural network as a computational unit whose interactions generate behavioral emergent properties; the word "embedding" refers to the unlumpability constraint. Many stages of model evolution have occurred since the mid-1960s and all of them have successfully built a foundation for their progeny. The present article will necessarily omit these modeling cycles and will instead discuss some of its results from the viewpoint of consciousness research.

1.3 How Do We Perceive Illusory Contours and Brightness?

Let me start by providing several examples of the diverse phenomena that ART clarifies. Consider the images in Fig. 1.1. Figure 1.1a shows an image called an Ehrenstein figure in which some radial black lines are drawn on a uniformly white paper. Remarkably, our minds construct a circular illusory contour that touches each line end at a perpendicular orientation. This illusory contour is a collective, emergent property of all the lines that only occurs when their positions relative to each other are suitable. For example, no illusory contour forms at the line ends in Fig. 1.1b even though they end at the same positions as the lines. Note also that the illusory contour in Fig. 1.1a surrounds a disk that seems uniformly brighter than its surround. Where does the brightness enhancement come from? It certainly does not always happen when illusory contours form, as can be seen by inspecting Fig. 1.1c. Here a vertical illusory contour can be recognized as interpolating the two sets of offset horizontal lines, even though neither side of the contour seems brighter

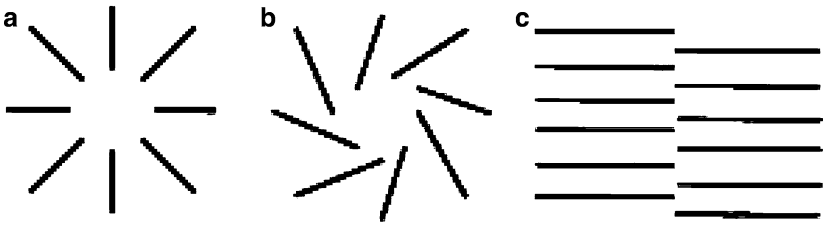


Fig. 1.1 (**a**) The Ehrenstein pattern generates a circular illusory contour that encloses a circular disk of enhanced illusory brightness. (**b**) If the endpoints of the Ehrenstein pattern remain fixed while their orientations are tilted, then both the illusory contour and the brightness vanish. (**c**) The offset pattern generates a vertical boundary that can be recognized even though it cannot be seen

than the other. How we can consciously *recognize* something that we cannot see and is thus perceptually invisible is a fascinating aspect of our conscious awareness about which quite a bit is now known. Such percepts are known as *amodal* percepts (Michotte et al. 1964) in order to distinguish them from modal, or visible, percepts. Amodal percepts are experienced in response to many naturalistic scenes, notably in response to scenes in which some objects are partially occluded by other objects. How both modal and amodal percepts can occur will be discussed below. Of particular interest from the viewpoint of ART processing is why the Ehrenstein disk looks bright, despite the fact that there are no local contrasts within the image itself that describe a disk-like object.

1.4 How Do We Learn to Recognize Visually Perceived Objects?

The Ehrenstein example concerns the process of visual perception. The next example concerns a process that goes on at a higher level of the visual system. It is the process whereby we visually recognize objects. A key part of this process concerns how we learn to categorize specific instances of an object, or set of objects, into a more general concept. For example, how do we learn that many different printed or script letter fonts can all represent the same letter A? Or how do we learn that several different combinations of patient symptoms are all due to the same disease? Moreover, how do we control how general our categories will become? For some purposes, like recognizing a particular face, we need highly specific categories. For others, like knowing that every person has a face, the categories are much more general. Finally, how does our learning and memory break down when something goes wrong in our brain? For example, it is known that lesions to the human hippocampal system can cause a form of amnesia whereby, among other properties, patients find it very hard to learn new information and hard to remember recently learned information, but previously learned information about which their memory has "consolidated" can readily be retrieved. Thus, an amnesic patient can typically carry out a perfectly intelligent conversation about experiences that occurred a significant time before the lesion that caused the amnesia occurred.

What computational properties do the phenomena of bright illusory disks and amnesic memory have in common? I will suggest below that their apparent differences conceal the workings of a general unifying principle.

1.5 How Do We Solve the Cocktail Party Problem?

To continue with our list, let us now consider a different modality entirely; namely, audition. When we talk to a friend in a crowded noisy room, we can usually keep track of our conversation above the hubbub, even though the sounds emitted by the friendly voice may be substantially overlapped by the sounds emitted by

other speakers. How do we separate this jumbled mixture of sounds into distinct voices? This is often called the cocktail party problem. The same problem is solved whenever we listen to a symphony or other music wherein overlapping harmonic components are emitted by several instruments. If we could not separate the instruments or voices into distinct sources, or auditory streams, then we could not hear the music as music or intelligently recognize a speaker's sounds. A striking and ubiquitous property of such percepts, and one which has not yet been understood by alternative modeling approaches, is how future events can alter our conscious percepts of past events in a context-sensitive manner.

A simple version of this competence is illustrated by the auditory continuity illusion (Bregman 1990). Suppose that a steady tone shuts off just as a broadband noise turns on. Suppose, moreover, that the noise shuts off just as the tone turns on once again; see Fig. 1.2a. When this happens under appropriate conditions, the tone seems to continue right through the noise, which seems to occur in a separate auditory "stream." This example shows that the auditory system can actively extract those components of the noise that are consistent with the tone and use them to track the "voice" of the tone right through the hubbub of the noise.

In order to appreciate how remarkable this property is, let us compare it with what happens when the tone does not turn on again for a second time, as in Fig. 1.2b. Then the first tone does not seem to continue through the noise. It is perceived to stop before the noise. How does the brain know that the second tone will turn on after the noise shuts off so that it can continue the tone through the noise, yet not continue the tone through the noise if the second tone does not eventually occur? Does this not seem to require that the brain can operate "backward in time" to alter its decision as to whether to continue a past tone through the noise based on future events?

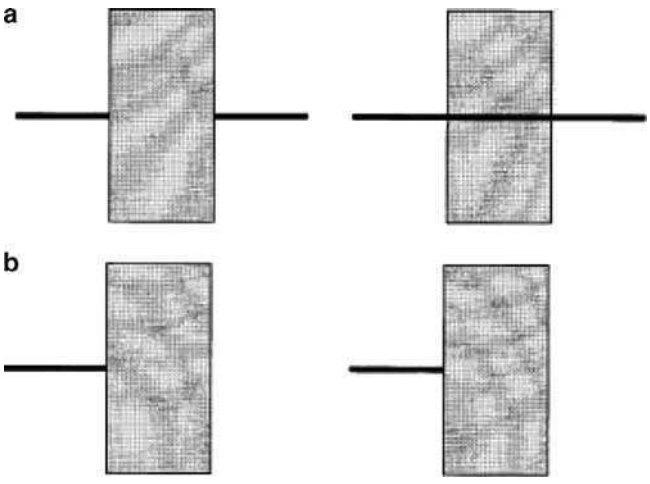


Fig. 1.2 (**a**) Auditory continuity illusion: When a steady tone occurs both before and after a burst of noise, then under appropriate temporal and amplitude conditions, the tone is perceived to continue through the noise. (**b**) This does not occur if the noise is not followed by a tone

Many philosophers and scientists have puzzled about this sort of problem. I argue that the process whereby we consciously hear the first tone takes some time to unfold so that, by the time we hear it, the second tone has already begun. To make this argument, we need to ask why does conscious audition take so long to occur after the actual sound energy reaches our brain? Just as important, why can the second tone influence the conscious percept so quickly, given that the first tone could not? Finally, I indicate what these auditory phenomena have to do with bright Ehrenstein disks and amnesia.

1.6 How Do We Consciously Perceive Speech?

The final examples also involve the auditory system, but at a higher level of processing. They concern how we understand speech. In these examples, too, the process whereby conscious awareness occurs takes a long time, on the order of 100 ms or more. An analysis of these percepts will also give us more clues about the nature of the underlying process. The first example is called phonemic restoration. Suppose that a listener hears a noise followed immediately by the words "eel is on the . . ." If this string of words is followed by the word "orange," then "noise-eel" sounds like "peel." If the word "wagon" completes the sentence, then "noise-eel" sounds like "wheel." If the final word is "shoe," then "noise-eel" sounds like "heel." This marvelous example, which was developed by Richard Warren and his colleagues more than 20 years ago (Warren 1984; Warren and Sherman 1974), vividly shows that the bottom-up occurrence of the noise is not sufficient for us to hear it. Somehow the sound that we expect to hear based upon our previous language experiences influences what we do hear, at least if the sentence is said quickly enough. As in the auditory continuity illusion, it would appear that the brain is working "backward in time" to allow the meaning imparted by a later word to alter the sounds that we consciously perceive in an earlier word.

I suggest that this happens because, as the individual words occur, they are stored temporarily via STM traces in a working memory. As the words are stored, they activate LTM traces which attempt to categorize the stored sound stream into familiar language units like words at a higher processing level. These list categories, in turn, activate learned top-down expectations that are *matched* against the contents of working memory to verify that the information expected from previous learning experiences is really there. This concept of bottom-up activation of learned categories by a working memory, followed by readout of learned top-down expectations, is illustrated in Fig. 1.3a.

What is the nature of this matching, or verification, process? Its properties have been clarified by experiments of Arthur Samuel (Samuel 1981a,b) and others in which the spectral content of the noise was varied. If the noise includes all the formants of the expected sound, then that is what the subject hears, and other spectral components of the noise are suppressed. If some formants of the expected sound are missing from the noise, then only a partial reconstruction is heard. If silence

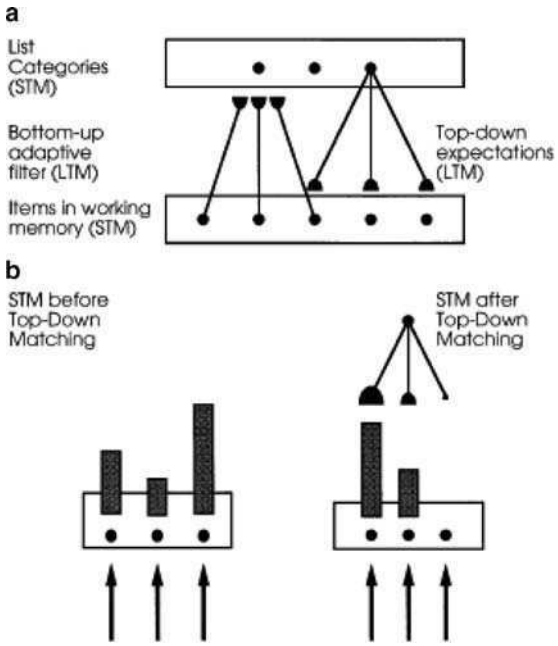


Fig. 1.3 (**a**) Auditory items activate STM traces in a working memory, which send bottom-up signals toward a level at which list categories, or chunks, are activated in STM. These bottom-up signals are multiplied by learned LTM traces which influence the selection of the list categories that are stored in STM. The list categories, in turn, activate LTM-modulated top-down expectation signals that are matched against the active STM pattern in working memory. (**b**) This matching process confirms and amplifies STM activations that are supported by contiguous LTM traces and suppresses those that are not

replaces the noise, then only silence is heard. The matching process thus cannot "create something out of nothing." It can, however, selectively amplify the expected features in the bottom-up signal and suppress the rest, as in Fig. 1.3b.

The process whereby the top-down expectation selectively amplifies some features while suppressing others helps to "focus attention" upon information that matches our momentary expectations. This focusing process helps to filter out the flood of sensory signals that would otherwise overwhelm us and to prevent them from destabilizing our previously learned memories. Learned top-down expectations hereby help to solve the stability–plasticity dilemma by focusing attention and preventing spurious signals from accidentally eroding our previously learned memories. In fact, Gail Carpenter and I proved mathematically in 1987 that such an ART matching rule assures stable learning of an ART model in response to rapidly changing environments wherein learning becomes unstable if the matching rule is removed (Carpenter and Grossberg 1987a).

What does all this have to do with our conscious percepts of speech? This can be seen by asking: If top-down expectations can select consistent bottom-up signals, then what keeps the selected bottom-up signals from reactivating their top-down

expectations in a continuing cycle of bottom-up and top-down feedback? Nothing does. In fact, this reciprocal feedback process takes awhile to equilibrate, and when it does, the bottom-up and top-down signals lock the STM activity patterns of the interacting levels into a resonant state that lasts much longer and is more energetic than any individual activation. ART hereby suggests how only resonant states of the brain can achieve consciousness and that the time needed for a bottom-up/top-down resonance to develop helps to explain why a conscious percept of an event takes so long to occur after its bottom-up input is delivered.

The example of phonemic restoration also clarifies another key point about the conscious perception of speech. If noise precedes "eel is on the shoe," we hear and understand the meaning of the sentence "heel is on the shoe." If, however, noise is replaced by silence, we hear and understand the meaning of the sentence "eel is on the shoe" which has a quite different, and rather disgusting, meaning. This example shows that the process of resonance binds together information about both meaning and phonetics. Meaning is not some higher-order process that is processed independently from the process of conscious phonetic hearing. Meaning and phonetics are bound together via resonant feedback into a global emergent state in which the phonetics that we hear are linked to the meaning that we understand.

1.7 ART Matching and Resonance: the Link Between Attention, Intention, and Consciousness

Adaptive resonance theory claims that, in order to solve the stability–plasticity dilemma, only resonant states can drive new fast learning. That is why the theory is called *adaptive* resonance theory. I explain how this works more completely below. Before doing so, let me emphasize some implications of the previous discussion that are worth reflecting about. The first implication provides a novel answer as to why, as philosophers have asked for many years, humans are "intentional" beings who are always anticipating or planning their next behaviors and their expected consequences. ART suggests that "stability implies intentionality." That is, stable learning requires that we have expectations about the world that are continually matched against world data. Otherwise expressed, without stable learning, we could learn very little about the world. Having an active top-down matching mechanism greatly amplifies the amount of information that we can stably learn about the world. Thus the mechanisms which enable us to know a changing external world, through the use of learned expectations, set the stage for achieving internal self-awareness.

It should be noted here that the word "intentionality" is being used, at once, in two different senses. One sense concerns the role of expectations in the anticipation of events that may or may not occur. The second sense concerns the ability of expectations to read-out planned sequences of behaviors aimed at achieving definite behavioral goals. The former sense will be emphasized first; the latter toward the end of the article. My main point in lumping them together is that ART provides a unified mechanistic perspective with which to understand both uses of the word.

The second implication is that "intention implies attention and consciousness." That is, expectations start to focus attention on data worthy of learning, and these attentional foci are confirmed when the system as a whole incorporates them into resonant states that include (I claim) conscious states of mind. Implicit in the concept of intentionality is the idea that we can get ready to experience an expected event so that, when it finally occurs, we can react to it more quickly and vigorously, and until it occurs, we are able to ignore other, less desired, events. This property is called priming. It implies that, when a top-down expectation is read out in the absence of a bottom-up input, it can subliminally sensitize the cells that would ordinarily respond to the bottom-up input, but not actually fire them, while it suppresses cells whose activity is not expected. Correspondingly, the ART matching rule computationally realizes the following properties at any processing level where bottom-up and top-down signals are matched: (1) bottom-up automatic activation: A cell, or node, can become active enough to generate output signals if it receives a large enough bottom-up input, other things being equal; (2) top-down priming: A cell can become sensitized, or subliminally active, and thus cannot generate output signals if it receives only a large top-down expectation input. Such a top-down priming signal prepares a cell to react more quickly and vigorously to subsequent bottom-up input that matches the top-down prime; (3) match: A cell can become active if it receives large convergent bottom-up and top-down inputs. Such a matching process can generate enhanced activation as resonance takes hold; (4) mismatch: A cell is suppressed even if it receives a large bottom-up input if it also receives only a small, or zero, top-down expectation input.

I claim that this ART matching rule and the resonance rule that it implies operate in all the examples that I have previously sketched and do so to solve the stability–plasticity dilemma. All the examples are proposed to illustrate how we can continue to learn rapidly and stably about new experiences throughout life by matching bottom-up signal patterns from more peripheral to more central brain processing stages against top-down signal patterns from more central to more peripheral processing stages. These top-down signals represent the brain's learned expectations of what the bottom-up signal patterns should be based upon past experience. The matching process is designed to reinforce and amplify those combinations of features in the bottom-up pattern that are consistent with the top-down expectations and to suppress those features that are inconsistent. This top-down matching step initiates the process whereby the brain selectively pays attention to experiences that it expects, and binds them into coherent internal representations through resonant knowledge about the world.

Given that such a resonant matching process occurs in the brain, how does the brain react when there is a mismatch situation? The ART matching rule suggests that a big enough mismatch between a bottom-up input and a top-down expectation can rapidly attenuate activity at the matching level. This collapse of bottom-up activation can initiate a rapid reset of activity at both the matching level itself and at the subsequent levels that it feeds, thereby initiating a memory search for a more appropriate recognition category or creating a new one.

1.8 Resonant Dynamics During Speech Categorization

Many examples of such a reset event occur during variable-rate speech perception. As one example, consider how people hear combinations of vowels (V) and consonants (C) in VC–CV sequences. Bruno Repp at Haskins Laboratories has studied perception of the sequences [ib]–[ga] and [ib]–[ba] when the silence interval between the initial VC syllable and the terminal CV syllable is varied (Repp 1980). If the silence interval is short enough, then [ib]–[ga] sounds like [iga] and [ib]–[ba] sounds like [iba]. Repp ran a number of conditions, leading to the several data curves displayed in Fig. 1.4. The main point for present purposes is that the transition from a percept of [iba] to one of [ib]–[ba] occurs after 100–150 ms more silence than the transition from [iga] to [ib]–[ga]. One hundred milliseconds is a very long time relative to the time scale at which individual neurons can be activated. Why is this shift so large?

My colleagues Ian Boardman and Michael Cohen and I have quantitatively simulated these data using a model, called the ARTPHONE model, of how a resonant

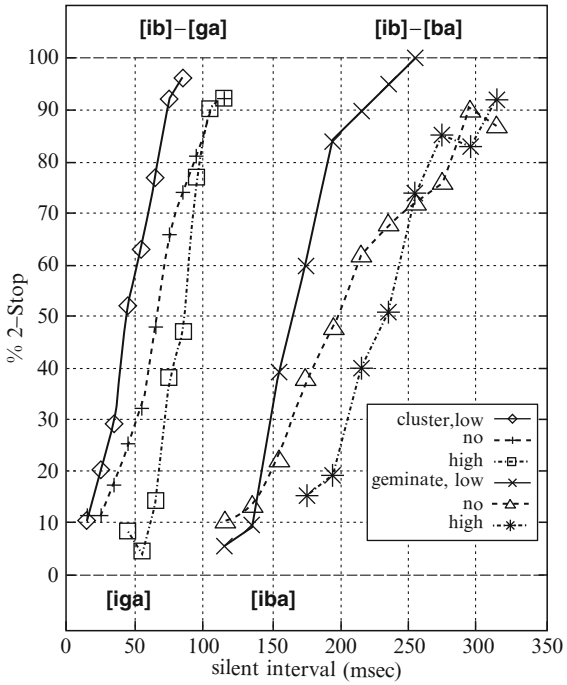


Fig. 1.4 The *left-hand curves* represent the probability, under several experimental conditions, that the subject will hear [ib]–[ga] rather than [iga]. The *right-hand curves* do the same for [ib]–[ba] rather than the fused percept [iba]. Note that the perception of [iba] can occur at a silence interval between [ib] and [ba] that is up to 150 ms longer than the one that leads to the percept [iga] instead of [ib]–[ga] (data are reprinted with permission from Repp BH (1980) Haskins Laboratories Status Report on Speech Research, SR-61, 151–165)

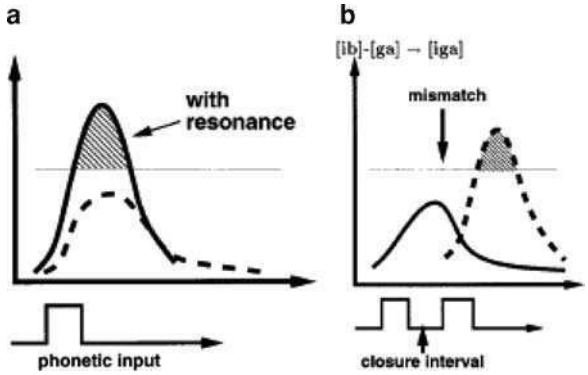


Fig. 1.5 (**a**) Response to a single stop, such as [b] or [g], with and without resonance. Suprathreshold activation is *shaded*. (**b**) Reset due to phonologic mismatch between [ib] and [ga]

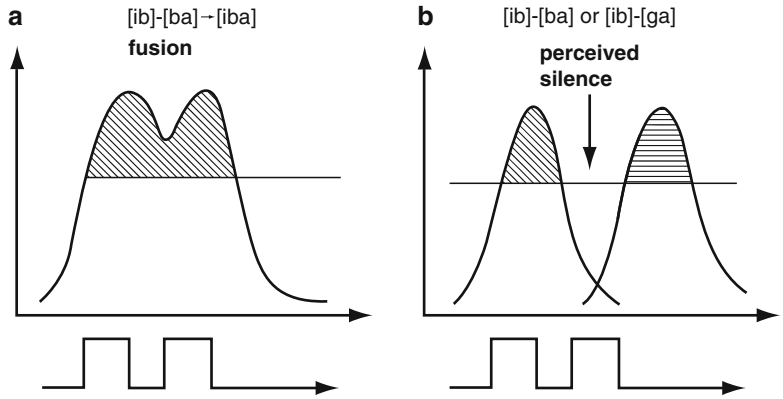


Fig. 1.6 (**a**) Fusion in response to proximal similar phones. (**b**) Perceptual silence allows a two stop percept

wave develops due to bottom-up and top-down signal exchanges between a working memory that represents the individual speech items and a list categorization network that groups them together into learned language units, or chunks (Grossberg et al. 1997a). We have shown how a mismatch between [g] and [b] rapidly resets the working memory if the silence between them is short enough, thereby preventing the [b] sound from reaching resonance and consciousness, as in Fig. 1.5. We have also shown how the development of a previous resonance involving [b] can resonantly fuse with a subsequent [b] sound to greatly extend the perceived duration of [iba] across a silence interval between [ib] and [ba]. Figure 1.6a illustrates this property by suggesting how the second presentation of [b] can quickly reactivate the resonance in response to the first presentation of [b] before the resonance stops.

This phenomenon uses the property that it takes longer for the first presentation of [b] to reach resonance than it does for the second presentation of [b] to influence the maintenance of this resonance.

If, however, [ib] can fuse across time with [ba], then how do we ever hear distinct [ib]–[ba] sounds when the silence gets long enough? Much evidence suggests that after a resonance fully develops, it spontaneously collapses after awhile due to a habituative process that goes on in the pathways that maintain the resonance via bottom-up and top-down signals. Thus, if the silence is long enough for resonant collapse of [ib] to occur, then a distinguishable [ba] resonance can subsequently develop and be heard, as in Fig. 1.6b.

Such a habituative process has also been used to explain many other data about perception, learning, and recognition, notably data about the reset of visual, cognitive, or motor representations in response to rapidly changing events. Relevant visual data include properties of light adaptation, visual persistence, aftereffects, residual traces, and apparent motion (Carpenter and Grossberg 1981; Francis and Grossberg 1996a,b; Francis et al. 1994). Abbott et al. (1997) have recently reported data from the visual cortex that they modeled using the same habituative law that was used in all of these applications. At bottom, such a habituative law is predicted to be found so ubiquitously across brain systems because it helps to rapidly adapt, reset, and rebalance neural circuits in response to rapidly changing input conditions, notably as part of an opponent process (Grossberg 1980).

The Repp (1980) data illustrate the important fact that the duration of a consciously perceived interval of silence is sensitive to the phonetic context into which the silence is placed. These data show that the phonetic context can generate a conscious percept of continuous sound across 150 ms of silence – that can be heard as silence in a different phonetic context. Our explanation of these data in terms of the maintenance of resonance in one case, but its rapid reset in another, is consistent with a simple, but revolutionary, definition of silence: Silence is a temporal discontinuity in the rate with which the auditory resonance evolves in time. Various other models of speech perception, having no concept like resonance on which to build, cannot begin to explain data of this type. Several such models are reviewed in Grossberg et al. (1997a).

1.9 Resonant Dynamics During Auditory Streaming

A similar type of resonant processing helps to explain cocktail party separation of distinct voices into auditory streams, as in the auditory continuity illusion of Fig. 1.2. This process goes on, however, at earlier stages of auditory processing than speech categorization. My colleagues Krishna Govindarajan, Lonce Wyse, Michael Cohen, and I have developed a model, called the ARTSTREAM model, of how distinguishable auditory streams are resonantly formed and separated (Grossberg 1999b; Grossberg et al. 2004). Here the two main processing levels (Fig. 1.7) are a spectral stream level at which the frequencies of the sound spectrum are represented across a spatial map, and a pitch stream level at which pitch nodes respond

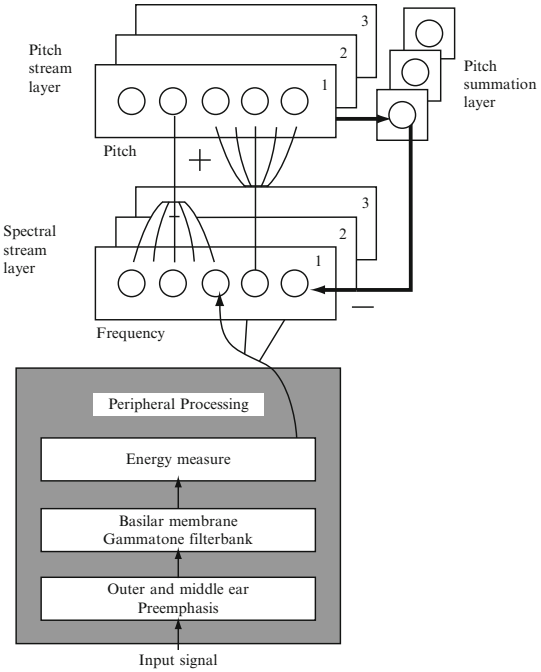


Fig. 1.7 Block diagram of the ARTSTREAM auditory streaming model. Note the nonspecific topdown inhibitory signals from the pitch level to the spectral level that realize ART matching within the network

to the harmonics at the spectral stream level that comprise a given pitch. After the auditory signal is preprocessed, its spectral, or frequency, components are redundantly represented in multiple spectral streams; that is, the sound's preprocessed frequency components are represented in multiple spatial maps, each one of which can subserve the percept of a particular auditory stream. Otherwise expressed, each frequency is represented by a *strip* of cells that can be cut into multiple streams by the network's cooperative-competitive interactions.

Each of these spectral streams is filtered by bottom-up signals that activate its own pitch stream representation at the pitch stream level; that is, there are multiple pitch streams, one corresponding to every spectral stream. This multiple representation of a sound's spectral components and pitch interact to break up the entire sound stream that is entering the system into distinct acoustic sources or voices. This happens as follows. A given sound spectrum is multiply represented at all the spectral streams and then redundantly activates all of the pitch nodes that are consistent with these sounds. These pitch representations compete to select a winner, which inhibits the representations of the same pitch across streams, while also sending top-down matching signals back to the spectral stream level. By the ART matching rule, the frequency components that are consistent with the winning pitch node are amplified, and all others are suppressed, thereby leading to a spectral-pitch resonance within the stream of the winning pitch node. In this way, the pitch layer coherently

binds together the harmonically related frequency components that correspond to a prescribed auditory source. All the frequency components that are suppressed by ART matching in this stream are freed to activate and resonate with a different pitch in a different stream. The net result is multiple resonances, each selectively grouping together into pitches those frequencies that correspond to distinct auditory sources.

Using the ARTSTREAM model, we have simulated many of basic streaming percepts, including the auditory continuity illusion of Fig. 1.2. It occurs, I contend, because the spectral stream resonance takes a time to develop that is commensurate to the duration of the subsequent noise. Once the tone resonance develops, the second tone can quickly act to support and maintain it throughout the duration of the noise, much as [ba] fuses with [ib] during perception of [iba]. Of course, for this to make sense, one needs to accept the fact that the tone resonance does not start to get consciously heard until just about when the second tone occurs.

1.10 A Circuit for ART Matching

Figure 1.7 incorporates one of the possible ways that Gail Carpenter and I proposed in the mid-1980s for how the ART matching rule can be realized (Carpenter and Grossberg 1987a). This matching circuit is redrawn in Fig. 1.8 for clarity. It is perhaps the simplest such circuit, and I have found it in subsequent studies to be the one that is implicated by data time and time again.

In this circuit, bottom-up signals to the spectral stream level can excite their target nodes if top-down signals are not active. Top-down signals try to excite those spectral, or frequency component, nodes that are consistent with the pitch node that activates them. By themselves, top-down signals fail to activate spectral nodes because the pitch node also activates a pitch summation layer that nonspecifically inhibits all spectral nodes in its stream. The nonspecific top-down inhibition hereby prevents the specific top-down excitation from supraliminally activating any spectral nodes. On the other hand, when excitatory bottom-up and top-down signals occur together, then those spectral nodes that receive both types of signals can be fully activated. All other nodes in that stream are inhibited, including spectral nodes

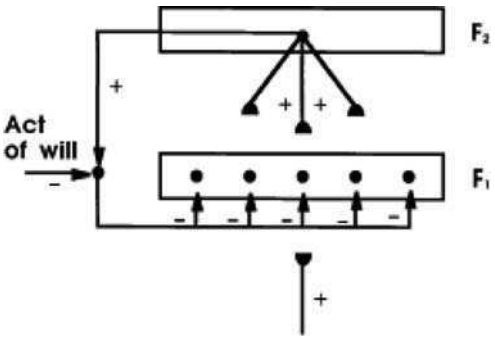


Fig. 1.8 One way to realize the ART matching rule using top-down modulatory on-center, off-surround network. See Carpenter and Grossberg (1987a)

that were previously activated by bottom-up signals but received no subsequent top-down pitch support. Attention hereby selectively activates consistent nodes while nonselectively inhibiting all other nodes in a stream.

1.11 Resonant Dynamics During Brightness Perception

Having come this far, let us review how ART matching and resonance help to explain the enhanced brightness of the Ehrenstein disk in Fig. 1.1a. This apparently simple percept has attracted a great deal of attention from vision scientists because one could imagine many reasons why no brightness difference or the reverse brightness difference might have been seen instead. John Kennedy (1979, 1988) has attempted to explain this percept by positing that "brightness buttons" occur at the ends of dark (low luminance) lines. The textbook mechanism for explaining these brightness buttons has, in turn, for decades been an appeal to the on-center, off-surround receptive fields of early visual processing. A cell that possesses such a receptive field is excited by inputs near the cell's location (the on-center) but inhibited by inputs to more distant locations (the off-surround).

An analysis of how such cells respond to dark lines shows, however, that they cannot, by themselves, explain brightness buttons. I show below why neither on-center off-surround cells (called ON cells below) nor off-center on-surround cells (called OFF cells below) can explain this phenomenon. Such ON and OFF cells occur in the lateral geniculate nucleus (or LGN), which is a way-station from the photosensitive retina in the eye to the visual cortex. Thus the ON and OFF cells that occur in the LGN, and that are the source of cortical brightness percepts, cannot explain brightness buttons without further processing. Figure 1.9 shows that whatever contribution to area contrast is generated at the ends of thin lines by ON or OFF cells must be less in magnitude than that generated along their sides. As explained below, this should make the Ehrenstein disk appear darker, rather than brighter, than its surround.

To see why this is so, assume, as in Fig. 1.9b, that the thin line is black (low luminance) and surrounded by a white (high luminance) background. Since OFF cells respond best to low luminance in their receptive field center and high luminance in their surround, OFF cells whose centers lie inside the line will be activated. Furthermore, OFF cells near the line end (but still inside the line) will be more strongly activated than OFF cells in the middle of the line because the line end is more like a black disk surrounded by a white background than the line middle is (Fig. 1.9b). That is, an OFF cell whose center lies in the line end receives less inhibition from its surround than does a cell centered in the middle of the line because a larger area of the former cell's surround lies in the white background.

A similar analysis can be applied to the ON cells. An ON cell is excited by high luminance in the center of its receptive field and low luminance in its surround. The ON cells that are active, then, are those centered outside the bar. An ON cell whose center is just outside the side of the line will respond more strongly than an ON cell centered just outside the end of the line (Fig. 1.9c).

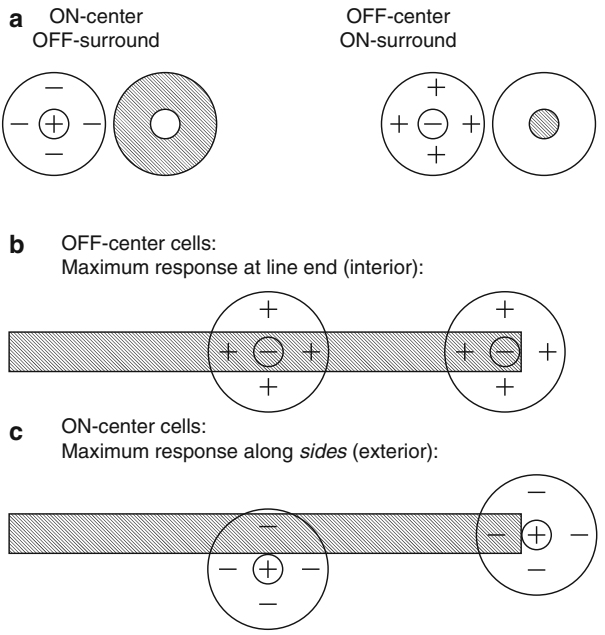


Fig. 1.9 Retinal center-surround cells and their optimal stimuli (**a**). The ON cell, on the *left*, responds best to a high-luminance disk surrounded by a low-luminance annulus. The OFF cell, on the *right*, responds best to a low-luminance disk surrounded by a high-luminance annulus (**b**). OFF cells respond to the inside of a *black line*. The OFF cell centered at the line end responds more strongly than the OFF cell centered in the middle because the surround region of the former cell is closer to optimal. In (**c**), ON cells respond to the white background just outside the *black line*. The amount of overlap of each ON cell's surround with the black line affects the strength of the cell's response. As seen in the ON cell's optimal stimulus (**a**), the more of the surround that is stimulated by a *black region*, the better the ON cell will respond. Thus, an ON cell centered just outside the side of the line will respond better than a cell centered just outside the end of the line because more of the off-surround is activated at the end of the line than along its side

Given that LGN ON and OFF cells, by themselves, cannot explain brightness buttons, an additional explanation needs to be found for how a brighter Ehrenstein disk could be generated. Clues were provided by John Kennedy, who analyzed a number of illusory contour stimuli. He argued that the effect of brightness buttons could often go unnoticed for isolated line segments, but could somehow be pooled and amplified in perceptual salience when several brightness buttons occurred in proximity or within a figurally complete region. In the mid-1980s, I worked with several colleagues to develop an analysis and interpretation of Kennedy's remarks by developing a neural model of visual boundary and surface representation (Cohen and Grossberg 1984; Grossberg and Mingolla 1985a,b; Grossberg and Todorovic' 1988).

In this model, the crucial mechanistic support for perceptually noticeable brightness buttons is a boundary segmentation that separates the region containing the buttons from other regions of a scene. Such a boundary segmentation may

be generated by image edges, textures, or shading and may give rise to illusory contours such as the Ehrenstein circle. We suggested how brightness buttons could, at a later processing stage, activate a diffusion process that could "fill-in" a uniform level of brightness within the bounding illusory contour. The model successfully explained and predicted many facts about illusory contours and brightness percepts, among other phenomena, but it incorrectly predicted that the Ehrenstein disk should look darker than its surround. Given that so many brightness data had been correctly predicted by the model, including data collected after its publication, the question arose of how the model's description was incomplete or incorrect. Such an analysis was recently carried out with Alan Gove and Ennio Mingolla (Gove et al. 1995). We showed how the addition of a feedback loop from the visual cortex to the LGN helps to explain brightness buttons without disturbing the model's previous explanations of other brightness phenomena.

The gist of this analysis can be summarized as follows. Brightness buttons are, by definition, an effect of an oriented structure such as a line or, more generally, a corner or sharp bend in a contour, on perceived brightness. Within the prior model, the computations leading to brightness perception were unoriented in the sense that they were initiated by ON and OFF cells with circularly symmetric receptive fields. How then could the effects of oriented filtering be used to modulate the inputs to the process that produces brightness buttons? Indeed, oriented filtering alone could not suffice. Interactions must exist among the oriented filters to determine the location of the ends of the lines at which the brightness buttons occur. A natural candidate for the latter interactions is the cortical endstopping process that has been known, since the Nobel-prize winning work of David Hubel and Thorstein Wiesel, to convert cortical complex cells into endstopped complex, or hypercomplex, cells (Hubel and Wiesel 1977). These oriented cells are selectively activated at and near the ends of lines. Where should the results of this endstopped processing have their effect on brightness processing?

Having come this far, it is plausible to propose that the cortex influences LGN cells via top-down feedback, which it is well known to do. It is not plausible, however, that this massive feedback pathway exists just to make Ehrenstein disks appear bright. I had, however, earlier predicted that corticogeniculate feedback exists for a potentially important functional reason; namely, to enhance the activity of LGN cells that support the activity of presently active cortical cells and to suppress the activity of LGN cells that do not (Grossberg, 1976a,b, 1980). In addition, bottom-up retinal input, by itself, was hypothesized to supraliminally activate LGN cells, but top-down corticogeniculate feedback, by itself, was not. In other words, corticogeniculate feedback was predicted to realize an ART matching and resonance rule in order to control and stabilize learned changes in cortical LTM traces in response to the flood of visual experience.

Figure 1.10 summarizes how this type of corticogeniculate feedback can produce brightness buttons. Figure 1.11b summarizes a computer simulation of brightness buttons. The model's boundary completion network generates the circular illusory contour of Fig. 1.11c. The brightness button activation pattern in Fig. 1.11b generates a topographic input to a filling-in domain, wherein the inputs diffuse freely

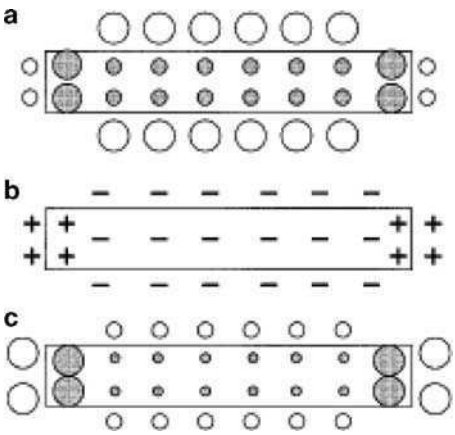


Fig. 1.10 Schematic diagram of brightness button formation in the model. In (**a**) the distribution of model LGN cell activities prior to receiving any feedback in response to a *black bar* is illustrated. Open circles code ON cell activity; *filled circles* code OFF cell activity. (**b**) The effect of feedback on bottom-up LGN activations. The *plus (minus)* signs designate the excitatory (inhibitory) top-down influence of an oriented endstopped cortical cell. (**c**) The LGN activity distribution after endstopped feedback, such as that in (**b**), combines with the direct effect of ON and OFF cell processing, such as that in (**a**). A brightness button is formed outside both ends of the line

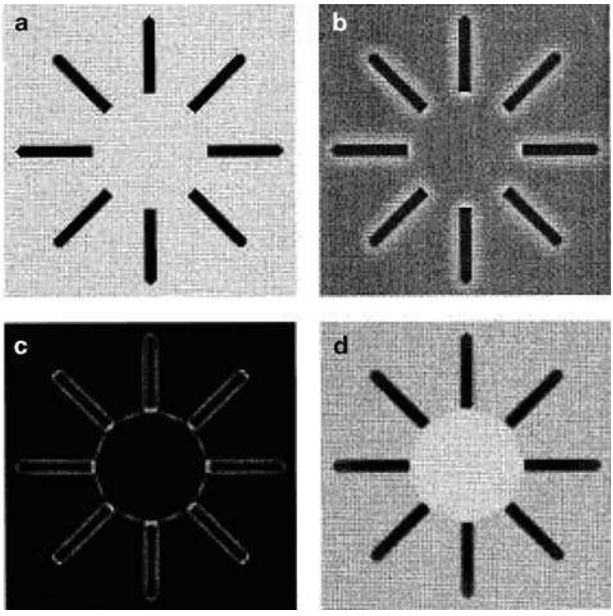


Fig. 1.11 (**a**) The Ehrenstein figure. (**b**) The LGN stage response. Both ON and OFF cell activities are coded as rectified deflections from a neutral gray. Note the brightness buttons at the line ends. (**c**) The equilibrium boundaries. (**d**) In the filled-in surface brightness, the central disk contains larger activities than the background, corresponding to the perception of increased brightness (reprinted with permission from Gove, Grossberg, Mingolla 1995)

in all directions until they hit a barrier to filling-in that is imposed by the circular boundary signals in Fig. 1.11c. The result is an Ehrenstein disk with uniformly enhanced brightness relative to its surround in Fig. 1.11d.

Is there direct experimental evidence that corticogeniculate feedback can alter LGN cell properties as desired? Murphy and Sillito (1987) showed that cortical feedback causes significant length-tuning in cat LGN cells. As in cortical endstopping, the response to a line grows rapidly as a function of line length and then abruptly declines for longer lines. The response to long lines is hereby depressed. Redies et al. (1986) found that cat dorsal LGN cells and strongly endstopped cortical complex cells responded best at line ends. In other words, the response of the LGN cells to line ends was enhanced relative to the response to line sides.

Is there direct experimental evidence for the prediction that corticogeniculate feedback supports ART matching and resonance? In a remarkable 1994 Nature article, Sillito et al. (1994) published neurophysiological data that strikingly support this prediction. They wrote in particular that "cortically induced correlation of relay cell activity produces coherent firing in those groups of relay cells with receptive field alignments appropriate to signal the particular orientation of the moving contour to the cortex . . . this increases the gain of the input for featurelinked events detected by the cortex . . . the cortico-thalamic input is only strong enough to exert an effect on those dLGN cells that are additionally polarized by their retinal input . . . the feedback circuit searches for correlations that support the 'hypothesis' represented by a particular pattern of cortical activity." In short, Sillito verified all the properties of the ART matching rule.

1.12 How Early Does Attention Act in the Brain?

If we take these results at face value, then it would appear that corticogeniculate feedback helps to "focus attention" upon expected patterns of LGN activity. However, it is typically argued that visual attention first acts at much higher levels of cortical organization, starting with the extrastriate visual cortex. Is there a contradiction here? The answer depends upon how you define attention. If attention refers only to processes that can be controlled voluntarily, then corticogeniculate feedback, being automatic, may not qualify. On the other hand, corticogeniculate feedback does appear to have the selective properties of an "automatic" attention process.

1.13 Attention at All Stages of Sensory and Cognitive Neocortex?

It has, in fact, been suggested how similar automatic attentional processes are integrated within the laminar circuits of visual cortex, notably the circuits of cortical areas V1 and V2 that are used to generate perceptual groupings, such as the illusory contours in Fig. 1.1 (Grossberg 1999a). In this proposal, the ART matching rule is

realized as follows. Top-down attentional feedback from cortical area V2 to V1 is predicted to be mediated by signals from layer 6 of cortical area V2. These top-down signals attentionally prime layer 4 of cortical area V1 via an on-center off-surround network within V1 from layer 6 to layer 4. In this conception, layer 6 of V2 activates layer 6 of V1, possibly via a multisynaptic pathway, which in turn activates layer 4 of V1 via an on-center off-surround network from layer 6 to layer 4. This analysis predicts that the layer-6-to-layer-4 on-center off-surround circuit can modulate layer 4 cells, but cannot fully activate them because the top-down attentional prime, acting by itself, is subliminal. Such a modulatory effect is achieved by appropriately balancing the strength of the on-center and off-surround signals within the layer-6-to-layer-4 network.

Related modeling work has shown how such balanced on-center off-surround signals can lead to self-stabilizing development of the horizontal connections within layer 2/3 of V1 and V2 that subserve perceptual grouping (Grossberg and Williamson 2001). It has also been shown how the top-down on-center off-surround circuit from area V1 to LGN can self-stabilize the development of disparity-sensitive complex cells in area V1 (Grunewald and Grossberg 1998). Other modeling work has suggested how a similar top-down on-center off-surround automatical attentional circuit from cortical area MST to MT can be used to generate coherent representations of the direction and speed with which objects move (Chey et al. 1997). Taken together, these studies show how the ART Matching Rule may be realized in known cortical circuits, and how it can self-stabilize development of these circuits as a precursor to its role in self-stabilizing later learning throughout life. Grossberg (1999a) has predicted that the same ART matching circuit exists within the laminar organization that is found universally in all sensory and cognitive neocortex, including the various examples of auditory processing that are reviewed above. This prediction does not, of course, deny that these circuits may be specialized in various ways to process the different types of information with which they are confronted.

Given that the cortical organization of top-down on-center off-surround attentional priming circuits seems to be ubiquitous in visual cortex, and by extension in other types of cortex, it is important to ask: What more does the brain need to add in order to generate a more flexible, task-dependent type of attention switching? This question leads us to our last example, that of visual object recognition, and how it breaks down during medial temporal amnesia. Various other models of object recognition, and their conceptual and explanatory weaknesses relative to ART, are reviewed in Grossberg and Merrill (1996).

1.14 Self-Organizing Feature Maps for Learned Object Recognition

Let us begin with a two-level network that illustrates some of the main ideas in the simplest possible way. Level F_1 in Fig. 1.12 contains a network of nodes, or cell populations, each of which is activated by a particular combination of sensory

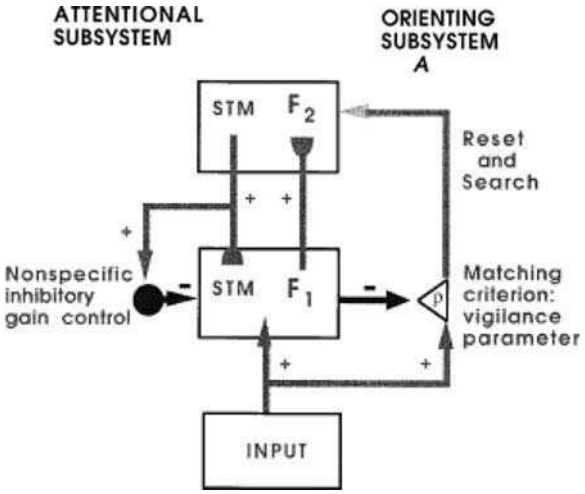


Fig. 1.12 An example of a model ART circuit in which attentional and orienting circuits interact. Level F_1 encodes a distributed representation of an event by a short-term memory (STM) activation pattern across a network of feature detectors. Level F_2 encodes the event using a compressed STM representation of the F_1 pattern. Learning of these recognition codes occurs at the long term memory (LTM) traces within the bottom-up and top-down pathways between levels F_1 and F_2. The top-down pathways read-out learned expectations whose prototypes are matched against bottom-up input patterns at F_1. The size of mismatches in response to novel events are evaluated relative to the vigilance parameter r of the orienting subsystem A. A large enough mismatch resets the recognition code that is active in STM at F_2 and initiates a memory search for a more appropriate recognition code. Output from subsystem A can also trigger an orienting response

features via inputs. Level F_2 contains a network of nodes that represent recognition codes, or categories, which are selectively activated by the activation patterns across F_1. Each F_1 node sends output signals to a subset of F_2 nodes. Each F_2 node thus receives inputs from many F_1 nodes. The thick bottom-up pathway from F_1 to F_2 in Fig. 1.12 represents in a concise way an array of diverging and converging pathways. Let learning take place at the synapses denoted by semicircular endings in the $F_1 \rightarrow F_2$ pathways. Pathways that end in arrowheads do not undergo learning. This bottom-up learning enables F_2 category nodes to become selectively tuned to particular combinations of activation patterns across F_1 feature detectors by changing their LTM traces.

Why is not bottom-up learning sufficient in a system that can autonomously solve the stability–plasticity dilemma? Why are learned top-down expectations also needed? To understand this, we consider a type of model that is often called a self-organizing feature map, competitive learning, or learned vector quantization. This type of model shows how to combine associative learning and lateral inhibition for purposes of learned categorization.

In such a model, as shown in Fig. 1.13a, an input pattern registers itself as a pattern of activity, or STM, across the feature detectors of level F_1. Each F_1 output signal is multiplied, or gated, by the adaptive weight, or LTM trace, in its respective

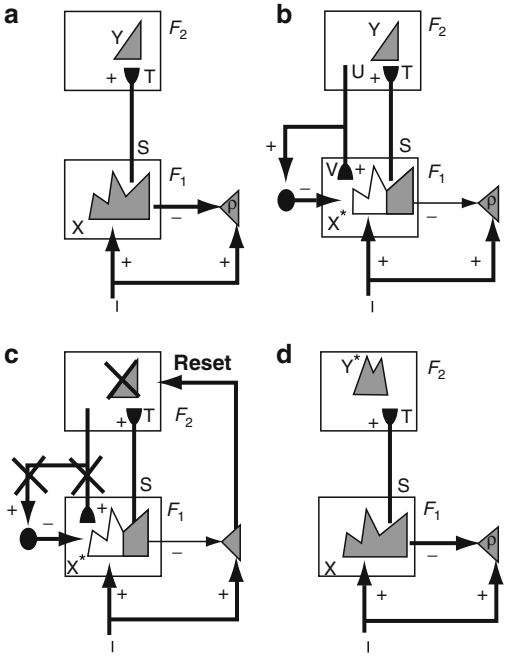


Fig. 1.13 ART search for a recognition code: (**a**) The input pattern I is instated across the feature detectors at level F_1 as a short-term memory (STM) activity pattern X. Input I also nonspecifically activates the orienting subsystem A; see Fig. 1.12. STM pattern X is represented by the hatched pattern across F_1. Pattern X both inhibits A and generates the output pattern S. Pattern S is multiplied by longterm memory (LTM) traces and added at F_2 nodes to form the input pattern T, which activates the STM pattern Y across the recognition categories coded at level F_2. (**b**) Pattern Y generates the top-down output pattern U, which is multiplied by top-down LTM traces and added at F_1 nodes to form the prototype pattern V that encodes the learned expectation of the active F_2 nodes. If V mismatches I at F_1, then a new STM activity pattern X^* is generated at F_1. X^* is represented by the hatched pattern. It includes the features of I that are confirmed by V. Inactivated nodes corresponding to unconfirmed features of X are unhatched. The reduction in total STM activity which occurs when X is transformed into X^* causes a decrease in the total inhibition from F_1 to A. (**c**) If inhibition decreases sufficiently, A releases a nonspecific arousal wave to F_2, which resets the STM pattern Y at F_2. (**d**) After Y is inhibited, its top-down prototype signal is eliminated, and X can be reinstated at F_1. Enduring traces of the prior reset lead X to activate a different STM pattern Y^* at F_2. If the top-down prototype due to Y^* also mismatches I at F_1, then the search for an appropriate F_2 code continues until a more appropriate F_2 representation is selected. Then an attentive resonance develops and learning of the attended data is initiated (reprinted with permission from Carpenter and Grossberg (1987a)

pathway. All these LTM-gated inputs are added up at their target F_2 nodes. The LTM traces hereby filter the STM signal pattern and generate larger inputs to those F_2 nodes whose LTM patterns are most similar to the STM pattern. Lateral inhibitory, or competitive, interactions within F_2 contrast-enhance this input pattern. Whereas many F_2 nodes may receive inputs from F_1, lateral inhibition allows a much smaller set of F_2 nodes to store their activation in STM. These are the F_2 nodes whose LTM

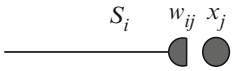
patterns are most similar to the STM pattern. These inhibitory interactions also tend to conserve the total activity that is stored in STM (Grossberg 1980, 1982), thereby realizing an interference-based capacity limitation in STM.

Only the F_2 nodes that win the competition and store their activity in STM can influence the learning process. STM activity opens a learning gate at the LTM traces that abut the winning nodes. These LTM traces can then approach, or track, the input signals in their pathways, a process called steepest descent. Such a learning law is thus often called gated steepest descent, or instar learning. This type of learning tunes the winning LTM patterns to become even more similar to the STM pattern and to thereby enable the STM pattern to more effectively activate the corresponding F_2 nodes. I introduced this learning law into neural network models in the 1960s (e.g., Grossberg 1969) and into ART models in the 1970s (Grossberg 1976a,b, 1978, 1980). Such an LTM trace can either increase (Hebbian) or decrease (anti-Hebbian) to track the signals in its pathway (Table 1.1). It has been used to model neurophysiological data about learning in the hippocampus (also called long-term potentiation and long-term depression) and about adaptive tuning of cortical feature detectors during early visual development (Artola and Singer 1993; Levy 1985; Levy and Desmond 1985; Rauschecker and Singer 1979; Singer 1983), thereby lending support to ART predictions that these systems would employ this type of learning.

Self-organizing feature map models were introduced and computationally characterized by Christoph von der Malsburg and me during the 1970s (Grossberg 1972, 1976a, 1978; von der Malsburg et al. 1973; Willshaw et al. 1976). These models were subsequently applied and further developed by many authors, notably Teuvo Kohonen (1984). They exhibit many useful properties, especially if not too many input patterns, or clusters of input patterns, perturb level F_1 relative to the number

TABLE 1.1

The Instar Learning, or Gated Steepest Descent Learning Rule, Embodies both Hebbian (LTP) and anti-Hebbian (LTD) Properties within a Single Process[a]



	Case 1	Case 2	Case 3	Case 4
State of S_i	+	−	+	−
State of x_j	+	+	−	−
State of w_{ij}	↑	↓	↔	↔

Note. Symbols: + = active; − = inactive; ↑ = increase; ↓ = decrease; ↔ = no change.

[a] Reprinted with permission from Grossberg and Merrill (1996).

of categorizing nodes in level F_2. I proved that, under these sparse environmental conditions, category learning is stable in the sense that its LTM traces converge to fixed values as learning trials proceed. In addition, the LTM traces track the statistics of the environment, are self-normalizing, and oscillate a minimum number of times (Grossberg 1976a). Also, the category selection rule, like a Bayesian classifier, tends to minimize error. I also proved, however, that under arbitrary environmental conditions, learning becomes unstable (Grossberg 1976b). Such a model could forget your parents' faces when it learns a new face. Although a gradual switching off of plasticity can partially overcome this problem, such a mechanism cannot work in a learning system whose plasticity is maintained throughout adulthood.

This memory instability is due to basic properties of associative learning and lateral inhibition, which are two processes that occur ubiquitously in the brain. An analysis of this instability, together with data about human and animal categorization, conditioning, and attention, led me to introduce ART models to stabilize the memory of self-organizing feature maps in response to an arbitrary stream of input patterns.

1.15 How Does ART Stabilize Learning of a Self-Organizing Feature Map?

How does an ART model prevent such instabilities from developing? As noted above, in an ART model, learning does not occur when some winning F_2 activities are stored in STM. Instead, activation of F_2 nodes may be interpreted as "making a hypothesis" about an input at F_1. When F_2 is activated, it quickly generates an output pattern that is transmitted along the top-down adaptive pathways from F_2 to F_1. These top-down signals are multiplied in their respective pathways by LTM traces at the semicircular synaptic knobs of Fig. 1.13b. The LTM-gated signals from all the active F_2 nodes are added to generate the total topdown feedback pattern from F_2 to F_1. It is this pattern that plays the role of a learned expectation. Activation of this expectation may be interpreted as "testing the hypothesis," or "reading out the prototype," of the active F_2 category. As shown in Fig. 1.13b, ART networks are designed to match the "expected prototype" of the category against the bottom-up input pattern, or exemplar, to F_1. Nodes that are activated by this exemplar are suppressed if they do not correspond to large LTM traces in the top-down prototype pattern. The resultant F_1 pattern encodes the cluster of input features that the network deems relevant to the hypothesis based upon its past experience. This resultant activity pattern, called X^* in Fig. 1.13b, encodes the pattern of features to which the network "pays attention."

If the expectation is close enough to the input exemplar, then a state of resonance develops as the attentional focus takes hold. The pattern X^* of attended features reactivates the F_2 category Y which, in turn, reactivates X^*. The network locks into a resonant state through a positive feedback loop that dynamically links, or binds,

X* with Y. The resonance binds spatially distributed features into either a stable equilibrium or a synchronous oscillation, much like the synchronous feature binding in visual cortex that has recently attracted so much interest after the experiments of Reinhard Eckhorn and Wolf Singer and their colleagues (Eckhorn et al. 1988; Gray and Singer 1989); also see Grossberg and Grunewald (1997).

In ART, the resonant state, rather than bottom-up activation, is predicted to drive the learning process. The resonant state persists long enough, at a high enough activity level, to activate the slower learning processes in the LTM traces. This helps to explain how the LTM traces can regulate the brain's fast information processing without necessarily learning about the signals that they process. Through resonance as a mediating event, the combination of top-down matching and attentional focusing helps to stabilize ART learning and memory in response to an arbitrary input environment. The stabilizing properties of top-down matching may be one reason for the ubiquitous occurrence of reciprocal bottom-up and top-down corticocortical and corticothalamic interactions in the brain.

1.16 How Is the Generality of Knowledge Controlled?

A key problem about consciousness concerns what combinations of features or other information are bound together into object or event representations. ART provides a new answer to this question that overcomes problems faced by earlier models. In particular, ART systems learn prototypes, rather than exemplars, because the attended feature vector X*, rather than the input exemplar itself, is learned. Both the bottom-up LTM traces that tune the category nodes and the top-down LTM traces that filter the learned expectation learn to correlate activation of F_2 nodes with the set of all attended X* vectors that they have ever experienced. These attended STM vectors assign less STM activity to features in the input vector I that mismatch the learned top-down prototype V than to features that match V.

Given that ART systems learn prototypes, how can they also learn to recognize unique experiences, such as a particular view of a friend's face? The prototypes learned by ART systems accomplish this by realizing a qualitatively different concept of prototype than that offered by previous models. In particular, Gail Carpenter and I have shown with our students how ART prototypes form in a way that is designed to conjointly maximize category generalization while minimizing predictive error (Bradski and Grossberg 1995; Carpenter and Grossberg 1987a,b; Carpenter et al. 1991, 1992). As a result, ART prototypes can automatically learn individual exemplars when environmental conditions require highly selective discriminations to be made. How the matching process achieves this is discussed below.

Before describing how this is achieved, let us note what happens if the mismatch between bottom-up and top-down information is too great for a resonance to develop. Then the F_2 category is quickly reset and a memory search for a better category is initiated. This combination of top-down matching, attention focusing, and

memory search is what stabilizes ART learning and memory in an arbitrary input environment. The attentional focusing by top-down matching prevents inputs that represent irrelevant features at F_1 from eroding the memory of previously learned LTM prototypes. In addition, the memory search resets F_2 categories so quickly when their prototype V mismatches the input vector I that the more slowly varying LTM traces do not have an opportunity to correlate the attended F_1 activity vector X^* with them. Conversely, the resonant event, when it does occur, maintains, amplifies, and synchronizes the matched STM activities for long enough and at high enough amplitudes for learning to occur in the LTM traces.

Whether a resonance occurs depends upon the level of mismatch, or novelty, that the network is prepared to tolerate. Novelty is measured by how well a given exemplar matches the prototype that its presentation evokes. The criterion of an acceptable match is defined by an internally controlled parameter ρ that Carpenter and I have called vigilance (Carpenter and Grossberg 1987a). The vigilance parameter is computed in the orienting subsystem A; see Fig. 1.12. Vigilance weighs how similar an input exemplar I must be to a top-down prototype V in order for resonance to occur. Resonance occurs if $\rho|I| - |X^*| \leq 0$. This inequality says that the F_1 attentional focus X^* inhibits A more than the input I excites it. If A remains quiet, then an $F_1 \leftrightarrow F_2$ resonance can develop.

Either a larger value of ρ or a smaller match ratio $|X^*||I|^{-1}$ makes it harder to satisfy the resonance inequality. When ρ grows so large or $|X^*||I|^{-1}$ is so small that $\rho|I| - |X^*| > 0$, then A generates an arousal burst, or novelty wave, that resets the STM pattern across F_2 and initiates a bout of hypothesis testing, or memory search. During search, the orienting subsystem interacts with the attentional subsystem (Fig. 1.13c and 1.13d) to rapidly reset mismatched categories and to select better F_2 representations with which to categorize novel events at F_1, without risking unselective forgetting of previous knowledge. Search may select a familiar category if its prototype is similar enough to the input to satisfy the resonance criterion. The prototype may then be refined by attentional focusing. If the input is too different from any previously learned prototype, then an uncommitted population of F_2 cells is selected and learning of a new category is initiated.

Because vigilance can vary across learning trials, recognition categories capable of encoding widely differing degrees of generalization or abstraction can be learned by a single ART system. Low vigilance leads to broad generalization and abstract prototypes. High vigilance leads to narrow generalization and to prototypes that represent fewer input exemplars, even a single exemplar. Thus a single ART system may be used, say, to learn abstract prototypes with which to recognize abstract categories of faces and dogs, as well as "exemplar prototypes" with which to recognize individual faces and dogs. A single system can learn both, as the need arises, by increasing vigilance just enough to activate A if a previous categorization leads to a predictive error. Thus the contents of a conscious percept can be modified by environmentally sensitive vigilance control.

Vigilance control hereby allows ART to overcome some fundamental difficulties that have been faced by classical exemplar and prototype theories of learning and recognition. Classical exemplar models face a serious combinatorial explosion,

since they need to suppose that all experienced exemplars are somehow stored in memory and searched during performance. Classical prototype theories face the problem that they find it hard to explain how individual exemplars are learned, such as a particular view of a familiar face. Vigilance control enables ART to achieve the best of both types of model by selecting the most general category that is consistent with environmental feedback. If that category is an exemplar, then a "very vigilant" ART model can learn it. If the category is at an intermediate level of generalization, then the ART model can learn it by having the vigilance value track the level of match between the current exemplar and the prototype that it activates. In every instance, the model tries to learn the most general category that is consistent with the data. This tendency can, for example, lead to the type of overgeneralization that is seen in young children until further learning leads to category refinement (Chapman et al. 1986; Clark 1973; Smith et al. 1985; Smith and Kemler 1978; Ward 1983). Many benchmark studies of how ART uses vigilance control to classify complex data bases have shown that the number of ART categories that is learned scales well with the complexity of the input data; see Carpenter and Grossberg (1994) for a list of illustrative benchmark studies.

1.17 Corticohippocampal Interactions and Medial Temporal Amnesia

As sequences of inputs are practiced over learning trials, the search process eventually converges upon stable categories. Carpenter and I mathematically proved (Carpenter and Grossberg 1987a) that familiar inputs directly access the category whose prototype provides the globally best match, while unfamiliar inputs engage the orienting subsystem to trigger memory searches for better categories until they become familiar. This process continues until the memory capacity, which can be chosen arbitrarily large, is fully utilized. The process whereby search is automatically disengaged is a form of memory consolidation that emerges from network interactions. Emergent consolidation does not preclude structural consolidation at individual cells, since the amplified and prolonged activities that subserve a resonance may be a trigger for learning-dependent cellular processes, such as protein synthesis and transmitter production. It has also been shown that the adaptive weights which are learned by an ART model at any stage of learning can be translated into IF-THEN rules (e.g., Carpenter et al. 1992). Thus the ART model is a self-organizing rule-discovering production system as well as a neural network.

The attentional subsystem of ART has been used to model aspects of inferotemporal (IT) cortex, and the orienting subsystem models part of the hippocampal system. The interpretation of ART dynamics in terms of IT cortex led Miller, Li, and Desimone (1991) to successfully test the prediction that cells in monkey IT cortex are reset after each trial in a working memory task. To illustrate the implications of an ART interpretation of IT–hippocampal interactions, I review how a lesion of the ART model's orienting subsystem creates a formal memory disorder with symptoms

much like the medial temporal amnesia that is caused in animals and human patients after hippocampal system lesions (Carpenter and Grossberg 1993; Grossberg and Merrill 1996). In particular, such a lesion in vivo causes unlimited anterograde amnesia; limited retrograde amnesia; failure of consolidation; tendency to learn the first event in a series; abnormal reactions to novelty, including perseverative reactions; normal priming; and normal information processing of familiar events (Cohen 1984; Graf et al. 1984; Lynch et al. 1984; Squire and Butters 1984; Squire and Cohen 1984; Warrington and Weiskrantz 1974; Zola-Morgan and Squire 1990).

Unlimited anterograde amnesia occurs because the network cannot carry out the memory search to learn a new recognition code. Limited retrograde amnesia occurs because familiar events can directly access correct recognition codes. Before events become familiar, memory consolidation occurs which utilizes the orienting subsystem (Fig. 1.13c). This failure of consolidation does not necessarily prevent learning per se. Instead, learning influences the first recognition category activated by bottom-up processing, much as amnesics are particularly strongly wedded to the first response they learn. Perseverative reactions can occur because the orienting subsystem cannot reset sensory representations or top-down expectations that may be persistently mismatched by bottom-up cues. The inability to search memory prevents ART from discovering more appropriate stimulus combinations to attend. Normal priming occurs because it is mediated by the attentional subsystem.

Similar behavioral problems have been identified in hippocampectomized monkeys. Gaffan (1985) noted that fornix transection "impairs ability to change an established habit . . . in a different set of circumstances that is similar to the first and therefore liable to be confused with it." In ART, a defective orienting subsystem prevents the memory search whereby different representations could be learned for similar events. Pribram (1986) called such a process a "competence for recombinant context-sensitive processing." These ART mechanisms illustrate how, as Zola-Morgan and Squire (1990) have reported, memory consolidation and novelty detection may be mediated by the same neural structures. Why hippocampectomized rats have difficulty orienting to novel cues and why there is a progressive reduction in novelty-related hippocampal potentials as learning proceeds in normal rats is also clarified (Deadwyler et al. 1979, 1981). In ART, the orienting system is automatically disengaged as events become familiar during the memory consolidation process. The ART model of normal and abnormal recognition learning and memory is compared with several other recent models of these phenomena in Grossberg and Merrill (1996).

At this point, it might also be useful to note that the processes of automatic and task-selective attention may not be independent in vivo. This is because higher-order attentional constraints that may be under task-selective control can in principle propagate downward through successive cortical levels via layer-6-to-layer-6 linkages. For example, recent modeling work has suggested how prestriate cortical areas may separate visual objects from one another and from their backgrounds during the process of figure–ground separation (Grossberg 1994, 1997; Grossberg and McLoughlin 1997). Such constraints may propagate top-down toward earlier cortical levels, possibly even area V1, to modulate the cells that get active there to be consistent with these figure–ground constraints. Still higher cortical processes, such

as those involved in learned categorization, may also propagate their modulatory constraints to lower levels. How the strength of such top-down modulatory influences depends upon the source cortical area and on the number of synaptic steps to the target cortical area is a topic that has yet to be systematically studied.

1.18 How Universal Are ART Processes in the Brain?

In all the examples discussed above – from early vision, visual object recognition, auditory streaming, and speech recognition – ART matching and resonance have played a central role in models that help to explain how the brain stabilizes its learned adaptations in response to changing environmental conditions. This type of matching can be achieved using a top-down nonspecific inhibitory gain control that downregulates all target cells except those that also receive top-down specific excitatory signals, as in Fig. 1.8. Are there yet other brain processes that utilize these mechanisms? John Reynolds and colleagues in Bob Desimone's laboratory (Reynolds et al. 1995) have reported neurophysiological data from cells in cortical areas V2 and V4 that are consistent with the ART attentional mechanism summarized in Fig. 1.8. Taken together with studies of the V1→LGN attention circuit and of attentional control by frontal and inferotemporal cortex during visual object recognition, it may be concluded that ART-like top-down matching occurs throughout the brain's visual system.

With my colleagues Mario Aguilar, Dan Bullock, and Karen Roberts, a neural model has been developed to explain how the superior colliculus learns to use visual, auditory, somatosensory, and planned movement signals to control saccadic eye movements (Grossberg et al. 1997b). This model uses ART matching and resonance to help explain behavioral and neural data about multimodal eye movement control. The model clarifies how visual, auditory, and planned movement signals use learning to form a mutually consistent movement map and how attention gets focused on a movement target location after all these signals compete to determine where the eyes will move.

Experiments from Marcus Raichle's lab at Washington University using positron emission tomography (PET) support the idea that ART top-down priming also occurs in human somatosensory cortex (Drevets et al. 1995). In their experiments, attending to an impending stimulus to the fingers caused inhibition of nearly cortical cells that code for the face, but not cells that code the fingers. Likewise, priming of the toes produced inhibition of nearby cells that code for the fingers and face, but not cells that code for the toes.

ART models have also been used to explain a great deal of data about cognitive–emotional interactions, notably about classical and instrumental conditioning (Grossberg 1987b) and about human decision making under risk (Grossberg and Gutowski 1987). In these examples, the resonances are between cognitive and emotional circuits and help to focus attention upon, and release actions toward, valued events and objects in the world.

Thus all levels of vision, visual object recognition, auditory streaming, speech recognition, attentive selection of eye movement targets, somatosensory representation, and cognitive–emotional interactions may incorporate variants of the circuit depicted in Fig. 1.8. These results suggest that a type of "automatic" attention operates even at early levels of brain processing, such as the lateral geniculate, but that higher processing levels benefit from an orienting subsystem that can be used to flexibly reset attention and to facilitate voluntary control of top-down expectations.

1.19 Internal Fantasy, Planned Movement, and Volitional Gating

Given this type of circuit, how could top-down priming be released from inhibition to enable us to voluntarily experience internal thinking and fantasies? This can be achieved through an "act of will" that activates inhibitory cells which inhibit the nonspecific inhibitory interneurons in the top-down on-center off-surround network of Fig. 1.8. This operation disinhibits the cells receiving the excitatory top-down signals in the on-center of the network. These cells are then free to generate supraliminal resonances. Such self-initiated resonances can, for example, be initiated by the read-out of top-down expectations from higher-order planning nodes into temporally organized working memories, say in the prefrontal cortex (Fuster 1996). It is, for example, well known that the basal ganglia can use such a disinhibitory action to gate the release of individual movements, sequences of movements, and even cognitive processes (Hikosaka 1994; Middleton and Strick 1994; Sakai et al. 1998).

These examples also help to understand how top-down expectations can be used for the control of planned (viz., intentional) behavioral sequences. For example, once such planning nodes read-out their top-down expectations into working memory, the contents of working memory can be read-out and modified by on-line changes in "acts of will." These volitional signals enable invariant representations of an intentional behavior to rapidly adapt themselves to changing environmental conditions. For example, Bullock et al. (1993b) have modeled how such a working memory can control the intentional performance of handwriting whose size and speed can be modified by acts of will, without a change of handwritten form. Bullock et al. (1993a) have shown how a visual target that is stored in working memory can be reached with a novel tool that has never been used before. The latter study shows how a such a model can learn its own parameters through a type of Piagetian perform-and-test developmental cycle.

Thus we arrive at an emerging picture of how the adaptive brain works wherein the core issue of how a brain can learn quickly and stably about a changing world throughout life leads toward a mechanistic understanding of attention, intention, thinking, fantasy, and consciousness. The mediating events are adaptive resonances that effect a dynamic balance between the complementary demands of stability and plasticity and of expectation and novelty and which are a necessary condition for consciousness.

1.20 What vs Where: Why Are Procedural Memories Unconscious?

Although the type of ART matching, learning, and resonance that have been reviewed above seem to occur in many sensory and cognitive processes, they are not the only types of matching and learning to occur in the brain. In fact, there seems to be a major difference between the types of learning that occur in sensory and cognitive processes versus those that occur in spatial and motor processes. In particular, sensory and cognitive processes are carried out in the What processing stream that passes through the inferotemporal cortex, whereas spatial and motor processes are carried out in the Where processing stream that passes through the parietal cortex. What processing includes object recognition and event prediction. Where processing includes spatial navigation and motor control. I suggest that the types of matching and learning that go on in the What and Where streams are different, indeed complementary, and that this difference is appropriate to their different roles.

First, consider how we use a sensory expectation. Suppose, for example, that I ask you to "Look for the yellow ball, and if you find it within three hundred milliseconds, I will give you a million dollars." If you believed me, you could activate a sensory expectation of "yellow balls" that would make you much more sensitive to yellow and round objects in your environment. As in ART matching, once you detected a yellow ball, you could then react to it much more quickly and with a much more energetic response than if you were not looking for it. In other words, sensory and cognitive expectations lead to a type of *excitatory matching*.

Now consider how we use a motor expectation. Such an expectation represents where we want to move (Bullock and Grossberg 1988). For example, it could represent a desired position for the hand to pick up an object. Such a motor expectation is matched against where the hand is now. After the hand actually moves to the desired position, no further movement is required to satisfy the motor expectation. In this sense, motor expectations lead to a type of *inhibitory matching*. In summary, although the sensory and cognitive matching process is excitatory, the spatial and motor matching process is inhibitory. These are complementary properties. Models such as ART quantify how excitatory matching is accomplished. A different type of model, called a Vector Associative Map, or VAM, model, suggests how inhibitory matching is accomplished (Gaudiano and Grossberg 1991; Grossberg et al. 1993; Guenther et al. 1994).

As shown in the discussions of ART above, learning within the sensory and cognitive domain is often a type of *match learning*. It takes place only if there is a good enough match of top-down expectations with bottom-up data to risk altering previously stored knowledge within the system, or it can trigger learning of a new representation if a good enough match is not available. In contrast, learning within spatial and motor processes, such as VAM processes, is *mismatch learning* that is used to either learn new sensory-motor, as in maps (e.g., Grossberg et al. (1993) or to adjust the gains of sensory-motor commands (e.g., Fiala et al. 1996). These types of learning are also complementary.

Why are the types of learning that go into spatial and motor processes complementary to those that are used for sensory and cognitive processing? My answer is that ART-like learning allows the brain to solve the stability–plasticity dilemma. It enables us to continue learning more about the world in a stable fashion throughout life without forcing catastrophic forgetting of our previous memories. On the other hand, catastrophic forgetting is a good property when it takes place during spatial and motor learning. We have no need to remember all the spatial and motor maps that we used when we were infants or children. In fact, those maps would cause us a lot of trouble if they were used to control our adult limbs. We want our spatial and motor processes to continuously adapt to changes in our motor apparatus. These complementary types of learning allow our sensory and cognitive systems to stably learn about the world and to thereby be able to effectively control spatial and motor processes that continually update themselves to deal with changing conditions in our limbs.

Why, then, are procedural memories unconscious? The difference between cognitive memories and procedural, or motor, memories has gone by a number of different names, including the distinction between declarative memory and procedural memory, knowing that and knowing how, memory and habit, or memory with record and memory without record (Bruner 1969; Miskin 1982, 1993; Ryle 1949; Squire and Cohen 1984). The amnesic patient HM dramatically illustrated this distinction by learning and remembering motor skills like assembly of the Tower of Hanoi without being able to recall ever having done so (Bruner 1969; Scoville and Milner 1957; Squire and Cohen 1984). We can now give a very short answer to the question of why procedural memories are unconscious: The matching that takes place during spatial and motor processing is often inhibitory matching. Such a matching process cannot support an excitatory resonance. Hence, it cannot support consciousness.

In this regard, Goodale and Milner (1992) have described a patient whose brain lesion has prevented accurate visual discrimination of object orientation, yet whose visually guided reaching behaviors toward objects are oriented and sized correctly. We have shown, in a series of articles, how head-centered and body-centered representations of an object's spatial location and orientation may be learned and used to control reaches of the hand–arm system that can continuously adapt to changes in the sensory and motor apparatus that is used to plan and execute reaching behaviors (Bullock et al. 1993; Carpenter et al. 1998; Gaudiano and Grossberg 1991; Grossberg et al. 1993; Guenther et al. 1994). None of these model circuits has resonant loops; hence, they do not support consciousness.

When these models are combined into a more comprehensive system architecture for intelligent behavior, the sensory and cognitive match-based networks in the What processing stream through the inferotemporal cortex provide self-stabilizing representations with which to continually learn more about the world without undergoing catastrophic forgetting, while the Where/How processing stream's spatial and motor mismatch-based maps and gains can continually forget their old parameters in order to instate the new parameters that are needed to control our bodies in their present form. This larger architecture illustrates how circuits in the self-stabilizing

match-based sensory and cognitive parts of the brain can resonate into consciousness, even while they are helping to direct the contextually appropriate activation of spatial and motor circuits that cannot.

1.21 Some Comments About Amodal and Modal Visual Percepts

There are many other aspects of perception and cognition, notably of vision and visual object recognition, which can be discussed in the light of recent modeling advances to shed light on consciousness. Here I make some summarizing remarks whose detailed analysis and justification can be found in the original articles. One issue of interest concerns the distinction between modal and amodal percepts. An amodal percept, such as the percept of a vertical boundary between the offset grating in Fig. 1.1c, is one which does not carry a visible perceptual sign. As noted above, it can be recognized without being seen; we are conscious of it even though it is perceptually invisible. A modal percept, such as a percept of brightness or color, does carry a visible perceptual sign. I believe that all theories of consciousness need to deal with how amodal percepts can occur because such percepts sharply distinguish between our consciously "knowing" that an event has occurred even though we do not consciously "perceive" it.

The FACADE theory of biological vision has provided an extensive analysis of some of the conditions that determine whether a percept will be modal or amodal (e.g., Francis et al. 1994; Grossberg 1994, 1997; Grossberg and McLoughlin 1997; Grossberg and Mingolla 1985b; Gove et al. 1995). A key contribution of this theory is to suggest how visual scenes are processed in parallel by cortical boundary and surface systems, which are proposed to be realized by the interblob and blob processing streams from the LGN to cortical area V4. Boundaries include illusory contours (Fig. 1.1), as well as the boundaries that are formed in response to texture, shading, and stereo cues.

A key insight of this theory is that "all boundaries are invisible" (i.e., amodal) within the boundary processing stream, and that visibility is a property of the surface processing stream. Boundaries are invisible within the boundary processing stream because like-oriented signals from cortical simple cells that are sensitive to opposite contrast polarities are pooled at complex cells. Complex cells can hereby respond to contrasts that are either dark/light or light/dark, as can all subsequent stages of the boundary system. As a result of this pooling process, a boundary can be formed around an object whose relative contrasts with respect to its background may reverse along its perimeter. A secondary consequence is that a perceptual boundary, by pooling across opposite contrast polarities (as well as all opponent colors), cannot represent any visible property that depends upon knowing the direction of a brightness or color change.

Modal percepts are predicted to occur within the surface processing stream. Surface representations arise through interactions with the boundaries. First, the

surface stream "discounts the illuminant," or compensates for variable illumination (Helmholtz 1962; Land 1977, 1986). This discounting process eliminates brightness or color signals within homogeneously bright or colored regions of a scene, which could otherwise cause serious confusions between variable lighting conditions and the surface properties of objects in the world. At subsequent processing stages, the boundaries interact with the discounted surface signals. Here, the boundaries suppress surface signals that are not spatially coincident with them. Boundaries select surface signals that are spatially coincident with them and initiate a process of filling-in whereby these selected signals can diffuse within the controlling boundaries.

FACADE theory predicts that the boundaries which exercise this control occur subsequent to the cortical processing stage at which visual inputs from both eyes are binocularly fused. It was suggested in Grossberg (1987a) that, although the binocular matching process is initiated in cortical area V1, the stage at which the binocular boundaries are completed occurs no earlier than cortical area V2.

During binocular rivalry, the inputs to the two eyes are mismatched in such a way that image regions from only one eye at a time can be consciously perceived. FACADE theory suggests how boundary signals from the two eyes compete in a cyclical fashion through time, with the boundaries from one eye winning at any time in each position. Such competition has been traced to the mechanisms whereby a winning boundary is selected from among many possible boundary groupings, even when the inputs to both eyes represent the same scene. The cyclicity of the percept was traced to the habituative mechanisms whereby boundaries are rapidly reset in response to rapidly changing imagery in order to prevent them from persisting too long (see Francis et al. 1994) for an analysis of how long perceptual boundaries do persist). Then the winning boundaries select those surface signals from the dominant eye which are spatially coincident with them while suppressing the spatially discordant surface signals from the losing eye. The first stage of such surface capture selects the surface properties from each eye separately. The selected surface representations are predicted to be amodal. These selected surface properties are then binocularly matched at a subsequent processing stage at which the modal, or visible, surface representation is predicted to form. This is also the processing stage at which visual figures are fully separated from one another and from their backgrounds.

Grossberg (1987a) predicted that this binocular modal, or visible, representation of the winning surface percept arises in cortical area V4, which resonates with inferotemporal cortex during consciousness. Logothetis et al. (1996) have reported consistent data on binocular rivalry from awake behaving monkeys. Schiller (1994) has reported data from awake behaving monkeys that is consistent with the prediction that figure–ground separation is completed in cortical area V4.

These results support the FACADE theory prediction that amodal percepts may form in cortical areas V2 or before and that modal representations of surfaces may first occur in cortical area V4. In further support of this hypothesis, Grossberg (1994) explained many data about 3D figure–ground separation in which, say, amodal representations of occluded object parts may be formed in cortical area V2 and

used to recognize these occluded objects, even though they are not seen. Modal representations of both occluding objects and the unoccluded parts of the objects that they occlude may not be formed until cortical area V4. This is proposed despite the fact that all of these cortical processing stages may incorporate the ART matching rule within their laminar circuits and may resonate using both the intercortical and intracortical feedback pathways that activate the layer-6-to-layer-4 on-center off-surround networks, the former during attentional priming and the latter during the selection of winning perceptual groupings.

Grossberg (1997) proposed that the modally conscious surface representations in V4 may be used to recognize and to control reaching toward physically accessible objects, especially in infants, whereas the amodally conscious representations – both of boundaries and of surfaces – in V2 may be used to recognize partially occluded objects and to reach toward them via more indirect motor planning and control circuits. This proposal provides a functional reason for making some visual representations visible and others not visible; in particular, being able to distinguish between modal (e.g., occluding) and amodal (e.g., occluded) representations helps to prevent efforts to reach through an occluding object to the object that it is occluding. On the other hand, the proposal does not explain how the property of visibility is achieved by one type of representation but not the other, particularly since both types of representation may be assumed to be resonant. This fact does not contradict the hypothesis that all conscious states are resonant states. It does show, however, that further mechanisms are needed to explain why some of these resonant representations are modal whereas others are merely amodal.

The need for further mechanisms is well-illustrated by the following modeling prediction. It was predicted in Grossberg (1987a), and then used extensively to explain much more perceptual data in Grossberg (1994, 1997), that a network of double-opponent cells forms an important mechanism in the process whereby boundaries select only those surface brightness and color signals that are spatially coincident with them. Double-opponent cells are often cited as a key mechanism of color perception (e.g., Livingstone and Hubel 1984). FACADE theory suggests that such networks are used to form both amodal and modal surface representations. In the amodal surface representations, double-opponent networks are predicted not to generate a percept of visible color. Some other factor must be sought, to whose discovery future research would be profitably directed.

The author thanks Cynthia E. Bradford and Diana Meyers for their valuable assistance in the preparation of this manuscript. This work was supported in part by the Defense Advanced Research Projects Agency and the Office of Naval Research (ONR N00014-92-J-1309, ONR N00014-95-1-0409, and ONR N00014-95-1-0657, by CELEST, an NSF Science of Learning Center (SBE-0354378), and by the SyNAPSE program of the Defense Advanced Research Projects Agency (HR0011-09-C-0011).

References

Abbott LF, Varela K, Sen K, and Nelson SB (1997) Synaptic depression and cortical gain control. Science 275:220–223

Artola A, Singer W (1993) Long-term depression of excitatory synaptic transmission and its relationship to long-term potentiation. Trends Neurosci 16:480–487

Bradski G, Grossberg G (1995) Fast learning VIEWNET architectures for recognizing 3-D objects from multiple 2-D views. Neural Netw 8:1053–1080

Bregman AS (1990) Auditory scene analysis: the perceptual organization of sound. MIT, Cambridge, MA

Bruner JS (1969) In: alland GA, Waugh NC (eds) The pathology of memory. Academic, New York

Bullock D, Grossberg S (1988) Neural dynamics of planned arm movements: emergent invariants and speed-accuracy properties during trajectory formation. Psychol Rev 95:49–90

Bullock D, Grossberg S, Guenther FH (1993a) A self-organizing neural model of motor equivalent reaching and tool use by a multijoint arm. J Cogn Neurosci 5:408–435

Bullock D, Grossberg S, Mannes C (1993b) A neural network model for cursive script production. Biol Cybernetics 70:15–28

Carpenter GA, Grossberg S (1981) Adaptation and transmitter gating in vertebrate photoreceptors. J Theor Neurobiol 1:1–42. Reprinted in Grossberg S (ed) The adaptive brain (1987, vol. II, pp. 271–310). Elsevier, Amsterdam

Carpenter GA, Grossberg S (1987a) A massively parallel architecture for a self-organizing neural pattern recognition machine. Comp Vis Graph Image Process, 37:54–115

Carpenter GA, Grossberg S (1987b) ART 2: Self-organization of stable category recognition codes for analog input patterns. Appl Optics 26:4919–4930

Carpenter GA, Grossberg S, Reynolds JH (1991) ARTMAP: supervised real-time learning and classification of nonstationary data by a self-organizing neural network. Neural Netw 4:565–588

Carpenter GA, Grossberg S, Markuzon N, Reynolds JH, Rosen DB (1992) Fuzzy ARTMAP: a neural network architecture for incremental supervised learning of analog multidimensional maps. IEEE Trans Neural Netw 3:698–713

Carpenter GA, Grossberg S (1993) Normal and amnesic learning, recognition, and memory by a neural model of cortico-hippocampal interactions. Trends Neurosci 116:131–137

Carpenter GA, Grossberg S (1994) Integrating symbolic and neural processing in a self-organizing architecture for pattern recognition and prediction. In: Honavar V, Uhr L (eds) Artificial intelligence and neural networks: Steps toward principled prediction. Academic, San Diego, pp 387–421

Carpenter GA, Grossberg S, Lesher GW (1998) The what-and-where filter: a spatial mapping neural network for object recognition and image understanding. Comp Vis Image Underst 69: 1–22

Chapman KL, Leonard LB, Mervis CG (1986) The effect of feedback on young children's inappropriate word usage. J Child Lang 13:101–107

Chey J, Grossberg S, Mingolla E (1997) Neural dynamics of motion grouping: from aperture ambiguity to object speed and direction. J Optical Soc Am 14:2570–2594

Clark EV (1973) What's in a word? On the child's acquisition of semantics in his first language. In: Morre TE (ed) Cognitive development and the acquisition of language. Academic, New York, pp 65–110

Cohen NJ, Squire LR (1980) Preserved learning and retention of a pattern-analyzing skill in amnesia: dissociation of knowing how and knowing that. Science 210:207–210

Cohen NJ (1984) Preserved learning capacity in amnesia: evidence for multiple memory systems. In: Squire L, Butters N (eds) The neuropsychology of memory. Guilford, New York, pp 83–103

Cohen MA, Grossberg S (1984) Neural dynamics of brightness perception: features, boundaries, diffusion, and resonance. Percept Psychophy 36:428–456

Deadwyler SA, West MO, Lynch G (1979) Activity of dentate granule cells during learning: differentiation of perforant path inputs. Brain Res 169:29–43
Deadwyler SA, West MO, Robinson JH (1981) Entorhinal and septal inputs differentially control sensory-evoked responses in the rat dentate gyrus. Science 211:1181–1183
Drevets WC, Burton H, Raichle ME (1995) Blood flow changes in human somatosensory cortex during anticipated stimulation. Nature 373:249
Eckhorn R, Bauer R, Jordan W, Brosch M, Kruse W, Munk M, Reitboeck HJ (1988) Coherent oscillations: a mechanism of feature linking in the visual cortex? Biol Cybernetics 60:121–130
Fiala JC, Grossberg S, Bullock D (1996) Metabotropic glutamate receptor activation in cerebellar Purkinje cells as substrate for adaptive timing of the classically conditioned eye-blink response. J Neurosci 16:3760–3774
Francis G, Grossberg S (1996a) Cortical dynamics of boundary segmentation and reset: persistence, afterimages, and residual traces. Perception 35:543–567
Francis G, Grossberg S (1996b) Cortical dynamics of form and motion integration: persistence, apparent motion, and illusory contours. Vision Res 36:149–173
Francis G, Grossberg S, Mingolla E (1994) Cortical dynamics of feature binding and reset: control of visual persistence. Vision Res 34:1089–1104
Fuster JM (1996) Frontal lobe and the cognitive foundation of behavioral action. In: Damasio AR, Damasio H, Christen Y (eds) The neurobiology of decision-making. Springer, New York, pp 47–61
Gaffan D (1985) Hippocampus: memory, habit, and voluntary movement. Philos Trans R Soc Lond B 308:87–99
Gaudiano P, Grossberg S (1991) Vector associative maps: unsupervised real-time error-based learning and control of movement trajectories. Neural Netw 4:147–183
Goodale MA, Milner D (1992) Separate visual pathways for perception and action. Trends Neurosci 15:20–25
Gove A, Grossberg S, Mingolla E (1995) Brightness perception, illusory contours, and corticogeniculate feedback. Visual Neurosci 12:1027–1052
Graf P, Squire LR, Mandler G (1984) The information that amnesic patients do not forget. J Exp Psychol: Learning Memory Cogn 10:164–178, 183
Gray CM, Singer W (1989) Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. Proc Natl Acad Sci 86:1698–1702
Grossberg S (1964) The theory of embedding fields with applications to psychology and neurophysiology. Rockefeller University, New York
Grossberg S (1969) On learning and energy–entropy dependence in recurrent and nonrecurrent signed networks. J Stat Phy 1:319–350
Grossberg S (1972) Neural expectation: cerebellar and retinal analogs of cells fired by learnable or unlearned pattern classes. Kybernetik 10:49–57
Grossberg S (1976a) Adaptive pattern classification and universal recoding. I. Parallel development and coding of neural feature detectors. Biol Cybernetics 23:121–134
Grossberg S (1976b) Adaptive pattern classification and universal recoding. II. Feedback, expectation, olfaction, and illusions. Biol Cybernetics 23:187–202
Grossberg S (1978) A theory of human memory: self-organization and performance of sensory-motor codes, maps, and plans. In: Rosen R, Snell F (eds) Progress in theoretical biology, vol 5. Academic, New York
Grossberg S (1980) How does a brain build a cognitive code? Psychol Rev 1:1–51
Grossberg S (1982) Studies of mind and brain: neural principles of learning, perception, development, cognition, and motor control. Kluwer, Norwell, MA
Grossberg S (1987a) Cortical dynamics of three-dimensional form, color, and brightness perception. II. Binocular theory. Percept Psychophy 41:117–158
Grossberg S (1987b) The adaptive brain, vol I. North-Holland, Amsterdam
Grossberg S (1988) Nonlinear neural networks: principles, mechanisms, and architectures. Neural Netw 1:17–61
Grossberg S (1994) 3-D vision and figure-ground separation by visual cortex. Percept Psychophy 55:48–120

Grossberg S (1997) Cortical dynamics of three-dimensional figureground perception of two-dimensional figures. Psychol Rev 104:618–658
Grossberg S (1999a) How does the cerebral cortex work? Learning, attention, and grouping by the laminar circuits of visual cortex. Spatial Vision 12:163–186
Grossberg S (1999b) Pitch based streaming in auditory perception. In: Griffith N, Todd P (eds) Musical networks: parallel distributed perception and performance. MIT, Cambridge, MA pp 117–140
Grossberg S, Mingolla E (1985a) Neural dynamics of form perception: boundary completion, illusory figures, and neon color spreading. Psychol Rev 92:173–211
Grossberg S, Mingolla E (1985b) Neural dynamics of perceptual grouping: textures, boundaries, and emergent segmentations. Percept Psychophys 38:141–171
Grossberg S, Gutowski W (1987) Neural dynamics of decision making under risk: affective balance and cognitive-emotional interactions. Psychol Rev 94:300–318
Grossberg S, Todorovic' D (1988) Neural dynamics of 1-D and 2-D brightness perception: a unified model of classical and recent phenomena. Percept Psychophy 43:241–277
Grossberg S, Merrill JWL (1996) The hippocampus and cerebellum in adaptively timed learning, recognition, and movement. J Cogn Neurosci 3:257–277
Grossberg S, Grunewald A (1997) Cortical synchronization and perceptual framing. J Cogn Neurosci 9:117–132
Grossberg S, McLoughlin NP (1997) Cortical dynamics of three-dimensional surface perception: binocular and half-occluded scenic images. Neural Netw 10:1583–1605
Grossberg S, Williamson JR (2001) A neural model of how horizontal and interlaminar connections of visual cortex develop into adult circuits that carry out perceptual grouping and learning. Cerebral cortex 11:37–58
Grossberg S, Guenther F, Bullock D, Greve D (1993) Neural representations for sensory-motor control, II: Learning a head-centered visuomotor representation of 3-D target position. Neural Netw 6:43–67
Grossberg S, Boardman I, Cohen MA (1997a) Neural dynamics of variable-rate speech categorization. J Exp Psychol: Hum Percept Perform 23:481–503
Grossberg S, Roberts K, Aguilar M, Bullock D (1997b) A neural model of multimodal adaptive saccadic eye movement control by superior colliculus. J Neurosci 17:9706–9725
Grossberg S, Govindarajan KK, Wyse LL, Cohen MA (2004) A neural network model of auditory scene analysis and source segregation. Neural Networks 17:511–536
Grunewald A, Grossberg S (1998) Self-Organization of binocular disparity tuning by reciprocal corticogeniculate interactions. J Cogn Neurosci 10:199–215
Guenther FH, Bullock D, Greve D, Grossberg S (1994) Neural representations for sensorymotor control. III. Learning a body-centered representation of 3-D target position. J Cogn Neurosci 6:341–358
Helmholtz HLF Von (1962) Treatise on physiological optics (Southall JPC translator). Dover, New York
Hikosaka O (1994) Role of basal ganglia in control of innate movements, learned behavior and cognition— a hypothesis. In: Percheron G, Mckenzie JS, Feger J (eds) The basal ganglia, vol IV. Plenum, New York, pp 589–595
Hubel DH, Wiesel TN (1977) Functional architecture of macaque monkey visual cortex. Proc R Soc Lond B 198:1–59
Kennedy JM (1979) Subjective contours, contrast, and assimilation. In Nodine CF, Fisher DF (eds) Perception and pictorial representation. Praeger, New York
Kennedy JM (1988) Line endings and subjective controls. Spatial Vision 3:151–158
Kohonen T (1984) Self-organization and associative memory. Springer, New York
Land E (1977) The retinex theory of color vision. Scientific Am 237:108–128
Land E (1986) Recent advances in retinex theory. Vision Res 26:7–21
Levy WB (1985) Associative changes at the synapse: LTP in the hippocampus. In: Levy WB, Anderson J, Lehmkuhle S (eds) Synaptic modification, neuron selectivity, and nervous system organization. Erlbaum, Hillsdale, NJ, pp 5–33

Levy WB, Desmond NL (1985) The rules of elemental synaptic plasticity. In: Levy WB, Anderson J, Lehmkuhle S (eds) Synaptic modification, neuron selectivity, and nervous system organization. Erlbaum, Hillsdale, NJ, pp 105–121

Livingstone MS, Hubel DH (1984) Anatomy and physiology of a color system in the primate visual cortex. J Neurosci 4:309–356

Logothetis NK, Leopold DA, Sheinberg DL (1996) What is rivalling during binocular rivalry? Nature 360:621–624

Lynch G, Mcgaugh JL, Weinberger NM (eds) (1984) Neurobiology of learning and memory. Guilford, New York

Michotte A, Thines G, Crabbe G (1964) Les complements amodaux des structures perceptives. Publications Universitaires de Louvain, Louvain

Middleton FA, Strick PL (1994) Anatomical evidence for cerebellar and basal ganglia involvement in higher cognitive function. Science 166:458–461

Miller EK, Li L, Desimone R (1991) A neural mechanism for working and recognition memory in inferior temporal cortex. Science 254:1377–1379

Mishkin M (1982) A memory system in the monkey. Philos Trans R Soc Lond B 298:85–95

Mishkin M (1993) Cerebral memory circuits. In: Poggio TA, Glaser DA (eds) Exploring brain functions: models in neuroscience. Wiley, New York, pp 113–125

Mishkin M, Ungerleider LG, Macko KA (1983) Object vision and spatial vision: two cortical pathways. Trends Neurosci 6:414–417

Murphy PC, Sillito AM (1987) Corticofugal feedback influences the generation of length tuning in the visual pathway. Nature 329:727–729

Parker DB (1982) Learning-logic (Invention Report 581–64, File 1) Office of Technology Licensing. October, Stanford University

Pribram KH (1986) The hippocampal system and recombinant processing. In: Isaacson RL, Pribram KH (eds) The hippocampus, vol 4. Plenum, New York, pp 329–370

Rauschecker JP, Singer W (1979) Changes in the circuitry of the kitten's visual cortex are gated by postsynaptic activity. Nature 280:58–60

Redies C, Crook JM, Creutzfeldt OD (1986) Neuronal responses to borders with and without luminance gradients in cat visual cortex and dLGN. Exp Brain Res 61:469–481

Repp BH (1980) A range-frequency effect on perception of silence in speech. Haskins Laboratories Status Report on Speech Research 61:151–165

Reynolds J, Nicholas J, Chelazzi L, Desimone R (1995) Spatial attention protects macaque V2 and V4 cells from the influence of nonattended stimuli. Soc Neurosci Abst 21(III):1759

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning internal representations by error propagation. In: Rumelhard DE, Mcclelland JL (eds) Parallel distributed processing. MIT, Cambridge, MA

Ryle G (1949) The concept of mind. Hutchinson Press

Sakai K, Hikosaka O, Miyauchi S, Takino R, Sasaki Y, Putz B (1998) Transition of brain activation from frontal to parietal areas in visuomotor sequence learning. J Neurosci 18:1827–1840

Samuel AG (1981a) The role of bottom-up confirmation in the phonemic restoration illusion. J Exp Psychol: Hum Percept Perform 7:1124–1131

Samuel AG (1981b) Phonemic restoration: insights from a new methodology. J Exp Psychol: General 110:474–494

Schiller PH (1994) Area V4 of the primate visual cortex. Curr Dir Psychol Sci 3:89–92

Scoville WB, Milner B (1957) Loss of recent memory after bilateral hippocampal lesion. J Neurol Neurosurg Psychiatry 20:11–21

Sillito AM, Jones HE, Gerstein GL, West DC (1994) Feature-linked synchronization of thalamic relay cell firing induced by feedback from the visual cortex. Nature 369:479–482

Singer W (1983) Neuronal activity as a shaping factor in the self-organization of neuron assemblies. In: Basar E, Flohr H, Haken H, Mandell AJ (eds) Synergetics of the brain. Springer, New York

Smith C, Carey S, Wiser M (1985) On differentiation: a case study of the development of the concept of size, weight, and density. Cognition 21:177–237

Smith LB, Kemler DG (1978) Levels of experienced dimensionality in children and adults. Cogn Psychol 10:502–532

Squire LR, Butters N (eds) (1984) Neuropsychology of memory. Guilford, New York

Squire LR, Cohen NJ (1984) Human memory and amnesia. In: Lunch G, Mcgaugh J, Weinberger NM (eds) Neurobiology of learning and memory. Guilford, New York, pp 3–64

Ungerleider LG, Mishkin M (1982) Two cortical visual systems: separation of appearance and location of objects. In: Ingle DL, Goodale MA, Mansfield RJW (eds) Analysis of visual behavior. MIT, Cambridge, MA, pp 549–586

von der Malsburg C (1973) Self-organization of orientation sensitive cells in the striate cortex. Kybernetik 14:85–100

Ward TB (1983) Response tempo and separable-integral responding: evidence for an integral-toseparable processing sequencing in visual perception. J Exp Psychol: Hum Percept Perform 9:1029–1051

Warren RM (1984) Perceptual restoration of obliterated sounds. Psychol Bull 96:371–383

Warren RM, Sherman GL (1974) Phonemic restorations based on subsequent context. Percept Psychophys 16:150–156

Warrington EK, Weiskrantz L (1974) The effect of prior learning on subsequent retention in amnesic patients. Neuropsychology 12:419–428

Werbos P (1974) Beyond regression: New tools for prediction and analysis in the behavioral sciences. Unpublished doctoral thesis, Harvard University, Cambridge, MA

Willshaw DJ, Malsburg C Von Der (1976) How patterned neural connections can be set up by self-organization. Proc R Soc Lond B 194:431–445

Zola-Morgan SM, Squire LR (1990) The primate hippocampal formation: evidence for a timelimited role in memory storage. Science 250:288–290

Chapter 2
Emergence of Intentional Procedures in Self-Organizing Neural Networks

Henri Atlan and Yoram Louzoun

We have used a neural network formalism in order to analyze under which conditions a positive answer could be given to the following question: can neural networks self-organize so that not only structures and functions not explicitly programmed emerge from their dynamics, but also goals for intentional actions, set up and achieved by themselves?

Such mechanistic models of intentional self-organization are useful in that they allow to circumvent the usual circular explanation of intentionality by causal effects of assumed intentional mental states on bodily movements.

From a mathematical and modeling point of view, we have presented a simulation model for the analysis of intentionality through the study of intentional actions (Louzoun and Atlan 2007). We limit ourselves in this paper to the cognitive interpretation and the philosophical analysis of the obtained results. Intentionality in the psycholinguistic sense of "meaning" – where there is no "goal" except for the content of a thought in an internal deliberation of a sentence *meant* to say something – is left outside the scope of this work. We have limited ourselves to intentionality in a pragmatic sense as it is observed in intentional actions to solve two problems of causality: the apparent time inversion involved in final causes and the "*mind–body*" causal relationship involved in the usual picture of a mental state being the case of bodily movements and actions.

The system we have developed is designed to devise new goals by itself and to reach these goals. The goals are determined by the capacity of a network to learn a relation between effects and the events that caused them. The model is a metaphor for the psychophysical goal learning process in cognitive beings. This process involves the ability to predict rapidly the result of a set of events, so that an initial

H. Atlan (✉)
Human Biology Research Center, Hadassah University Hospital, Ein Karem, Jerusalem, Israel
and
Ecole des Hautes Etudes en Sciences Sociales, 54 bd Raspail, 75270 Paris, France
e-mail: atlan@ehess.fr

Y. Louzoun
Math. Department, Bar Ilan University, Ramat Gan, Israel, 52900
e-mail: louzouy@math.biu.ac.il

A. Carsetti (ed.), *Causality, Meaningful Complexity and Embodied Cognition*,
Theory and Decision Library A 46, DOI 10.1007/978-90-481-3529-5_2,

event is reproduced knowing its expected result. In other words, prediction (which is knowledge) and intentional action are closely related. That is why this capacity is modeled using a non-supervised learning network associated to a recurrent neural network. However, while the prediction capacity is obviously based on memory of previous experience, this knowledge must be allowed some degrees of freedom, which produce new predictions of new events and the achievement of new goals. In our model, this capacity is simply the result of network dynamics where closely related but different states are associated in basins of attraction.

To summarize, the recurrent network represents a mechanistic causal process that develops from a random initial state to a steady state. There is no trivial relation between the steady state and the initial state of the network. The very indirect relation between initial and steady state represents the complexity of the causal relation in real environments.

The feed-forward learning network creates a link between final and initial states, allowing the time inversion occurring in goal-directed action. The input to the network is the steady state of the recurrent network, and the output of the feed-forward network is an initial state of the recurrent network, which is equivalent to a dynamic memory.

The selection mechanism chooses which final states are defined as goals, and works like a non-programmed satisfaction function, emerging from the partially random history of the system in its environment.

Our model is obviously not directly related to mental processes in its details. It only represents a plausibility analysis to show that the self emergence of meaningful *actions* is possible and can be explained by a relatively simple mechanism. The model allows us to study, which mechanisms are essential to have such a representation. The same question can also be addressed from the point of view of Spinozist monism as will be further discussed. The combination of a simple model and Spinoza's propositions enable us to provide plausible answers to shed new light on experimental results and propose ways to treat some of the most basic questions in cognition.

2.1 Minimal Necessary Requirements

Goals emerge in our simulation from a combination of four elements: A seemingly random process relating the initial and final states (which is actually a deterministic process too complex to be directly deduced from the initial and final states), a limited memory capable of remembering the relation between some initial and final states, a learning algorithm that invents a systematic relation between final and initial states, and an evolving set of required final states selected semi-randomly according to the frequencies of their appearance. Note that goals would not emerge in the absence of any of these elements. Thus, we think that our networks represent a minimal structure where such goals can be obtained. Obviously one can extensively alter the details and even completely replace the mechanistic aspect of each component. However the same general elements must prevail in order for goals to emerge.

- The first element required is an indirect dynamical link between initial and final states. Learning the relation between a state and its direct result is not defined as goal emergence. We define a learnt goal as a relation between an initial state and a final state that cannot be directly guessed from the initial state. Another aspect of the required dynamics is a difference between the probabilities to reach different final states. If all final states are reached with equal probabilities, the goal emerging would only be a mirror of the network history and would not represent an inherent property of the network.
- Memory is obviously required; it actually is the most important element of the network. The seemingly minor role of remembering a relation during the learning process is actually essential. In the absence of memory, the network would not be able to retrieve an initial state from a final state. This "time inversion" is the element giving the network a future prediction capacity. In other words, the network is able to predict the future in certain conditions, since it has seen similar evolutions and has learnt (either "erroneously" or "correctly") a relation between an initial state and the final state it led to. A similar conclusion can be drawn for human behavior. Humans predict the future, since they have seen similar evolutions in the past and have learnt (either erroneously or correctly) a relation between a situation and its results.
- The learning algorithm is required since the capacity to attain goals depends on the ability to find a "simple" rule relating some of the initial states to the appropriate final states. Again one can infer from the network to human behavior, one can predict the future, only in cases similar to past events. These past events and their results were learnt and a time inversion mechanism is used to relate new situations to their future.
- Finally, the evolving set of goals allowing for both stability and newness is required in order to distinguish between goals that can and cannot be learnt, and goals for which no simple rule can be obtained. A specific aspect of the goals that we have requested in the current application is stability (i.e. we required that goals should change slowly compared to the network dynamics). This request is not essential. One could imagine rapidly changing goals (e.g. the mind of a small child). However, most aspects of human behavior are based on a set of relatively long term goals. The emergence of these "long term goals" is equivalent to the emergence of stable goals in the current application. This element is thus not required for the goal emergence per-se, but it adds an aspect of *meaning* to the goals. In addition, the possibility of newness is embedded into the role of small random variations in the definition of goals.

2.2 Externally Versus Internally Defined Goals

In the current application, we minimized the model and merged together two different tasks. The memory device which allows for goal directed action and the learning device which allows for goal definition by the system itself are merged into the operation of the feed-forward, perceptron-like, network.

Of course, the two different tasks performed by the feed-forward network can be separated, especially if the model is designed in a more trivial fashion to achieve predefined goals, assigned from outside the system. Contrary to a goal defined from outside, a self-generated, internally defined goal is not a goal because it has some inherent value from the beginning. It is a goal because it represents a properly learnt and stable link between initial and final states. A set of such goals is an emerging and stable property of the network's structure and the history it underwent. Dependence on history represents how the system adapts itself and generates new goals accordingly. On the other hand, externally predefined goals can be learnt more simply. Each of them must be coded into an attractor state of the recurrent network; and then kept in memory as one of the final states to be eventually retrieved with an initial state leading to it from its basin of attraction. (It is clear that only attractor states of the recurrent network can be established as goals, either by external imposition or by non-supervised learning, since a state can be stored as a goal only if the system can reach it with a high enough probability.)

2.2.1 At the Beginning

For example, one could consider that the initiation of the learning process needs not start from scratch, as in the present model. Before learning, a basic set of goals may have been stored as an initial set of "instinctual" goals with which to start. This may be the result, in the real world, of *long term evolutionary processes*, which may be simulated, for example, by genetic algorithms driven by selection for survival. Such processes must be distinguished from the mechanisms of setting oneself cognitive goals that is studied in the present work. Such a priori goals may produce built in, basic drives to start with, like biochemical signals for hunger, sexual drive, tissue damage repair and so on. These signals would affect only the initial set up of goals, but not the general mechanism of goal development. One can even set a permanent "vital" set of goals selected through a long evolutionary process. These goals can be hard wired not to change. Another possibility would be that some goals have an inherent higher score than others. We have tested models to include such initial goals or preferred goals, and the subsequent picture emerging from these models is similar to what we currently report.

According to our model, intention and action appear to be one and the same realization, simply represented in different ways. This implies that an intention to act is always normally associated with its execution. In other words, both the action and the intention are represented by links between initial and final states. The difference between the action and the intention is actually the difference between an action actually performed and its initiation, as indicated by neurophysiological data discussed further. This difference results in our capacity to stop an action once initiated. We would call an action interrupted after being initiated, an intention to do an action and invent a mental state to represent it. This view is opposed to the usual mentalist assumption that an intention exists first in the mind as a "pure" mental

state, possibly, but not normally associated to its execution. In our model, as in the work of Anscombe (1957), intentions are not defined as pure subjective states of the mind, but as properties of some sets of actions, which make them intentional and different from non-intentional ones. The fact that a subjective intention to act may not be followed by its execution is not to be seen as the normal flow. Rather, it must be related to an external obstacle to the execution or to any other kind of superimposed inhibition preventing the iterative process to reach completion.

One does not need to invent intentional mental states as causes of teleological actions. This is of course in contrast with common sense or folk psychology based on our initial insight of the causal relation between will and action. However, neurophysiology data on voluntary movements contradict this commonly accepted picture as well and support our model in showing that the conscious will to trigger an action does not necessarily precede the action.

2.3 Neurophysiology of Voluntary Movements

Following observations by Benjamin Libet and his co-workers (Libet et al. 1983; Libet 1985, 1992), recently confirmed and expanded (Haggard and Eimer 1999; Haggard et al. 2002), spontaneous short-term conscious decision to act with no preplanning does not precede but follows by approximately 300 ms the *initiation* of movement, as measured by the Readiness Potential cortical activity. Thus, initiation of a voluntary action is triggered by some unconscious activity, and the following awareness is interpreted as its cause. When asked about the timing of their decision, subjects perceive it, by antedating, before the initiation of the action. However, the motor activity itself follows by 150–200 ms the conscious decision to act, which means that a conscious "veto" is possible, as an inhibition of the movement after its inhibition.

Most of the controversy around this work was triggered by the difficulties to reconcile these data with the traditional Cartesian concept of free will and to integrate these data within the commonly accepted mentalist causal theories of action. The model presented in our work contributes to make these data intelligible within an alternative monist theory of action. Mentalist theories of action, based on the idea that mental representations described as subjective states of the mind, can cause objective brain states able to trigger physical movements, were extensively analyzed and criticized already in 1957 in a philosophical and psychological context (Anscombe 1957). This criticism, as well as our model, contradicts our common sense representation of free will as a direct cause of voluntary actions. However, the general question of free will as an illusion or a reality remains open, because the model, as do Libet's data, allows believers in free will to relate it in an indirect way, to a possibility of vetoing a movement after its initiation, rather than to the initiation itself.

Antedating the conscious decision to act may be thought of as a temporal illusion (analogous to a spatial visual illusion), with a possible adaptive value whereby

voluntary actions are linked to our memory-based capacity of prediction and self-awareness (e.g. (Llinas 2001)). As in our model, inhibition of movement completion after initiation explains intentional action with no execution. However, this does not necessary infer that the problem of free will is solved in one way or another: if one can relate it to vetoing the execution of a movement, one cannot exclude, on the other hand, that vetoing itself would be caused by a non-conscious event, in spite of our spontaneous subjective conscious experience.

Thus, the question of whether free will is an illusion or not is definitely left outside the scope of this study. Similarly, long-term deliberation leading to intentions to do something in principle with no specific timing for the actual decision to act, are left outside Libet's observation. In the experimental setting, the patients were asked to perform some movement and to decide upon the timing. It is clear that their very participation in the experiment indicates their agreement and intention to do it before their decision.

2.4 Philosophical Interpretation

One feature of the views presented here is the monist ontology involved in the approach to the mind–body problem. Spinozist philosophy is certainly the most radical monist attitude towards this problem. This is apparent, for example, in propositions such as

> "*Body cannot determine mind to think, neither can mind determine body to motion or rest or any state different from these, if such there be*" (The Ethics, III, 2), where Spinoza denies the possibility of causal relationships between the mind and the body, not because they would pertain to two different substances, as in Descartes, but precisely because they are "*one and the same thing, though expressed in two ways*" (Ibid. II, 7, note).

The analysis of some aspects of this psycho-physical monism will help to better understand the philosophical counterintuitive implications of our model, as well as of the neurophysiological data on voluntary movements briefly reported in the previous section.

Let us first note that this Spinozist denial of a causal relationship between mind and body states, just mentioned, implies that the cause of a voluntary bodily movement must always be some previous bodily (brain) event or set of events, and not a conscious decision viewed as a mental event as described by subjective reports about conscious experiences. The difference from a non-voluntary movement is the nature and degree of conscious experience accompanying it. But in any case, a conscious mental event in this context may accompany the brain event *but not be its cause, being in fact identical with it,* although not describable in the same language. Results from neurophysiology support this view: unconscious initiation of voluntary action precedes the conscious decision to trigger the movement. Thus, our model may provide Spinozist monism, however counterintuitive, with some theoretical and philosophical interpretation.

This kind of counterintuitive identity between different properties or events, identical but not describable by synonymous enunciations, was called a "synthetic

identity of properties" (Putnam 1981), to be distinguished from the usual analytical identity, where synonymous descriptions can replace one another. Hilary Putnam found an example of synthetic identity in the notion of physical magnitudes, which we employ in physics, such as "temperature" and "mean molecular kinetic energy" being synthetically, but not analytically, identical. In the same context, Putnam explicitly related the Spinozist psycho-physical identity to such a synthetic identity, as a way to overcome many well known difficulties in understanding this approach to the mind–body problem (see also (Atlan 1998a)).

Similar results on affects and emotions, indicating a lack of causality between body and mind, have been proposed by A. Damasio, with the same reference to Spinozist monism as its philosophical interpretation (Damasio 2003).

This stance, as well as the elaborated Wittgensteinian view of intentional descriptions (Wittgenstein 1953), has been neglected by most philosophers and cognitive scientists, mostly because it contradicts our common sense experiences and the commonly accepted ethical implications of free will which go with them. Thus, under the influence of mentalist theories in psychology (for analysis and criticism see e.g. Anscombe 1957; Davidson 1970; Fodor 1981; Shanon 1993; Chalmers 1995), intentions are viewed as some kinds of conscious mental states, able to cause bodily movements whenever an intentional action is executed. These theories raise several difficult questions, such as:

1. How can a mental state be the cause of a physical movement?
2. More generally, what is the conscious intentional experience made of?

The first question has been addressed, more or less successfully, by several philosophers. Among them, Donald Davidson's theory of action may be the most comprehensive (Davidson 1970, 1999), especially in view of his definite monist attitude, which he explicitly relates to *The Ethics* of Spinoza. However, his willingness to stick to common sense conscious subjective and ethical experiences does not allow him to overcome serious difficulties in trying to reconcile the Spinozist explicit denial of causal relationship between subjective states of mind as such and objective bodily movements, with his "anomalous monism" (Davidson 1991; Atlan 1998a).

The second question covers several problems related with different aspects of what we call consciousness. According to David Chalmers (1995), some of these problems are "easy", although not trivial: they deal with specific cognitive aspects of consciousness, related with objective mechanisms accounting for cognitive properties, such as memory, learning, adaptation, etc. However, what he calls the "hard problem" is the "question of how physical processes in the brain give rise to subjective experience". This question is the same in the opposite direction as that of intentional actions, where subjective intentions are supposed to cause physical movements.

In our work, we depart from mentalist causal theories of action and we try to come back to a more objective approach to the question of causality (Atlan 1998b). The model presented here exhibits one of the main features outlined by Anscombe in order to circumscribe the logical difficulties of these theories, namely the approach of intentionality through the study of intentional *actions*. As mentioned above, this implies that intentions and actions are not dissociated to start with, and that the

normal state of affairs is the execution of the intention. Such a dissociation, which may occur when an intention is not accompanied by an action, is the result of an obstacle or inhibition of the execution.

In this view, the "hard problem" of causality between the mental and the physical is eliminated: there is no causal relationship between an intention as a mental state and action as a bodily movement, because "roughly speaking, a man intends to do what he does" (Anscombe 1957). Because this view seems counter-intuitive and raises new questions, following the quest initiated by Wittgenstein about the status of intentional statements language games, Anscombe feels compelled to add: "But of course that is *very* roughly speaking. It is right to formulate it, however, as an antidote against the absurd thesis which is sometimes maintained: that a man's intended action is only described by describing his *objective*". In many instances the objective of the agent is a description after the fact, aiming at answering the question: "Why did you do it?".

Let us conclude with several features of the non-mentalist model of intentions presented in this work, which appear almost literally in Spinoza's writings, at the point that one could speak of a "Spinozist neurophysiology".

1. Decision to act and previous knowledge allowing prediction are two different aspects of the same process associated with voluntary actions, although the former seems directed towards the future and the latter towards the past. That is the case because intentions are described by means of intentional *actions* and not of intentional mental states as causes of the actions. "*Will and understanding are one and the same*" ((Spinoza 1677), II, 49, corollary) seems to be an abrupt statement of this counterintuitive concept.
2. In our model, general sets of goals are memorized from learning by experience. The aquired knowledge results from the interaction between the internal structure of the network and the history of its most frequent encounters with classes of stimuli from its environment.

In the context of the classical controversy about the reality of "Universals", we read:

> ... these general notions (called Universals) are not formed by all men in the same way, but vary in each individual according as the point varies, whereby the body has been most frequently affected and which the mind most easily imagines or remembers. For instance, those who have most often regarded with admiration the stature of man, will by the name of man understand an animal of erect stature; those who have been accustiomed to regard some other attribute, will form a different general image of man, for instance, that man is a laughing animal, a two-footed animal without feathers, a rational animal, and thus, in other cases, everyone will form general images of things according to the habit (disposition) of his body ((Spinoza 1677), II, 40, note).

Thus, this "disposition of the body" is made by the way the cognitive system (mind–body) is assembled and also by the way it has been most frequently affected.

3. According to the neurophysiological data on voluntary movements reported before, as well as in our model, voluntary action is triggered by some unconscious stimulus, accompanied but not caused by a conscious state of the mind. A conscious observation with an understanding of our action accompanies that action

but is not its cause. And we can interpret it as a decision of our will which determines the action, because we do not know the unconscious events in our body which are the real causes.

> Now all these things clearly show that the decision of the mind and the desire or decision of the body are simultaneous in nature, or rather one and the same thing, which when considered under the attribute of Thought and explained through the same we call a decision, and when considered under the attribute of Extension and deduced from the laws of motion and rest we call determination ((Spinoza 1677), III, 2, note).

4. As noted in Libet's observations there is a slight delay between the triggering of action and our being conscious of it, because consciousness and understanding take time: as in our model, they need to be retrieved from *memory*. In other words,

> we can do nothing by a decision of the mind unless we recollect having done so before ((Spinoza 1677), III, 2, note).

5. In the stance adopted here, we obviously *lose* something, namely common sense about free will and causation of actions by decisions of a non-bodily mind. However, we *gain* understanding of intentional actions without resorting to hidden causal properties of mental states. Let us note that the reality of free will is not necessarily denied, although its content is modified. According to Libet, it can be located in a kind of *veto* function, i.e. a possible inhibition of movement after it has been initiated. In addition, nothing is said here about the possible effects of long term deliberations and decisions to act "in principle", with a more or less extended period of time until the decision is made to start the action. Spinoza's stance about free will is more radical:

> ...men think themselves free on account of this alone, that they are conscious of their actions and ignorant of the causes of them; and, moreover, that the decisions of the mind are nothing save their desires, which are accordingly various according to various dispositions of their and other interacting bodies ((Spinoza 1677), note on proposition III, 2, mentioned above).

6. At last, the picture of intentional actions presented in this work helps to better understand what "desire" in the practical syllogism is about[1]: an unconscious drive with awareness of the goal which one is driven to.

[1] Let us recall the classical description of intentional actions by the practical syllogysm:

- Agent A desires to be in state S.
- A knows or believe that C is a cause for S.
- Therefore A performs C.

This description assumes intentional mental states from the beginning, such as desire, knowledge, belief. In our model, knowledge or belief are just retrieved memories of previous causal events. In addition, as Elizabeth Anscombe rightfully noticed, the first proposition of the syllogism may be conflated with the third. Contrary to the usual demonstrative syllogism (Men are mortal, Socrates is a man, etc.), the first proposition here does not add information: it is contained in the "therefore" of the third proposition. Our model may be seen as a computer simulation of this modified syllogism, where intentional mental states causing intentional actions and different from them are not needed.

This definition of desire has been extended further by Spinoza to the realm of moral judgements:

> Desire is appetite with consciousness thereof. It is thus plain from what has been said, that in nocase do we strive for, wish for, long for, or desire anything, because we deem it to be good, but on the other hand we deem a thing to be good, because we strive for it, wish for it, long for ir, or desire it ((Spinoza 1677), III, 9, note).

References

Anscombe GEM (1957) Intention. Basic Blackwell, London
Atlan H (1998a) Immanent causality: a spinozist viewpoint on evolution and theory of action. In: Vijver GVd (ed) Evolutionary systems. Kluwer, The Netherlands, pp 215–231
Atlan H (1998b) Intentional self-organization. Emergence and reduction. Towards a physical theory of intentionality, Thesis Eleven 52:5–34
Chalmers DJ (1995) The puzzle of conscious experience. Scientific Am 12:62–68
Damasio A (2003) Looking for Spinoza. Joy, sorrow, and the feeling brain. Harvest Books, Harvest edition, Washington
Davidson D (1970) Mental events experience and theory. In: Foster L, Swanson J (eds) University of Massachusetts, Amherst, pp 79–81; reprinted in Davidson D (1980), Essays on actions and events. Oxford University Press, New York
Davidson D (1991, 1999). Spinoza's causal theory of the affects. In: Yovel Y (ed) Ethica III. Desire and affect. Spinoza as psychologist. Little Room, New York, pp 95–111
Fodor JA (1981) Representations. MIT, Cambridge, MA
Haggard P, Eimer M (1999) On the relation between brain potentials and the awareness of voluntary movements. Exp Brain Res 126:128–133
Haggard P, Clark S, Kalogeras J (2002) Voluntary action and conscious awareness. Nat Neurosci 5(4):382–385
Libet B (1985) Unconscious cerebral initiative and the role of conscious will in voluntary action. Behav Brain Sci 8:529–566
Libet B (1992) Models of conscious timing and the experimental evidence (Commentary/Dennett and Kinsbourne: Time and the Observer). Behav Brain Sci 15:213–215
Libet B, Gleason CA, Wright EW, Pearl DK (1983) Time of conscious intention to act in relation to onset of cerebral activity (readiness potential): the unconscious initiation of a freely voluntary act. Brain 106:623–642
Llinas RR (2001) I of the vortex: from neurons to self. MIT, Cambridge, MA
Louzoun Y, Atlan H (2007) The emergence of goals in a self-organizing network: a non-mentalist model of intentional actions. Neural Networks 20:156–171
Putnam H (1981) Reason, truth and history. Cambridge University Press, Cambridge
Shanon B (1993) The representational and the presentational. Harvester Wheatsheaf, Simon and Schuster, New York
Spinoza B (1677) The Ethics (engl. translation H.M. Elwes 1955). Dover, New York

Chapter 3
Action Goal Representation and Action Understanding in the Cerebral Cortex

Leonardo Fogassi

3.1 Introductory Remarks

Classically it has been assumed that the sector of frontal cortex devoted to motor control has the main aim of coding movements, that is the parameters necessary to accomplish joints displacements such as amplitude and direction. More recently, another view has been proposed, that maintains that the main function of the motor cortex is that of coding goal directed actions (see Rizzolatti et al. 2000, 2004). According to this view, motor neurons would define a limited number of motor goals, instead of computing an exponential number of movements. Once these goals are coded, their implementation, including movement parameters, would be performed by areas with more executive functions, such as MI. The neurophysiological data of the last twenty years show that the goal interpretation can be applied to many areas of the motor cortex. Actions are indeed coded in several premotor areas and, more extensively, in the parieto-frontal circuits linking specific premotor and parietal areas (see Rizzolatti et al. 1988, 1998, 2000, see also below). In particular, several studies carried out in the ventral premotor cortex of the macaque monkey (areas F4 and F5) showed that area F4 code axial and proximal actions toward three-dimensional objects in space (Gentilucci et al. 1988; Fogassi et al. 1996), while area F5 code different types of hand and mouth actions (Gentilucci et al. 1988; Rizzolatti et al. 1988; Ferrari et al. 2003).

3.2 Perception and Action Are Strictly Inter-Related

In the classical view of the motor cortex, and in general of cortical functioning, it was maintained that the flow of information was directed from parietal to frontal

L. Fogassi (✉)
Department of Neuroscience and Department of Psychology, University of Parma,
V.Volturno 39, B.go Carissimi 10, 43100 Parma
e-mail: leonardo.fogassi@unipr.it

A. Carsetti (ed.), *Causality, Meaningful Complexity and Embodied Cognition*,
Theory and Decision Library A 46, DOI 10.1007/978-90-481-3529-5_3,

cortex, where the parietal cortex represented the higher stage of sensory elaboration, the outcome of which (the percept) was considered to be fed to motor cortex that, passively receiving this outcome, would have used it in order to execute movements in relation to sensory stimuli (see Goodale and Milner 1992). The discover that motor cortex contains a mosaic areas, many of which coding goal-directed actions, and the demonstration that each of these areas is reciprocally connected to a specific area of the parietal cortex, challenged this concept. There is now good consensus that different parieto-premotor circuits are involved in specific types of sensorimotor transformations, each dependent on a specific effector (arm, hand, eye) (Rizzolatti et al. 1998; Rizzolatti and Luppino 2001; Rizzolatti and Matelli 2003). The reciprocity of the connections between the parietal and the frontal area constituting each circuit is against a rigid separation between perceptual and motor properties, but rather points to a sharing of these functions inside each circuit. Moreover, as it will be discussed in the next chapters, these circuits are not only used for sensorimotor transformations, but provide the basis for the emergence of different types of cognitive functions, such as space perception, action understanding, coding of intention.

In the next chapter I will show how actions are coded by single neurons in ventral premotor and inferior parietal cortex of the monkey and how action coding can be exploited for understanding actions made by others. Then I will present evidence that also in humans there are circuits for action coding and understanding similar to those found in monkeys. Finally, I will suggest, based on experimental evidence, that these cortical circuits and the mechanisms conveyed by the neurons belonging to them provide the basis for higher cognitive functions such as imitation and intention understanding.

3.3 A Vocabulary of Actions in Ventral Premotor Area F5

Area F5 is located in the rostral part of ventral premotor cortex. This area, anatomically identified by means of the cytochrome oxidase staining technique (Matelli et al. 1985), when microstimulated evokes hand and mouth movements (Gentilucci et al. 1988) Single neurons recording experiments demonstrated that neurons in this area discharge when a monkey executes goal-related hand and mouth actions (Rizzolatti et al. 1988) such as grasping, manipulating, holding, tearing objects. Most of them discharge during grasping actions. Some of them discharge, for example, when the monkey grasps food with the hand or with the mouth, thus coding the action “grasp” in an abstract way, independently of the effector used for executing that action. Other F5 motor neurons code actions in a more specific way discharging, for example, when the monkey grasps a small object using a precision grip and not when it grasps food using a whole hand prehension.

Beyond purely motor neurons, which constitute overall the majority of all F5 neurons, area F5 contains also two categories of visuomotor neurons, the motor properties of which are indistinguishable from those of the former category.

However, their visual properties are peculiar. The first category of visuomotor neurons is formed by neurons responding to the presentation of objects of particular size, shape and orientation. The size or the shape of the object effective in triggering the neurons discharge is very often congruent with the specific type of action they code (Rizzolatti et al. 1988). These neurons were named "canonical" neurons (Rizzolatti and Fadiga 1998; Rizzolatti et al. 2000, 2004). Recently we investigated in detail the properties of these neurons (Murata et al. 1997; Raos et al. 2006), by using a behavioral task in which the monkey was presented with several objects of different shape and size, and it had to observe the object and, after a variable delay, to grasp it (Movement in light condition). In another condition of the same task the monkey had only to observe the presented object, without grasping it (Object fixation condition). The results confirmed that the motor properties of purely motor and canonical neurons are specific for the type of grip used for grasping the different objects. During object observation performed in the "Movement in light" condition and in "Object fixation" condition canonical neurons presented a visual response, the specificity of which was congruent with that found during movement execution. Very interestingly, a cluster analysis revealed that the visual responses of canonical neurons grouped according to the type of prehension, and not according to the objects visual properties. For example, a typical cluster was constituted by cone, cube and sphere that, although different in their shape, were grasped in the same way. This suggests that the visual response of canonical neurons can be better interpreted as a kind of motor representation of the object. In other words, while the pictorial object description necessary for recognizing and discriminating objects is represented in the inferior temporal cortex, the premotor cortex codes a pragmatic object description. This object motor representation can be used for executing the type of grip necessary to grasp it or can remain in the status of a potential motor act, thus subserving a more cognitive function, that of object knowledge.

The second category of F5 visuomotor neurons is constituted by "mirror" neurons, that will be described in a later section.

3.4 Goal Representation in the Inferior Parietal Cortex

The inferior parietal cortex has been traditionally considered as an association cortex, in which polymodality would subserve cognitive functions, such as, for example, space coding. However, the single neuron recording studies of Mountcastle et al. (1975) and Hyvarinen et al. (1982) showed that the inferior parietal cortex is endowed also with the fundamental property of coding eye, hand and arm movements. These pioneering studies were confirmed by other more recent studies showing that areas belonging to these regions are crucial for sensorimotor transformation for visually guided hand actions (see for example Taira et al. 1990; Murata et al. 2000) and for visually guided eye movements (see for example Andersen et al. 1990). Furthermore, lesion and inactivation studies clearly demon-

strated the occurrence of specific motor deficits after damage to posterior parietal cortex, including misreaching, ocular dysmetria and disruption of hand shaping in monkeys (Faugier-Grimaud et al. 1978; Gallese et al. 1994; Li et al. 1999; see Hyvarinen et al. 1982). Note that also lesions of posterior parietal cortex in humans determine motor deficits, such as optic ataxia, impairment in grasping movements, apraxia, motor neglect and directional hypokinesia in humans (Perenin and Vighetto 1988; Binkofski et al. 1998; see De Renzi and Faglioni 1999).

Recently Rozzi et al. (2008) explored the functional properties of the inferior parietal lobule (IPL), and found that a large percent of neurons respond during execution of motor acts. The representation of these acts in the lobule follows a gross somatotopy, with the mouth motor field located rostrally, the hand and arm motor fields in an intermediate position and, finally, the eye field located caudally. Very interestingly, the responses of most IPL motor neurons code the goal of the motor acts and not specific movement parameters. For example, there are neurons responding when the monkey grasps a piece of food with the hand or with the mouth; other neurons respond during reaching for grasping, but not during reaching for moving away, although a similar arm extension is performed in both conditions.

3.5 Motor Organization in the Inferior Parietal Lobule

Before describing the motor organization found in IPL, it is necessary to define more strictly what is a motor act and what is an action. By motor acts we mean movements that have a goal, but whose goal is only partial (e.g. grasping a piece of food). By motor action we mean a series of motor acts that, as their final outcome, lead to a reward (e.g., eating a piece of food after reaching it, grasping it and bringing it to the mouth).

Recently, we recorded motor neurons from the hand representation of IPL in order to assess whether grasping neurons were equally active when grasping is part of different actions leading to different goals (Fogassi et al. 2005). The recorded neurons were tested while the monkey performed a task involving two main conditions. In one, the monkey reached and grasped a piece of food located in front of it and brought it to its mouth. In the other, the monkey reached and grasped an object and placed it into a container. In the second condition the experimenter gave the monkey a reward if it performed correctly the task. Note that in this task the first motor act of both conditions is the same (grasping).

The results showed that the majority of grasping neurons discharged differently according to the intended goal of the action in which grasping was embedded. Neurons coding grasping for eating discharged strongly when grasping preceded bringing to the mouth than when it preceded placing in the container. Neurons coding grasping for placing showed the opposite behavior. These data suggest that: (a) the IPL contains pre-wired or learned chains of motor neurons, each coding a specific final goal; (b) the discharge of IPL grasping neurons reflects the intention of the performing agent.

This organization of IPL appears to be appropriate for providing fluidity in action execution. Each neuron codes a specific motor act, but at the same time (being embedded into a specific action) is linked, and possibly facilitates, the next motor act according to the action goal. In addition, this motor organization includes also the concept of intention. This does not mean that intention is *directly* coded by IPL motor neurons, because other areas, for example prefrontal cortex, could have a major role in this function.

3.6 Mirror Neurons

Mirror neurons constitute the second category of F5 visuomotor neurons. They discharge when a monkey observes another individual (a human being or another monkey) performing a hand action in front of it (here the term "action" will be used in a more general sense, not as strictly defined in the previous chapter) Differently from canonical neurons, they do not discharge to the simple presentation of food or of other interesting objects. They also do not discharge, or discharge much less, when the observed action is mimicked without the target object. The response is generally weaker or absent when the effective action is executed by using a tool instead of the hand. Thus, the only visual stimulus effective in triggering mirror neurons response is a hand-object interaction (Gallese et al. 1996; Rizzolatti et al. 1996a).

By using the coded observed action as a classification criterion, it appears that mirror neurons code actions that generally coincide with or are very similar to those "motorically" coded by F5 motor neurons, e.g. grasping, manipulating, tearing, holding objects. More than half of F5 mirror neurons responds to the observation of only one action, while the remaining neurons respond to the observation of two or more actions. Among neurons responding to the observation of grasping action (by far the most effective in driving the visual response of mirror neurons) there are some very specific, since they code also the *type* of observed grip. Thus, mirror neurons can present different types of visual selectivity: selectivity for the observed action, and selectivity for the way in which the observed action is performed.

3.6.1 Mouth Mirror Neurons

After the discovery of mirror neurons that discharge to the observation and execution of hand actions, another category of mirror neurons, activated by the observation and execution of mouth actions were described ("mouth mirror neurons", Ferrari et al. 2003). Most mouth mirror neurons respond to observation of ingestive actions such as biting, tearing with the teeth, sucking, licking, etc., showing a high specificity, similarly to hand mirror neurons. They do not respond to simple object

presentation or to mouth mimed actions. When the congruence between the visual and the motor response is analysed, most of these neurons (about 90%) show a very good congruence. A smaller but significant percent of mouth mirror neurons respond specifically to the observation of mouth communicative actions belonging to the monkey repertoire, such as lips-smacking, lips protrusion or tongue protrusion. Mouth mirror neurons of this sub-category do not respond, or respond very weakly, to the observation of ingestive actions.

3.6.2 Motor Properties of F5 Mirror Neurons

The most important property of mirror neurons is that their "visual" responses are matched, at the single neuron level, with motor responses which, as emphasized above, are virtually indistinguishable from that of F5 purely motor or canonical neurons.

Most mirror neurons show a good congruence between visual and motor responses. However there are two major categories: "strictly congruent" neurons, in which observed and executed actions coincide (about 30% of all F5 mirror neurons), and "broadly congruent" neurons, in which the coded observed action and the coded executed action are similar but not identical (60%). In some cases the congruence could be defined according to a logical or "causal" sense: for example a neuron could respond when the monkey observed an experimenter placing a piece of food on a tray and when the monkey grasped the same piece of food. The two actions can be considered to be part of a logical sequence.

The congruence found between the visual and motor responses of mirror neurons suggests that every time an action is observed, there is an activation of the motor circuits of the observer coding a similar action. Strictly congruent mirror neurons could be more involved in a detailed analysis of the observed action. These neurons could be suitable for imitation (see below). In contrast, broadly congruent neurons could have the capacity to generalize across different ways of achieving the same goal, thus probably enabling a more abstract type of action coding. Moreover, these neurons could be very important for appropriately react within a social environment and for communicating, by responding with gestures to other individuals gestures. In fact, these neurons "recognize" one or more observed actions, and produce an output that can be ethologically related to them.

3.7 Mirror Neurons and Action Understanding

An important property of mirror neurons is that when the agent mimics the action in absence of the target, the response of mirror neurons is much weaker or absent. We can suppose that as monkeys do not act in absence of a target, they do not interpret observed mimicking as a goal-directed action. These observations suggest that

mirror neurons may play a crucial role in *understanding the goal of another individual action*. This understanding occurs because of the activation, in the observer, of his/her own motor representation of the goal.

In everyday life we can understand the goal of an action made by another person also when visual information about the observed action is incomplete, for example when part of the observed action occurs out of sight. A series of experiments was recently carried out to address the issue of whether mirror neurons become active also during the observation of partially hidden actions (Umiltà et al. 2001). The experiments consisted of two basic experimental conditions. In one the monkey observed a fully visible action directed toward an object ("Full vision" condition). In the other it observed the same action, but its final crucial part (hand–object interaction) was hidden behind an occluding screen ("Hidden" condition). Note that in this condition the monkey knew that an object was present behind the screen. In two control conditions ("Mimicking in Full vision", and "Hidden mimicking") the same action was mimed without object, both in full vision and behind the occluding screen. Note that in 'Hidden mimicking' condition the monkey knew that there was no object behind the screen.

The results showed that the majority of tested F5 mirror neurons responded to the observation of hand actions even when the final part of the action, i.e. the part triggering the response in full vision, was hidden from the monkey's vision. However, when the hidden action was mimed, with no object present behind the occluding screen, there was no response. It appears therefore that the mirror neurons responsive in the Hidden condition are able to generate a motor representation of an observed action, not only when the monkey sees that action, but also when it "knows" its outcome without seeing its most crucial part (i.e. hand–object interaction). These results corroborate the hypothesis, previously suggested, that the mirror neurons mechanism is at the basis of action understanding (Gallese et al. 1996; Rizzolatti et al. 1996, 2004).

Another demonstration of the involvement of the mirror neuron system in action understanding is represented by the presence of another category of mirror neurons, named audio-visual mirror neurons. These neurons become active when monkeys not only observe, but also hear the sound of an action (Kohler et al. 2002). The response of these neurons are specific for the type of action seen and heard. For example, they respond to peanut breaking when the action is only observed, only heard or both heard and observed, and do not respond to the vision and sound of another action, or to unspecific sounds. Note that often the neuron discharge to the simultaneous presentation of both the visual and the acoustic inputs is higher than the response to either of the inputs, when presented alone. These data show that the acoustic input has access to the motor cortex of a listener allowing him to retrieve the action representation present in this area, thus accessing to the action content. This content is coded by audiovisual mirror neurons, independently of whether these actions are observed, listened or executed. Interestingly enough, the capacity of representing action content independently of the modality used to access this content is typical of language.

3.8 The Mirror Neuron Circuit

Mirror neurons are endowed with both visual and motor properties. What is the origin of their visual input? Data from Perrett and coworkers (Perrett et al. 1989, 1990) show that in the anterior part of the superior temporal sulcus (STSa) there are neurons responding to the sight of biological motion, that is to the observation of other individuals' movements, performed with different body parts, such as head, legs, body. One category of these neurons is specific for the observation of hand-object interactions but, differently from mirror neurons, apparently do not discharge when the monkey executes the same hand actions. These neurons could provide the visual information to the cortical circuit involved in matching action observation with action execution. STSa has no direct connections with ventral premotor cortex, where area F5 is located. Thus, a functional connection between STSa and F5 could be possibly established only indirectly by means of two pathways: one throughout the prefrontal cortex, the other through the inferior parietal lobule, since STSa is connected with both these cortical regions (Cavada and Goldman-Rakic 1989; Seltzer and Pandya 1994; Rozzi et al. 2006). Of these two pathways the first one seems the most unlike, since the connections between area F5 and prefrontal cortex are present but very weak (Matelli et al. 1986). In contrast, the connections between the inferior parietal lobule and ventral premotor cortex are very strong (Matsumura and Kubota 1979; Muakkassa and Strick 1979; Petrides and Pandya 1994; Matelli et al. 1986; Cavada and Goldman-Rakic 1989; Rozzi et al. 2006).

On the basis of this anatomical evidence, we recently re-investigated the properties of the inferior parietal lobule (IPL), looking for the possible presence of mirror properties. First of all we could confirm that in this area, in particular in its rostral half, there are neurons responding to visual or somatosensory stimulation or both (see Hyvärinen 1981; Graziano and Gross 1995). As described in a previous chapter, we also found many neurons responding during arm, hand and mouth actions (Gallese et al. 2002; Rozzi et al. 2008). In addition, we found, very likely in area PFG, also neurons responding to the sight of hand-object interactions (Fogassi et al. 1998, 2005; Gallese et al. 2002; Rozzi et al. 2008). Of them, 70% had also motor properties, being activated when the monkey performed mouth or hand actions or both. These neurons were called "parietal mirror neurons" (Gallese et al. 2002; Fogassi et al. 2005).

Parietal mirror neurons, similarly to F5 mirror neurons, respond to the observation of several types of single or combined actions. Grasping action, alone or in combination with other actions, is the most represented one. Differently from F5 mirror neurons, a high number of parietal mirror neurons are activated by the observations of two hands interacting with an object. Parietal mirror neurons responded during the execution of hand, mouth, or hand and mouth actions and the vast majority of them present either a strict or a broad congruence between observed and executed action, accordingly to the same criterion used for analyzing the congruence of F5 mirror neurons. Among broadly congruent neurons, a significant number entered in the category "logically related", i.e. neurons discharging during the execution of an action that could be seen as the logical prolongation of

the effective observed one. For example, the effective observed action could be placing a piece of food on a tray, while the effective executed action could be grasping the piece of food. The possible role of these neurons will be discussed below, in the section concerning intention understanding.

Finally, in IPL there are neurons that respond only to the observation of hand actions, but are devoid of motor properties. These neurons, similar to those described in STSa, confirm that parietal cortex could be the link necessary for matching observed and executed actions.

Summing up, the presence of mirror neurons in both parietal and ventral premotor cortex strongly suggests that the mirror neuron system is formed through the anatomical temporo-parieto-premotor circuit. As it will be described below, a similar circuit is also present in humans.

3.9 The Mirror System in Humans

The first evidence that a mirror system exists also in humans was provided by a transcranial magnetic stimulation experiment (Fadiga et al. 1995), showing that in subjects observing hand actions made by an experimenter there was an enhancement of motor evoked potentials in those muscles that subjects normally used to execute the observed actions. Subsequent TMS, electroencephalographic (EEG) and magnetoencephalographic (MEG) investigations confirmed this finding (Hari et al. 1998; Cochin et al. 1999; Strafella and Paus 2000; Nishitani and Hari 2000, 2002; Gangitano et al. 2004). Brain imaging experiments (Rizzolatti et al. 1996b; Grafton et al. 1996; Grèzes et al. 1998, 2003; Iacoboni et al. 1999, 2001; Buccino et al. 2001; Koski et al. 2002, 2003; Manthey et al. 2003; Johnson-Frey et al. 2003) showed that action observation activates a temporo-parieto-frontal circuit, namely the STS region, the inferior parietal lobule and the lower part of the precentral gyrus (ventral premotor cortex) plus the posterior part of the inferior frontal gyrus (IFG) The parietal and frontal regions form the core of the mirror neuron system in humans. It is important to note that the activation of IFG involves the Broca's region, previously conceived as a "speech" area. This region becomes active not only during action observation, but also during the execution of hand-related tasks (Parsons et al. 1995; Grafton et al. 1996; Binkofski et al. 1999; Iacoboni et al. 1999; Buccino et al. 2004a). The posterior part of it, Brodmann's area 44, is considered to be homologue of monkey's F5 (see Rizzolatti and Arbib 1998).

3.10 Possible Functions Derived from the Mirror Neuron System: Imitation, Language, Intention Understanding

The properties of mirror neurons reported above show that they have a crucial role in action understanding. The next important question to address is whether they can also constitute the neural circuit for other functions. The functions most

likely related to the mirror system are imitation, intention understanding, communication/language. Some of these functions, such as imitation and language, are exclusively or mainly present in humans, others appears already in monkeys. I will examine briefly the involvement of the mirror system in all these functions.

3.11 Imitation

Imitation is the first function that comes to mind when one thinks to the possible use of mirror neurons, because they possess the property enabling the observer to immediately translate the visual information on observed action into the motor parameters necessary for reproducing it. However, in monkeys, the capacity to imitate is weak or even absent. Thus, mirror neurons cannot be primarily tools for imitation, although they could represent the building blocks on which imitation takes place in humans. Indeed, experiments in humans confirm this suggestion.

In an fMRI experiment, Iacoboni et al. (1999) showed that in volunteers required to observe and imitate a finger lifting, there was an activation of the left inferior frontal gyrus (IFG) during observation and, more strongly, during imitation. The importance of IFG for imitation was also shown by Nishitani and Hari (2000) using the event-related neuromagnetic (MEG) technique. In these experiments individuals were asked to repeat highly practiced actions done by another individual. Buccino et al. (2004b) recently addressed the issue of which cortical areas become active when individuals are required to learn, on the basis of action observation, a *novel motor pattern*. Naive participants were required to imitate guitar chords played by an expert guitarist. By using an event-related fMRI paradigm, cortical activations were mapped during the following events: (a) observation of the chords made by the expert player, (b) pause, (c) execution of the observed chords, and (d) rest. Control conditions involved pure observation and non imitative motor activity. The results showed that during observation for imitation there was activation of the inferior parietal lobule and the dorsal part of ventral premotor cortex plus the *pars opercularis* of IFG. It is interesting to note that during the pause in imitation condition, when subjects are preparing a program to reproduce the observed chord, there was a strong activation of the middle frontal cortex (area 46). It has been hypothesized that the role of this area is that of re-combining the motor representations corresponding to the different motor acts, in order to fit the observed model.

Summing up, imitation in humans appears to require the involvement of the mirror neuron circuit, with the additional activation of prefrontal areas when recombination of already existing motor representation in novel sequences is required.

It remains to be explained how imitation in monkeys is minimal, inspite of the presence of a well-developed mirror neuron system? There are probably many reasons for this apparent contradiction. First, in monkeys a lower percent of mirror neurons show a strict congruence between observed and executed action, the majority coding the action goal. Second, as shown above, in humans a crucial role in imitation learning is played by the prefrontal cortex, a region that is much more developed in the human brain in respect to that of monkeys.

3.12 A Pathway from Monkey F5 to Human Broca's for Language Evolution

The mirror neuron mechanism appears to be very close to the mechanism that, during inter-individual communication, enables the listener/observer to understand the meaning of the message emitted by the sender. The central point here is that both sender and receiver share the same motor programs necessary to produce a message and the pathway that allows to access these programs. The proposed homology between F5 and Broca's area is in favor of the idea that language can be derived from a system involved in action and, lately, in gesture understanding. This homology is based on several data. (1) Cytoarchitectonically, both area 44 (part of Broca's area) and area F5 are dysgranular (see Petrides and Pandya 1994; Rizzolatti and Arbib 1998; Nelissen et al. 2005). (2) Both Broca's area and F5 have a mouth and hand representation. Many brain imaging experiments demonstrated that Broca's area, beyond its classical role in speech production, is also involved in hand movement tasks. For example it is activated by the execution of hand movements, mental imagery of grasping actions, hand mental rotation and imitation tasks (Parsons et al. 1995; Grafton et al. 1996; Iacoboni et al. 1999; Buccino et al. 2004). (3) Area F5 is endowed with a system for the control of laryngeal muscles and of orofacial synergisms (Hast et al. 1974). (4) Both area F5 and Broca's area are activated during observation of hand and mouth actions (see for refs. Rizzolatti et al. 2001; Rizzolatti et al. 2004). In particular, recent fMRI experiments demonstrated that the inferior frontal gyrus is activated both when subjects observe biting action (Buccino et al. 2001) and when they observe other individuals performing silent speech (Campbell et al. 2001; Calvert and Campbell 2003; Buccino et al. 2004a), in agreement with the presence in F5 of mouth mirror neurons for ingestive and communicative actions (Ferrari et al. 2003). (5) Both F5 and Broca's area are reached by an acoustic input related to action semantic content. As described above, there are in F5 mirror neurons responding both to the sight and the sound of actions. In humans it has been recently demonstrated that (a) listening to sentences related to actions made with different effectors activate Broca's area and premotor cortex (Tettamanti et al. 2005) and (b) listening to words and pseudo-words containing a consonant requiring a marked tongue muscles involvement to be pronounced determines a significant increase of the amplitude of motor evoked potentials (MEPs) recorded from the tongue muscles with respect to listening to words and pseudo-words containing consonants not requiring such tongue involvement (Fadiga et al. 2002).

All these data corroborate the idea that an ancient observation/execution matching system, as that found in monkeys, may have paved the way to the evolution of human language. This process occurred through many steps, two of which, however, are assumed to be very important (see Rizzolatti and Arbib 1998; Fogassi and Ferrari 2004; Arbib 2005). The first is the transition from a motor system coding actions to one with the capacity to encode also intransitive actions, probably through a process of ritualization of goal-directed actions (Van Hoof 1967). This transition, probably in its primitive form, could be found in communicative mirror neurons, that respond to observation of communicative actions and during execution not only

of the same actions, but also of ingestive actions. The second is represented by the association between a gesture and a sound. The possibility to use facial and brachiomanual gestures in association with utterances provides a higher combinatorial power, allowing to create a richer vocabulary. The presence in monkey area F5 of a large population of neurons coding both hand and mouth actions and its access to auditory input could have been important elements, in evolution, for facilitating the occurrence of the proposed association gesture/action-sound.

3.13 Intention Understanding

When we observe somebody else performing goal-directed action, in most cases we are able to infer his/her intended goal, even though the action is not completely accomplished. In other words we have the capacity to understand the intention of other individuals. Since mirror neurons provide a mechanism to understand the goal of motor acts performed by others, it is natural to raise the issue of whether they can also play a role in intention detection. In a recent experiment, the visual response of *parietal* mirror neurons was studied in the same conditions, described in a previous section, that were used for studying motor properties of IPL grasping neurons. Briefly, in one condition the experimenter grasped a piece of food and brought it to the mouth, in the other he grasped the same piece of food and placed it into a container. Mirror neuron activity was recorded while the monkey observed the two conditions. The crucial part of the activity was that related to observation of grasping.

The results showed that the majority of IPL mirror neurons were differently activated when the observed grasping motor act was followed by bringing to the mouth or by placing. The remaining mirror neurons did not show any selectivity.

A characterizing property of all mirror neurons is the congruence of their motor and visual responses. The data just described show a further level of congruence. Mirror neurons that discharged more intensely during grasping for eating than during grasping for placing discharged more intensely also during the observation of grasping for eating. Conversely, neurons selective for grasping to place discharged strongly during the observation of this motor act.

Thus, IPL mirror neurons, in addition to recognizing the observed motor act, are able to discriminate among identical motor acts according to the context in which it is executed. Because the discriminated motor acts are part of chains, each of which leading to a specific final goal, this capacity allows the monkey to predict what is the goal of the observed action and, in this way, to "read" the intention of the acting individual. If grasping neurons belonging to a particular chain fire, the observed acting individual is going to bring the food to the mouth; if, in contrast, another set of grasping neurons belonging to another chain fire, the observed acting individual is going to put the food away.

The selection of a particular group of grasping mirror neurons may be determined by many factors. One of these is the type of object grasped. The sight of food very likely would trigger neurons coding grasping for eating than grasping for other

purposes. However, if food is near a container, this different context could trigger neurons coding grasping for placing.

Another factor that can be very important for discriminating between different intentions is the previous action made by the observed agent. For example, if the trials are run in blocks, very soon the monkey can guess that the next trial, very likely, will be the same as the previous one, so the neuron discharge during grasping observation will reflect the higher probability of occurrence of an action instead of the other one.

In agreement with the monkey data suggesting that the mirror system can provide the mechanism for intention understanding, a recent fMRI study in humans (Iacoboni et al. 2005) indicates that also our species uses the mirror neuron system in order to understand the intention of others. In this study subjects had to observe hand actions performed in two different contexts The results showed that hand actions performed in contexts, compared with other two control conditions (actions without context or context only), produced a higher activation of the inferior frontal gyrus.

Summing up, the mirror neuron system in monkeys provides the first neural substrate for a primitive understanding of others' intentions, that probably paved the way for the evolution of the more sophisticated aspects of mind reading present in humans. Probably many of these aspects still rely on the automatic activation of the parieto-frontal mirror neuron circuit.

3.14 Conclusions

In this article it has been shown that a vocabulary of actions coded in the motor system form the core of an internal, "first person" knowledge on the top of which many cognitive functions, such as action understanding, intention understanding, imitation and language can be built. The possibility to show in humans mechanisms similar to those studied in details in monkeys will allow, in the future, to assess more in depth which cognitive functions can emerge from the motor circuits and which other cortical and subcortical structures can be added to these basic substrates, in order to have a more complete picture of the anatomo-functional network involved in each function. Language is a good example of one of these complex functions that probably originated from motor cortical structures.

References

Andersen RA, Bracewell RM, Barash S, Gnadt JW, Fogassi L (1990) Eye position effects on visual, memory and saccade-related activity in area LIP and 7a of macaque. J Neurosci 10:1176–1196

Arbib M A (2005) From monkey-like action recognition to human language: an evolutionary framework for neurolinguistics. Behav Brain Sci 28:105–124

Binkofski F, Dohle C, Posse S, Stephan K, Hefter H, Seitz R, Freund H (1998) Human anterior intraparietal area subserves prehension: a combined lesion and functional MRI activation study. Neurology 50:1253–1259

Binkofski F, Buccino G, Stephan KM, Rizzolatti G, Seitz RJ, Freund H-J (1999) A parieto-premotor network for object manipulation: evidence from neuroimaging. Exp Brain Res 128:210–213

Buccino G, Binkofski F, Fink GR, Fadiga L, Fogassi L, Gallese V, Seitz RJ, Zilles K, Rizzolatti G, Freund H-J (2001) Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. European J Neurosci 13:400–404

Buccino G, Lui F, Canessa N, Patteri I, Lagravinese G, Benuzzi F et al. (2004a) Neural circuits involved in the recognition of actions performed by nonconspecifics: an FMRI study. J Cogn Neurosci 16:114–126

Buccino G, Vogt S, Ritzl A, Fink GR, Zilles K, Freund HJ, Rizzolatti G (2004b) Neural circuits underlying imitation of hand actions: an event related fMRI study. Neuron 42:323–334

Calvert GA, Campbell R (2003) Reading speech from still and moving faces: Neural substrates of visibile speech. J Cogn Neurosci 15:57–70

Campbell R, MacSweeney M, Surguladze S, Calvert GA, Mc Guire P, Suckling J, Brammer MJ, David AS (2001) Cortical substrates for the perception of face actions: An fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). Brain Research Cognitive Brain Research, 12:233–243

Cavada C, Goldman-Rakic PS (1989) Posterior parietal cortex in rhesus monkey: I. Parcellation of areas based on distinctive limbic and sensory corticocortical connections. J Comp Neurol 287:393–421

Cochin S, Barthelemy C, Roux S, Martineau J (1999) Observation and execution of movement: similarities demonstrated by quantified electroencephalograpy. Eur J Neurosci 11:1839–1842

De Renzi E, Faglioni P (1999) Apraxia. In: Denes G, Pizzamiglio L (eds) Clinical and experimental neuropsychology. Psychology Press, Hove, UK

Fadiga L, Fogassi L, Pavesi G, Rizzolatti G (1995) Motor facilitation during action observation: A magnetic stimulation study. J Neurophysiol 73:2608–2611

Fadiga L, Craighero L, Buccino G, Rizzolatti G (2002) Speech listening specifically modulates the excitability of tongue muscles: a TMS study. Eur J Neurosci 15:399–402

Faugier-Grimaud S, Frenois C, Stein D (1978) Effects of posterior parietal lesions on visually guided behavior in monkeys. Neuropsychologia 16:151–168

Ferrari PF, Gallese V, Rizzolatti G, Fogassi L (2003) Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex. Eur J Neurosci 17:1703–1714

Fogassi L, Ferrari PF (2004) Mirror neurons, gestures and language evolution. Interact Stud 5:345–363

Fogassi L, Gallese V, Fadiga L, Luppino G, Matelli M, Rizzolatti G (1996) Coding of peripersonal space in inferior premotor cortex (area F4). J Neurophysiol 76:141–157

Fogassi L, Gallese V, Fadiga L, Rizzolatti G (1998) Neurons responding to the sight of goal-directed hand/arm actions in the parietal area PF (7b) of the macaque monkey. Soc Neurosci Abst 275:5

Fogassi L, Ferrari PF, Gesierich B, Rozzi S, Chersi F, Rizzolatti G (2005) Parietal lobe: from action organization to intention understanding. Science 308:662–667

Gallese V, Murata A, Kaseda M, Niki N, Sakata H (1994) Deficit of hand preshaping after muscimol injection in monkey parietal cortex. Neuroreport 5:1525–1529

Gallese V, Fadiga L, Fogassi L, Rizzolatti G (1996) Action recognition in the premotor cortex. Brain 119:593–609

Gallese V, Fadiga L, Fogassi L, Rizzolatti G (2002) Action representation and the inferior parietal lobule. In: Prinz W, Hommel B (eds) Common mechanisms in perception and action: attention and performance, vol 19. Oxford University Press, Oxford, pp 334–355

Gangitano M, Mottaghy FM, Pascual-Leone A (2004) Modulation of premotor mirror neuron activity during observation of unpredictable grasping movements. Eur J Neurosci 20:2193–2202

Gentilucci M, Fogassi L, Luppino G, Matelli M, Camarda R, Rizzolatti G (1988) Functional organization of inferior area 6 in the macaque monkey: I. Somatotopy and the control of proximal movements. Exp. Brain Res 71:475–490

Goodale MA, Milner (1992) Separate visual pathways for perception and action. Trends Neurosci 15:20–25

Grafton ST, Arbib MA, Fadiga L, Rizzolatti G (1996) Localization of grasp representations in humans by PET: 2. Observation compared with imagination. Exp Brain Res 112:103–111

Graziano MSA, Gross CG (1995) The representation of extrapersonal space: a possible role for bimodal visual-tactile neurons. In: Gazzaniga MS (ed) The cognitive neurosciences. MIT, Cambridge, MA, pp 1021–1034

Grèzes J, Costes N, Decety J (1998) Top-down effect of strategy on the perception of human biological motion: a PET investigation. Cogn Neuropsychol 15:553–582

Grèzes J, Armony JL, Rowe J, Passingham RE (2003) Activations related to "mirror" and "canonical" neurones in the human brain: an fMRI study. Neuroimage 18:928–937

Hari R, Forss N, Avikainen S, Kirveskari S, Salenius S, Rizzolatti G (1998) Activation of human primary motor cortex during action observation: a neuromagnetic study. Proc Natl Acad Sci USA 95:15061–15065

Hast MH, Fischer JM, Wetzel AB, Thompson VE (1974) Cortical motor representation of the laryngeal muscles in Macaca mulatta. Brain Res 73:229–240

Hyvärinen J (1981) Regional distribution of functions in parietal association area 7 of the monkey. Brain Res 206:287–303

Hyvärinen J (1982) Posterior parietal lobe of the primate brain. Physiol Rev 62:1060

Iacoboni M, Woods RP, Brass M, Bekkering H, Mazziotta JC, Rizzolatti G (1999) Cortical mechanisms of human imitation. Science 286:2526–2528

Iacoboni M, Koski LM, Brass M, Bekkering H, Woods RP, Dubeau MC, Mazziotta JC, Rizzolatti G (2001) Reafferent copies of imitated actions in the right superior temporal cortex. Proc Natl Acad Sci USA 98:13995–13999

Iacoboni M, Molnar-Szakacs I, Gallese V, Buccino G, Mazziotta JC, Rizzolatti G (2005) Grasping the intentions of others with one's own mirror neuron system. PloS Biology 3:529–535

Johnson-Frey SH, Maloof FR, Newman-Norlund R, Farrer C, Inati S, Grafton ST (2003) Actions or hand-objects interactions? Human inferior frontal cortex and action observation. Neuron 39:1053–1058

Kohler E, Keysers C, Umiltà MA, Fogassi L, Gallese V, Rizzolatti G (2002) Hearing sounds, understanding actions: action representation in mirror neurons. Science 297:846–848

Koski L, Wohlschlager A, Bekkering H, Woods RP, Dubeau MC (2002) Modulation of motor and premotor activity during imitation of target-directed actions. Cerebral Cortex 12:847–855

Koski L, Iacoboni M, Dubeau MC, Woods RP, Mazziotta JC (2003) Modulation of cortical activity during different imitative behaviors. J Neurophysiol 89:460–471

Li C-SR, Mazzoni P, Andersen RA (1999) Effect of reversible inactivation of macaque lateral intraparietal area on visual and memory saccades. J Neurophysiol 81:1827–1838

Manthey S, Schubotz RI, von Cramon DY (2003) Premotor cortex in observing erroneous action: an fMRI study. Brain Res Cogn Brain Res 15:296–307

Matelli M, Camarda R, Glickstein M, Rizzolatti G (1986) Afferent and efferent projections of the inferior area 6 in the macaque monkey. J Comp Neurol 251:281–298

Matelli M, Luppino G, Rizzolatti G (1985) Patterns of cytochrome oxidase activity in the frontal agranular cortex of the macaque monkey. Behav Brain Res 18:125–137

Matsumura M, Kubota K (1979) Cortical projection of hand-arm motor area from post-arcuate area in macaque monkeys: a histological study of retrograde transport of horseradish peroxidase. Neurosci Lett 11:241–246

Muakkassa KF, Strick PL (1979) Frontal lobe inputs to primate motor cortex: evidence for four somatotopically organized 'premotor' areas. Brain Res 177:176–182

Mountcastle VB, Lynch JCGA, Sakata H, Acuna C (1975) Posterior parietal association cortex of the monkey: command functions for operations within extrapersonal space. J Neurophysiol 38:871–908

Murata A, Fadiga L, Fogassi L, Gallese V, Raos V and Rizzolatti G (1997) Object representation in the ventral premotor cortex (area F5) of the monkey. J Neurophysiol 78:2226–2230

Murata A, Gallese V, Luppino G, Kaseda M, Sakata H (2000) Selectivity for the shape, size, and orientation of objects for grasping in neurons of monkey parietal area AIP. J Neurophysiol 83:2580–2601

Nelissen K, Luppino G, Vanduffel W, Rizzolatti G, Orban G (2005) Observing others: multiple action representation in the frontal lobe. Science 310:332–336

Nishitani N, Hari R (2000) Temporal dynamics of cortical representation for action. Proc Natl Acad Sci USA 97:913–918

Nishitani N, Hari R (2002) Viewing lip forms: cortical dynamics. Neuron 36:1211–1220

Parsons LM, Fox PT, Hunter Down J, Glass T, Hirsch TB, Martin CC, Jerabek PA, Lancaster JL (1995) Use of implicit motor imagery for visual shape discrimination as revealed by PET. Nature 375:54–58

Perenin M-T, Vighetto A (1988) Optic ataxia: a specific disruption in visuomotor mechanisms. Brain 111:643–674

Perrett, DI, Harries MH, Bevan R, Thomas S, Benson PJ, Mistlin AJ, Chitty AK, Hietanen JK, Ortega JE (1989) Frameworks of analysis for the neural representation of animate objects and actions. J Exp Biol 146:87–113

Perrett DI, Mistlin AJ, Harries MH, Chitty AJ (1990) Understanding the visual appearance and consequence of hand actions. In: Goodale MA (ed) Vision and action: the control of grasping. Ablex, Norwood, NJ, pp 163–342

Petrides M, Pandya DN (1994). Projections to the frontal cortex from the posterior parietal region in the rhesus monkey. J Comp Neurol 228:105–116

Petrides M, Pandya DN (1997). Comparative architectonic analysis of the human and the macaque frontal cortex. In Boller F, Grafman J (eds) Handbook of neuropsychology, vol IX. Elsevier, New York, pp 17–58

Raos V, Umiltá MA, Murata A, Fogassi L, Gallese V (2006). Functional properties of grasping-related neurons in the ventral premotor area F5 of the macaque monkey. J Neurophysiol 95:709–729

Rizzolatti G, Arbib MA (1998) Language within our grasp. Trends Neurosci 21:188–194

Rizzolatti G, Fadiga L (1998) Grasping objects and grasping action meanings: the dual role of monkey rostroventral premotor cortex (area F5). Novartis Foundation Symp 218:81–95

Rizzolatti G, Luppino G (2001) The cortical motor system. Neuron 31:889–901

Rizzolatti G, Matelli M (2003) Two different streams form the dorsal visual system: anatomy and functions. Exp Brain Res 153:146–157

Rizzolatti G, Camarda R, Fogassi L, Gentilucci M, Luppino G, Matelli M (1988) Functional organization of inferior area 6 in the macaque monkey: II. Area F5 and the control of distal movements. Exp Brain Res 71:491–507

Rizzolatti G, Fadiga L, Fogassi L, Gallese V (1996a) Premotor cortex and the recognition of motor actions. Cogn Brain Res 3:131–141

Rizzolatti G, Fadiga L, Matelli M et al. (1996b) Localization of grasp representation in humans by PET: 1. Observation versus execution. Exp Brain Res 111:246–252

Rizzolatti G, Luppino G, Matelli M (1998) The organization of the cortical motor system: new concepts. Electroencephalogr Clin Neurophysiol 106:283–296

Rizzolatti G, Fogassi L, Gallese V (2000) Cortical mechanisms subserving object grasping and action recognition: a new view on the cortical motor function. In: Gazzaniga MS (ed) The new cognitive neurosciences, 2nd edn. MIT, Cambridge, USA, pp 539–552

Rizzolatti G, Fogassi L, Gallese V (2001) Neurophysiological mechanisms underlying the understanding and imitation of action. Nat Rev Neurosci 2:661–670

Rizzolatti G, Fogassi L and Gallese V (2004) Cortical mechanism subserving object grasping, action understanding and imitation. In: Gazzaniga MS (ed) The cognitive neuroscience, 3rd edn. A Bradford Book. MIT, Cambridge, MA, pp 427–440

Rozzi S, Calzavara R, Belmalih A, Borra E, Gregoriou GG, Matelli M, Luppino G (2006) Cortical connections of the inferior parietal cortical convexity of the macaque monkey. Cerebral Cortex 16:1389–1417 Epub. 2005

Rozzi S, Ferrari PF, Bonini L, Rizzolatti G, Fogassi L. (2008) Functional organization of inferior parietal lobule convexity in the macaque monkey: Electrophysiological characterization of motor, sensory and mirror responses and their correlation with cytoarchitectonic areas. Eur J Neurosci 28:1569–88

Seltzer B, Pandya DN (1994) Parietal, temporal, and occipital projections to cortex of the superior temporal sulcus in the rhesus monkey: a retrograde tracer study. J Comp Neurol 343:445–63

Strafella AP, Paus T (2000) Modulation of cortical excitability during action observation: a transcranial magnetic stimulation study. NeuroReport 11:2289–2292

Taira M, Mine S, Georgopoulos AP, Murata A and Sakata H (1990) Parietal cortex neurons of the monkey related to the visual guidance of hand movement. Exp Brain Res 83:29–36

Tettamanti M, Buccino G, Saccuman MC, Gallese V, Danna M, Scifo P, Fazio F, Rizzolatti G, Cappa S, Perani D (2005) Listening to action-related sentences activates fronto-parietal motor circuits. J Cogn Neurosci 17:273–281

Umiltà MA, Kohler E, Gallese V, Fogassi L, Fadiga L, Keysers C, Rizzolatti G (2001). I know what you are doing. a neurophysiological study. Neuron 31:155–65

Van Hoof Jaram (1967) The facial displays of the catarrhine monkeys and apes. In: Morris D (ed) Primate ethology. Weidenfield and Nicolson, London, pp 7–68

Part II
Truth, Randomness and Impredicativity

Chapter 4
The Genesis of Mathematical Objects, Following Weyl and Brouwer

Dirk van Dalen

Almost a century ago, Brouwer launched his first intuitionistic programme for mathematics. He did so in his dissertation of 1907, where he formulated the basic act of creation of mathematical objects, known as the *ur-intuition* of mathematics. Mathematics, in Brouwer's view, was an intellectual activity of men (of the *subject*), independent of language and logic. The objects of mathematics come first in the process of human cognition, and description and systematization (in particular logic) follow later. The formulation of the ur-intuition is somewhat hermetic, but in view of its fundamental role, let us reproduce it here.

> Ur-intuition of mathematics (and every intellectual activity) as the substratum, divested of all quality, of any perception of change, a unity of continuity and discreteness, a possibility of thinking together several entities, connected by a 'between' that by the interpolation of new entities never gets exhausted.

As we see, Brouwer sees the ur-intuition as the genesis of both the discrete part of mathematics, let us say, the natural numbers, and of the continuous part, i.e., the continuum. Neither of these can be reduced to the other.

A more refined analysis was given in the Vienna lectures (although it is foreshadowed in the so-called 'rejected parts' of the thesis), where the notion of the falling apart of a moment of life is introduced. In the final presentation, *Consciousness, Philosophy and Mathematics* (CPM) [Brouwer 1949a], this phenomenon is described as the *move of time*: 'By a move of time a present sensation gives way to another present sensation in such a way that consciousness retains the former one as a past sensation and moreover, through this distinction between present and past, recedes from both and from stillness and becomes mind.' Thus the subject has created a 'twoity' of a past and present sensation. The process evidently can be iterated, and complexes and strings of sensation become the object of attention.The sensation

D. van Dalen (✉)
Department of Philosophy, Utrecht University, Heidelberglaan 8, P.O. Box 80126, 3508 TC Utrecht The Netherlands
e-mail: dirk@dalenwolwever11.demon.nl

A. Carsetti (ed.), *Causality, Meaningful Complexity and Embodied Cognition*, Theory and Decision Library A 46, DOI 10.1007/978-90-481-3529-5_4,

complexes form a bewildering mixture, in which a certain order is introduced by the *causal attention*. This carries out a process of *identification*. One may think of the identification of 'similar' complexes, or of *abstraction*.

In CPM the notion of causal sequence is further refined: 'An iterative complex of sensations whose elements have an invariable order of succession in time, whilst if one of its elements occurs, all following elements are expected to occur likewise, in the right order of succession, is called a causal sequence'. It might be tempting to explain these, let us say 'strongly causal sequences', *scs*, by a causality, independent of the will of the subject. This, however, is rejected by Brouwer. On the contrary, causality is explained by the notion of strong causal sequence. A scs can be put to use by the subject in order to realize events that are not immediately obtainable. One only has to realize the first event of a scs, or an intermediate one, in order to obtain the final event. The procedure of realizing the final (and desirable) event by realizing a preceding event was called the '*jump from end to means*', and later the *mathematical* or *cunning act*. The jump from end to means is a useful and convenient tool for the subject to dominate nature and for the protection of his personal sphere.

Assuming that in a scs $a_0, \ldots, a_k, \ldots, a_n$ the realization of all stages is indeed of a fixed determined nature, one may recognize in the jump from end to means the germ of the constructive implication. The transition from, say, a_k to a_n is completely lawlike and thus the proof interpretation of $A \to B$ is foreshadowed by the automatic and algorithmic transition from (the building for) A to (the building for) B. Of course, the subject may and will add much more regularity to causal sequences than the primitive spontaneous sequence of sensations offers.

By abstracting from all accidental features of twoities, the *empty twoity* is obtained. In other words, by identifying all twoities one obtains the object where only order and distinction are recognized. This empty twoity then can take the place of the number 2. From there it is not difficult to generalize to the individual natural numbers, and the next step – the recognition of the iteration of the 'next number' step as a legitimate mental construction, together with the corollary, the (potentially infinite) set of natural numbers – is mentioned in passing by Brouwer. He speaks of 'unlimited unfolding' (CPM, p. 1237), see also [van Dalen 2008].

Thus the basic material of 'discrete mathematics' is at the disposition of the subject. This part of the process of creating is later called *the first act of intuitionism*. We should note that the aspect of simultaneous creation of discrete and continuous, is played down, but as late as the Vienna lectures (1928) Brouwer pointed out that both acts of intuitionism are grounded in the ur-intuition. The continuum is given in the move-of-time act as the 'between'. In his Rome lecture (1908) Brouwer explicitly points out that 'the first and the second are thus kept together, and the intuition of the continuous (continere = keeping together) consists of this keeping together'. And he adds: 'This mathematical ur-intuition is nothing but the contentless abstraction of the sensation (experience) of time'. Time is thus created by the subject through the 'move of time', together with the continuum and the natural numbers. *The second act of intuitionism* is the creation of 'more or less freely proceeding infinite sequences of mathematical entities previously acquired' and of 'species', i.e., 'properties supposable for mathematical entities previously acquired'.

In CPM the two acts are tacitly lumped together under the act of 'unlimited unfolding'. The process of creation of causal sequences and complexes does extend beyond the realm of mathematics; indeed the physical world, as well as the social one is made up of those objects. If we look for a minute at the physical phenomena, then we can see the role of mathematics as follows. The objects of the physical world are obtained by abstraction from sensation complexes, a further abstraction gets the subject to mathematical objects and structures. And hence there is a natural connection between the physical universe and the mathematical, something like a projection. Although this does not explain the success of mathematics in full, it shows that the connections do not come out of the blue.

By and large, the above sketches the genesis of Brouwer's mathematical universe. In the dissertation Brouwer goes to great lengths to determine the possible sets in mathematics on the basis that there are no sets but those we can create ourselves. After the introduction of choice sequences (cf. the second act) he revised his views. The extent of the mathematical universe is modest compared to the traditional Cantorian universe, from a classical point of view, Brouwer's universe does not get beyond ω_1. But what it lacks in 'height' is compensated by the extra fine structure which is inherent to the intuitionistic approach (and its logic).

The most spectacular part of the universe is the second-order part, let us say second-order arithmetic with sequences, species, or both. Where the first-order part yields more-or-less a subtheory of classical arithmetic, the second-order part has certain specific properties that are incompatible with classical mathematics.

We will look at a few of these principles. The first and most striking principle was introduced by Brouwer in his courses on pointset theory of 1915–1917. The principle appeared in print in 1918; in modern formulation it reads 'A mapping F from choice sequences to natural numbers has the property that each $F(\alpha)$ is determined by an initial segment $\overline{\alpha}k(= (\alpha 0, \alpha 1, \alpha 2, \ldots, \alpha(k-1))$ Formalized: $\forall\alpha\exists x\forall\beta(\overline{\alpha}x = \overline{\beta}x \to F(\alpha) = F(\beta))$

The principle finds a more general form in the *Principle of weak continuity*

$$WC \qquad \forall\alpha\exists x A(\alpha, x) \to \forall\alpha\exists x\exists y\forall\beta(\overline{\alpha}y = \overline{\beta}y \to A(\beta, x))$$

Brouwer formulated his functional version in a proof, giving no argument for it. A first attempt at a justification could run as follows: in order to compute the natural number $F(\alpha)$ a finite number of steps is required, when the computation is finished only finitely many members of the sequence α have been generated, and so only this initial segment enters into the computation. Hence any sequence β with the same initial segment yields the same value under F. This argument only works in the case that only numerical information of α is used. In general, however, information of a different kind may be used.

Here is an example, formulated as a game (Brouwer introduced game formulations in his Groningen Lectures, 1930). There are two players, I and II. I provides successively information about α and II has an algorithm for computing $F(\alpha)$. At each step II may ask for more information or show the output. In our example II simply takes $F(\alpha) = 100$

	I	II
0	7	?
1	2	?
2	301	?
⋮	⋮	⋮
13	5 and α becomes stationary	$F(\alpha) = 5$

Note that I may (and perhaps *must*) give more information than just the numerical values of α. Indeed, if one accepts the idea of mathematics as a solitary play of the subject, then I and II are no more than puppets controlled by the subject. Thus the availability of full information is obvious.

Now there obviously are β's with the same initial segment $\overline{\beta}14 = \overline{\alpha}14 = (7, 2, 301, \ldots, 5)$ with $F(\beta) \neq 5$. This failure of the simple argument is caused by the fact that suddenly a condition of a higher order is put on α. And higher order condition cannot be avoided, if only because one wants to allow lawlike sequences (think of the difference between the decimals of π and those determined by flipping a coin). Hence a better argument is required. One was provided by Mark van Atten in a setting which slightly, but justifiably, extended Brouwer's framework. Brouwer demanded that once one has introduced a condition on future choices (of values or conditions), one sticks to it. However, it is fairly clear that his main stipulation was that each finite sequence of choices has at least one immediate successor. By allowing higher order conditions to be repealed, the extendibility condition is observed, and the extra flexibility certainly does not restrict the practical aspects of choice sequences. Now the possible ephemeral nature of higher order conditions, disqualifies them for use in the computation of the output of F on input α, see [van Atten–van Dalen 2002]. The analysis lays down certain conditions on the class of sequences for the validity of the continuity principle. The principle is in fact justified for the holistic universe, but we can see that there is a new problem for research: for which universes does *WC* hold? A simple example of a universes that violates the continuity principle is the one in which each sequence eventually becomes constant. The function F assigns this constant value to α; F is obviously not continuous. There is a rich literature on the continuity principle, see for example [van Dalen–Troelstra 1988a, van Dalen–Troelstra 1988b]. The continuity principle has striking consequences in everyday mathematics e.g., *Brouwer's continuity theorem – all real functions are continuous* and the *indecomposability of the continuum* – $\mathbb{R}$ *cannot be split into two non-empty parts*. Both results confirm the above mentioned incompatibility, in particular the latter shows that the principle of the excluded middle is false: $\neg\forall x \in \mathbb{R}(x = 0 \vee x \neq 0)$.

Weyl, in his basic paper, *On the new foundational crisis in mathematics* [Weyl 1921], adopted Brouwer's intuitionistic programme, adding his own interpretations to it. In particular Weyl did not give the same status to choice sequences Brouwer did. For Weyl choice sequences did not belong to mathematics proper; all he accepted was the real status of initial segments. As a consequence arbitrary reals were replaced by generating intervals. Such an interval, say (a, b) for rational a and b, represents for Weyl the open horizon of 'the reals that are potentially given by the

interval'. Concrete real numbers are given by lawlike sequences of intervals, and arbitrary ones by choice sequences, in the representing interpretation. Hence there is on Weyl's approach a fundamental distinction between existential quantification (over lawlike reals), and universal quantification (over choice reals). Apart from everything else, this destroys the hope of salvaging the principle of the excluded middle. Here Brouwer's and Weyl's roads separated. For Weyl quantified statements were 'judgement abstracts', not to be taken for real judgements, whereas Brouwer recognized quantified statements as ordinary statements with ordinary proof conditions. Hence for Weyl the continuity of all real functions was an obvious consequence of the notion of arbitrary real number (approximations follow from approximations), whereas for Brouwer there was a hard theorem to be proved. For more on the Brouwer-Weyl views, see [van Atten–van Dalen–Tieszen 2002].

A further analysis, making use of transfinite principles (*the principle of Bar Induction*, established the *bar theorem, the fan theorem*, and the *locally uniform continuity theorems* (real functions on intuitionistically compact subsets of $\mathbb{R}$ are uniformly continuous). For the practical consequences of these properties of Brouwer's universe see [van Dalen–Troelstra; van Dalen–Troelstra 1988a; 1988b].

So far the treatment of the universe was completely uniform, but in the twenties Brouwer started to make the distinction between the lawlike and the full continuum. Equivalently, between the set of lawlike sequences and the set of (all) choice sequences. Historically speaking, there was a perfect reason to do so. When dealing with infinite processes algorithms are the first things that come to mind, for the law is the thing that guarantees infinite continuation. The first Brouwerian counterexamples, were, not surprisingly, based on an algorithm: the decimal expansion of π. However, once choice sequences were recognized by him as legitimate objects (the subject is free to make choices), it was natural to look for a counterpart of the (lawlike) Brouwerian counter examples where one uses a decidable property of a lawlike sequence, which has neither been proved, nor rejected. One should fully exploit the choice-character of sequences in the hope of exploiting the properties of the full Brouwerian universe. In 1927 there are the first signs of the new method, which was published some 20 years later, and which goes by the name of the 'creating subject'. The underlying idea is that the subject investigates some particular property, while he carries out a convenient bookkeeping at the same time: if at moment n A has not yet been established, put down a 0, otherwise a 1. Brouwer uses the expression 'the creating subject experiences the truth of A'. Here it is tacitly assumed that 'the creating subject experiences the truth or he does not', the simple argument being that 'in doubt, one does not experience the truth'. A reasonable assumption. In view of the fact that the ur-intuition, in its function as a time-measuring and -introducing principle, provides the subject with a sequence of moments ordered like the natural numbers, the time parameter n is a natural one. The effect of the activity of the creating subject is that a choice sequence α is in the following way associated to a proposition A:

$$\exists\alpha(A \leftrightarrow \exists x(\alpha x \neq 0))$$

This formalization of Brouwer's argument is due to Kripke and is called *Kripke's Schema, KS*. Note that KS is an extra condition on the richness of the Brouwerian universe. It asserts the existence of particular sequences, compare the role of the axiom of choice. Thus it is not automatically seen that the old principles still hold. It has in fact been shown that KS is consistent with most principles. Kreisel formulated an interesting 'tensed modal' extension of the existing theories which captures the properties of the creating subject, and which is equivalent to the extension by *KS* (Kreisel 1967; van Dalen 1978).

The classically inclined logician will note that KS is a very weak comprehension principle, which is provable in the classical setting. So whatever strength one can expect from KS, it has to come from suitable extra principles, such as the continuity principle.

We will now proceed to show a number of consequences of KS in practical mathematics, consequences which are not mere curiosities, but which make manifest certain features of the universe one would expect, and some unexpected phenomena to boot. The proofs are carried out under the assumption of the continuity principle and Kripke's Schema. It turns out to be convenient to reformulate Kripke's Schema, such that there is at most one 1 in the sequence $\alpha : \forall x(\sum_{y\leq x}\alpha(y) \leq 1)$. Let us call such a sequence satisfying $A \leftrightarrow \exists x(\alpha x = 1)$, a *Kripke sequence for A*.

(1) $\neg\forall xy \in \mathbb{R}(x \neq y \rightarrow x\#y)$

(2) $\neg\forall xy \in \mathbb{R}(\neg\neg x < y \rightarrow x < y)$

(2) was shown by Brouwer in [Brouwer 1949b], and (1) follows by a completely similar argument.

(3) *The Principle of* $\forall\alpha\exists\beta$*-continuity fails* (Myhill 1966).
Proof: consider the statement $r \in \mathbb{R}$. We apply KS to $\forall x(\alpha(x) = 0)$:

$$\exists\beta(\forall x(\alpha(x) = 0 \leftrightarrow \exists y(\beta(y) = 1))$$

Hence $\forall\alpha\exists\beta(\ldots)$; by $\forall\alpha\exists\beta$-continuity there should be a continuous functional $G\colon \mathbb{N}^{\mathbb{N}} \rightarrow \mathbb{N}^{\mathbb{N}}$ such that $\forall\alpha((\forall x(\alpha(x) = 0 \leftrightarrow \exists y(G(\alpha)(y) = 1))$. Hence we have a continuous functional G testing if an α is the zero-sequence $\mathbf{0}$. I.e. G is $\mathbf{0}$ on all sequences distinct from $\mathbf{0}$, and non-zero on $\mathbf{0}$. This functional is clearly discontinuous.
Note that therefore there is a real foundational choice to be made here: adopt KS or $\forall\alpha\exists\beta$-continuity, but not both.

(4) *All negative dense subsets of* $\mathbb{R}$ *are indecomposable.*

By a negative subset X we mean one for which $X = X^{cc}$ (in particular the complement of a set is negative).

Proof. This theorem follows from two lemmas. Let X be negative and dense in $\mathbb{R}$.

(4.1) If $X = A \cup B$, with $A \cap B = \emptyset$, then converging sequences (a_i) and (b_i) in respectively A and B cannot have the same limit.

Assume $\forall k \exists m \forall n(|a_{m+n} - b_{m+n}| < 2^{-k})$. We consider the Kripke sequences α for $r \in \mathbb{Q}$ and β for $r \notin \mathbb{Q}$, where r is an arbitrary real number.
We define new sequences γ and c_i by

$$\begin{cases} \gamma(2n) = \alpha(n) \\ \gamma(2n+1) = \beta(n) \end{cases} \quad \text{and} \quad \begin{cases} c_{2n} = a_n \\ c_{2n+1} = b_n \end{cases}$$

Now we introduce a new sequence (d_i)

$$d_n = \begin{cases} c_n \text{ if } \forall k \leq n(\gamma(k) = 0) \\ c_k \text{ if } k \leq n \text{ and } \gamma(k) = 1 \end{cases}$$

Claim: $d \in X$.
If $d \notin X$, then $d \notin A$; hence (d_n) does not become stationary in A. So $\alpha(n) = 0$ for all n. And by the definition of Kripke sequence we get $r \notin \mathbb{Q}$.
Similarly $d \notin B$; hence (d_n) does not become stationary in B. Therefore $\beta(n) = 0$ for all n, and thus $r \notin \mathbb{Q}^c$. Contradiction.
So $\neg\neg d \in X$. But since X is negative, we find $d \in X$.
As $X = A \cup B, d \in A \vee d \in B$. If $d \in A$ then (d_n) does not become stationary in B, hence $\forall n \beta(n) = 0$. By the definition of β this implies $\neg\neg r \in \mathbb{Q}$. A similar argument shows that $\neg r \in \mathbb{Q}$ if $d \in B$. As a result we get $\neg r \in \mathbb{Q} \vee \neg\neg r \in \mathbb{Q}$. As r was en arbitrary real, we have established $\forall r \in \mathbb{R}(\neg r \in \mathbb{Q} \vee \neg\neg r \in \mathbb{Q})$, which contradicts the indecomposability of $\mathbb{R}$. Therefore $lim(a_n) \neq lim(b_n)$.
(4.2) If the above sets A and B are inhabited (i.e., contain an element), then there are sequences in A and B converging to the same point. The proof is a piece of elementary analysis, see [van Dalen 1999].
Conclusion: X is indecomposable.
This theorem shows that there are lots of indecomposable subsets of the continuum, for example the irrationals, $\mathbb{Q}^c$, and the not-not-rationals, $\mathbb{Q}^{cc}$. The continuum is clearly extremely 'connected'; even if we punch holes in it, it still remains indecomposable. Note that classically $\mathbb{Q}^c$ is *not* topologically connected. It is even zero-dimensional. Intuitionistically it has dimension 1. The moral is that the intuitionistic continuum is very tight, and that its topology will offer unknown surprises and difficulties.

(5) *The powerset of* $\mathbb{N}$ *exists*.
More precisely: each subset of $\mathbb{N}$ can be represented by a suitable $0-1$ choice sequence.
The basic idea of the proof is that, given a subset X there is for each n a Kripke sequence α_n such that $n \in X \leftrightarrow \exists x(\alpha_n(x) = 1)$ All these α_n's can be glued together to form one α that tests membership for X. For the technical details, see [van Dalen 1977].

(6) *If* $\mathbb{R}$ *is indecomposable, then there are no discontinuous functions* (van Dalen 2001).
The converse is obvious, and it allows one to conclude the indecomposability on the basis of Brouwer's negative version of the continuity theorem (cf. [Brouwer 1927).

Proof: Let f be discontinuous, say in 0. It is no restriction to assume $f(0) = 0$. Then $\exists k \forall n \exists x (|x| < 2^{-n} \wedge |f(x)| > 2^{-k})$

After determining k we can find a sequence (x_n) with $|f(x_n)| > 2^{-k}$ and $|x_n| < 2^{-n}$.

Let α and β again be Kripke sequences for $r \in \mathbb{Q}$ and $r \notin \mathbb{Q}$. Put

$$\begin{cases} \gamma(2n) = \alpha(n) \\ \gamma(2n+1) = \beta(n) \end{cases} \quad \text{and} \quad c_n = \begin{cases} x_n \text{ if } \forall k \leq n(\gamma(k) = 0) \\ x_k \text{ if } k \leq n \text{ and } \gamma(k) = 1 \end{cases}$$

(c_n) converges, say to c. As $0 < 2^{-k}$, we get $f(c) < 2^{-k} \vee f(c) > 0$. If $f(c) < 2^{-k}$, then $f(c) = 0$, so $\forall p(\gamma(p) = 0)$, which is impossible. So $f(c) > 0$, and therefore $r \in \mathbb{Q} \vee r \notin \mathbb{Q}$. As before we see that this yields a non-trivial decomposition of the continuum. Contradiction.

This result establishes an equivalence between a certain characteristic of a function and the nature of its domains. Results of this kind are familiar from recursion theory and descriptive set theory.

In our description of Brouwer's universe we have discusses a few basic principles which have unusual consequences in practical mathematics. One of the challenges of constructive mathematics, is to find new principles that embody certain specific phenomena that shed new and unexpected light on the universe. Markov's principle is one of those principles, but unfortunately, one cannot justify it on the basis of a strong notion of 'constructive'. Kripke's schema is a good candidate. What we need is more experience with its applications, furthermore it would be desirable to find a realistic mathematical principle equivalent to KS, in the tradition of reverse mathematics.

References

Brouwer LEJ (1927) Über Definitionsbereiche von Funktionen. Math Ann 97:60–75

Brouwer LEJ (1949a) Consciousness, philosophy and mathematics. In: Proceedings of the 10th international congress of philosophy, Amsterdam 1948, 3:1235–1249

Brouwer LEJ (1949b) De non-aequivalentie van de constructieve en negatieve orderelatie in het continuum. Indag Math 11:37–39. Transl. "The non-equivalence of the constructive and the negative order relation in the continuum" in CW 1, pp 495–496

Kreisel G (1967) Informal rigour and completeness proofs. In: Musgrave A, Lakatos I (eds) International colloquium of philosophy of science. North-Holland, Amsterdam, pp 138–186

Myhill J (1966) Notes towards an axiomatization of intuitionistic analysis. Logique et Analyse 9:280–297

van Atten M, van Dalen D (2002) Arguments for the continuity principle. Bull. Symb. Logic 8: 329–347

van Atten M, van Dalen D (2002) Intuitionism. In: Jaquette D (ed) A companion to philosophical logic. Blackwell, Oxford, pp 513–530

van Atten M, van Dalen D, Tieszen R (2002) Brouwer and Weyl: The phenomenology and mathematics of the intuitive continuum. Philos Math 10:203–226

van Dalen D (1977) The use of Kripke's schema as a reduction principle. J Symb Logic 42:238–240
van Dalen D (1978) An interpretation of intuitionistic analysis. Ann Math Log 13:1–43
van Dalen D (1999) From Brouwerian counter examples to the creating subject. Studia Logica 62:305–314
van Dalen D (2001) Intuitionistic logic. In: Gobble L (ed) Philosophical logic. Blackwell, Oxford, pp 224–257
van Dalen D (2008) Another look at Brouwer's dissertation. In: van Atten M et al. (eds) One hundred years of intuitionism (1907–2007), pp 3–20, Basel, Birkhäuser
van Dalen D, Troelstra AS (1988a) Constructivism in mathematics, vol 1. North-Holland, Amsterdam
van Dalen D, Troelstra AS (1988b) Constructivism in mathematics, vol 2. North-Holland, Amsterdam
Weyl H (1921) Über die neue Grundlagenkrise der Mathematik. Math Zeitschr 10:39–79

Chapter 5
Randomness, Determinism and Programs in Turing's Test*

Giuseppe Longo

5.1 Introduction

In a famous 1950 article, Alan Turing proposes, in order to operate a functional comparison between brain and machine, a game he calls "imitation game". This text is, in many respects, as fundamental as his other writings, but in a completely different field since this time it consists of an article in philosophy and human cognition. These philosophical musings divide Turing's intellectual trajectory into two parts: the first moment of it being devoted to the modeling of the action executed by calculating thought, the "Human Computer" by means of the machine that tradition has endowed with Turing's own name[1] the second moment is devoted to the analysis, from 1950 on, of the morphogenetic potentialities of phenomena of chemical diffusion ((Turing 1952). From as early as his first article of 1936, Turing had thus described his computing/deducting machine, a discrete-state machine, as he himself rightfully reminds: a record/playback head moves right or left, writes 1 or 0 on the tape, erases them. The fundamental idea: the machine consists of software (the instructions) and hardware (the material: the read/write head and the tape). This distinction, purely conceptual at the time, is the true beginning of modem Computer Science (you may recognize your Macintosh). This abstract machine can compute anything; there lies the extraordinary result of the years 1936–1937.

In fact, Turing himself, Kleene, and a few other pioneers demonstrate that all formalisms for computability, since the works of Herbrand and Gödel (1930–1931), are equivalent to Turing's machine: using lambda-calculus (Church 1932; another

G. Longo (✉)
CNRS & Département. d'Informatique Ecole Normale Supérieure 45,
Rue dUlm 75230,Paris France
e-mail: Giuseppe.Longo@ens.fr

*Invited conference, Colloquium on *Cognition, Meaning and Complexity*, Roma, June 2002 (version française dans *Intellectica*, n. 35, 2002/2, pp. 131–162, suivi par une réponse aux articles de commentaires, pp. 199–216). Reprinted also in "Parsing the Turing Test" Epstein et al. (eds.), Springer, 2008.

[1] The term "Turing machine" is traceable to A. Church, review of (Turing 1936) in Journal of Symbolic Logic, 2, 42–43, 1937. The expression employed by Turing to designate his machine is "logical computing machine".

A. Carsetti (ed.), *Causality, Meaningful Complexity and Embodied Cognition*,
Theory and Decision Library A 46, DOI 10.1007/978-90-481-3529-5_5,

fundamental formalism for computability, see (Barendregt 1984) and Section 5.4 below), they translate the various processes of arithmetic calculus the ones into the others. Consequently, all systems calculate the same class of functions on integers. That "we have an absolute" was clamored at the time (see the comment Gödel makes in 1963 on the re-edition of his 1931 article, reappearing in ((Gödel et al. 1989)): this absolute is the class of calculable (partial) functions, of integers into integers, as locus of all which is effective, calculable, in fact thinkable ("... the laws of arithmetic govern all which is enumerable. This one is the vastest of all disciplines, since it contains nor only the actual and the intuitive, but all which is thinkable." (Frege 1884)). The lambda-calculus, its types, their semantic categories are extremely rich syntactical and mathematical structures (see Hindley and Seldin 1986; Girard et al. 1990; Krivine 1990; Asperti and Longo 1991; Amadio and Curien 1998): they are still at the heart of contemporary logic and theoretical Computer Science, although there are other problems today. These formalisms have indeed been the result of a remarkable conceptual and mathematical journey, the notion of logico-formal system and language, a pillar of the mathematics of the twentieth century. In fact, a project of foundations of mathematics and of human knowledge.

Among the pioneers of this "formalist-linguistic turn" one must include the mathematicians Peano and Padoa: for them, mathematical certainty, in fact the certainty of thought and therefore thought itself, would situate itself among the "potentially mechanisable". So the first thing needing to be done was to reduce mathematics to a formal calculus, a numerical calculus that a machine should be capable of completely reproducing (hence the preliminary step: to encode mathematics in Peano's arithmetic). But which is this machine? One may also find a first intuition of it with Hilbert: he refers to "finite sequences of signs, constructed according to a finite number of rules", or to "laws of formal deduction" also written under the form of finite series of signs and, therefore, under the form of integers (and Hilbert knows what he's talking about, since he encodes, in his 1899 book, all the geometries, Euclidean and non-Euclidean, within Arithmetic by analytic means). Between 1930 and 1936, at last the intuition of these great pioneers will be formalized and, modulo a remarkable idea, goedelization,[2] extended to an arithmetical encoding of all which is finite, Turing's machine replaces Vaucanson's and Diderot's automatons: potentially, it is able to simulate any human function, thought in particular (or primarily) (Gandy 1988).

5.2 The Game, the Machine and the Continuum

In 1950, Turing had the courage to submit Peano's and Padoa's program to a sort of scientific-mental experiment: to demonstrate that a discrete-state machine, a DSM (his universal machine), is undistinguishable from a human brain, or, at least, that it is able to play and win what he calls the "imitation game", by playing against

[2] Crucial technical aspect of Gödel's proof, 1931: it allows the encoding of the formal-deductive meta theory of Arithmetic in Arithmetic itself (see Gödel et al. 1989).

a man (are, rather, a woman?). In this text, we shall not discuss the specific question raised by this game between a man, a woman and a machine, but its general and dominant interpretation: as alleged proof of a "functional equivalence" between digital machine and human brain. And we shall address the issue within a purely physico-mathematical conceptual framework.

Turing's proof is cautious: it is based on mathematical hypotheses carefully made explicit, as shall be seen. Also to be noted is a capital difference from the modem claimants of "all is program", this "all" being replaced depending of the author by evolution, the genome, the brain, etc. (in fact, in this slogan, no hypothesis is formulated, it consists solely of a description of "reality", of the Universe, itself identified to a Discrete-State Machine). Turing is to the contrary aware of the strong hypotheses that are necessary to his reasoning. The conclusion, the success of the machine in the imitation game, is also very cautious. However, the central hypothesis as well as the conclusion is not corroborated. And, today, it can be proved for this great mathematician had well exhibited hypotheses and conclusions. There lies the interest of the article: explicit premises and rich arguments. We shall therefore play Turing's game from a mathematical viewpoint, with its hypotheses, without engaging into any discussion in Philosophy of Mind: it is not necessary in order to be certain of winning against any DSM.

In a DSM, Turing observes, "... it is always possible to predict all future states". And he continues: "This is reminiscent of Laplace's view... The prediction which we are considering is, however, rather nearer to practicability than that considered by Laplace" (Turing 1950; p 47). In fact, he explains, the Universe and its processes are "sensitive to initial conditions", should we say in modern terminology. (Turing uses the following example: "The displacement of a single electron by a billionth of a centimeter at one moment might make the difference between a man being killed by an avalanche a year later, or escaping".) To the contrary, and there lies the greatest effectiveness of his approach, "It is an essential property of... (DSMs) that this phenomenon does not occur. Even when we consider the actual physical machines instead of the idealized machines," prediction is possible (Turing 1950; p 47). Thus Turing has no doubt: his machine is an ideal machine, indeed a logical. one, as he called it, with a laplacian behavior. And he is absolutely right: the notion of program and the mathematical structure of its implementation are deterministic in Laplace's sense, that is, the determination, by a finite number of rules (or equations, for laplacian mechanics), implies predictability. Of course, there may be some endowed indeterminacy (the machine can make steps which lead to an arbitrary element of a finite set of possible discrete states, instead of leading to a single one – we are then dealing with an indeterministic DSM), but it consists of probabilistic type of abstract indeterminacy already well studied by Laplace, and which is not the same mathematical concept as the unpredictability of deterministic dynamical systems, in the modern sense which we shall discuss in length.

Though, as Turing understands well, "the nervous system is surely not a DSM" (ah, if only everyone would at least agree with that!). And he specifies: "a small error in the information about the size of the nervous impulse ..." (p 57). Once again, and in modem terminology, the brain rather is a dynamical system (Turing calls these

systems "continuous"). Then how to compare a DSM with the brain? The comparison is functional and relative to the only possible access to the machine, during the imitation game: the finite sequences of a teleprinter's signs (your keyboard in front of your screen today, or mouse clicks, which start off a small program, a finite sequence of signs). Under these conditions, according to Turing, we would be unable to distinguish a continuous system, as the brain, or "... a more simple one, a differential analyzer...", from a DSM; if the continuous machine makes its response though a printer, it will be undistinguishable from a DSM's response, even if obtained by different means (continuous variations instead of discrete steps). So there is Turing's central hypothesis: if the interface with the dynamical system is given by a "discrete access grid", then it will be undistinguishable from a DSM.

In fact, today's physical DSM, our computers, simulate dynamical systems in a more than remarkable way! They develop finite approximations of the equations which model them with great efficiency: now here may we better see the "form" of an attractor than on the screen of a powerful enough machine. Their applications to aerodynamics (simulation of turbulence), for example, has considerably lowered the price of airplanes (almost no more need for wind tunnels). But... what are the conceptual, mathematical, physical differences?

Let's first evacuate any confusion between mathematical modeling and imitation, in Turing's sense. Take the discrete logistic equation $x_{n+1} = kx_n(1 - x_n)$, where $2 \leq k \leq 4$. Many physical systems (and even biological ones) are very well modeled by this function: typically in presence of an antagonist coupling, such as an x_n action coupled to a symmetric reaction $(1 - x_n)$. For some values of k this obviously deterministic transformation from [0, 1] to [0, 1], has a chaotic behavior. A slightest variation of x_0, and the evolution will radically differ moreover, except for a countable subset of initial points x_0(or a subset of "measure 0"), when $k = 4$ and n goes to infinity the sequence $\{x_n\}$ is dense in [0, 1]: its behavior is thus said to be ergodic (or quasi ergodic, to be precise, as it is so with respect to a non-standard measure – not with respect to Lebesgue-measure). However, if you start your machine a second time on the same numerical value for x_0, you will obtain the same sequence, that is what a DSM is. Conversely, in a physical (classical!) system, it makes no sense to say: "start with the exact same initial situation", for the physical measurement will always be an interval. And the dynamic is such that, as it happens, a perturbation beneath the possible measure, that is, within the interval, can shift the system towards very different evolutions.

In short, the trajectories, the portrait of the attractors (their geometrical structures), caused by variations beneath the finite grid measurement, can be very different. Now that is the complexity, from the Santa Fe' Institute to the CenECC of the ENS: it is in the possible bifurcations, in the richness of the attractors' geometrical structures, in their various forms of structural stability, up to the synchronization phenomena (in an epileptic's brain, for example) of which they might be the origin. The stakes are of geometrical Nature.

So here we are with a first approximation of the winning strategy, if we endow "imitation", the word used by Turing, with a strong meaning, usually restricted to the notion of simulation: computational model or, more precisely, computational

realization of the physico-mathematical modeling. In this case, a true physical dynamical system always wins the imitation game against a DSM, because it needs only to say: "let's start over with the same initial conditions and then let's compare the evolution of our phase portraits".

Measurement by interval and sensitivity to the initial conditions will mark the difference between the DSM and the physical system. If the system is a turbulent river, for example, it will win at its first turn and in few instants. A forced or double pendulum needs only a little more time. Start off, for example, your double penduium[3] and the computer on, say, the values 3 and 7, twice in a row: the latter will use these exact values for the numerical simulation, each time. It will then obtain the same rounded values and, except in quite exceptional cases that shall be discussed, it will describe the same trajectory. However, there is no way of starting off the physical pendulum on 3 and 7, exactly: it can only be launched upon an interval, however small it may be, around those values. After a sufficiently long moment, the physical system shall follow a second different trajectory, very different indeed, from the first with regards to its phase space (the structure engendered by all the positions and speeds compatible with the system's data). Thus "*more geometrico*", a continuous system shows the unpredictability of its evolution in comparison to a DSM, even for an observer of the "linguistic turn", who swears but by a teleprinter, because no discrete reading grid, however fine it may be, allows to stabilize a system with an unstable dynamic.

For now, we have only applied Turing's statement concerning the sensitivity of dynamical systems to initial conditions, which is at the origin of the unpredictability, and his observation that "one of the essential properties of the... DSM is that this phenomenon does not occur". Obviously, this game strategy is only a first mathematical response to what has been called, quite beyond Turing's thinking, "Turing's test", and to the myth of the machine as brain's model; it consists of a response within the framework Turing's mathematical hypotheses, which defines in several instances the brain as being "a continuous system" and his DSM, a discrete state machine, as a "laplacian machine".

Before refining the game strategy and thoroughly discussing functional imitation, let's briefly sum up the terms of this first confrontation between the machine and a physical system. We have thus supposed, as first approximation, that the machine attempts to simulate at best a dynamical system, by using a mathematical model designed on the basis of its deterministic nature (thus described by a finite number of equations, or formal rules of deduction for a logicist who wants to model thought[4]). At the first turn, it may be impossible to distinguish between the evolution of the

[3] A mathematical description of a forced pendulum can be found in Lighthill (1986).

[4] A system is deterministic, if we know to (or think we can) write a finite number of equations or rules of inference that will determine its evolution. In classical physics, determinism is inherent to the construction of scientific objectivity: the possibility to "determine" a system by a finite number of equations or of rules is intrinsic to its theoretical approach. Within this classical framework, Poincaré has demonstrated that equational determinism does not imply the predictability of the physical system. But we will come back to this, during an intermission.

DSM and that of the physical system, of which a teleprinter or a screen's pixels inform us of the numerical measurements: of course, the two evolutions are in general different, but neither is more realistic than the other (in physics, at least). However, the iteration of the simulation-modeling from the same initial conditions reveals the machine: if a DSM restarts upon the same numerical values, necessarily discrete, it will describe. the exact same evolution in the phase space; however, the dynamical instability of a physical system, necessarily restarted within an approximating interval, will cause the second trajectory to differ from the first, after a sufficiently long time, and, moreover (see Section 5.3 for more details), even the discrete reading of the physical measurements will display this difference. To conclude, we have shown that a DSM is surely not a model of the brain, at least if we consider the latter, with Turing, a continuous system, as opposed to what is pleaded in the field of classical Artificial Intelligence and by many modern cognitivists. But can a DSM imitate the brain? And what does this word mean, exactly, when referring to modeling? Turing's game allows to clarify these important concepts.

So let's continue with our game. In order to thwart this first sketch of the iteration strategy that has just been proposed, the machine (the programmer) could in fact use the trick suggested by a comment by Turing on p 58; he proposes to trick a continuous system's and a DSM's observer–comparator by having the latter produce a series of random numbers. This idea is at the center of a difference that demonstrates the philosophical and mathematical depth of the imitation game. In the concerned comment, Turing displays this radical difference which is of interest to us, and of which he is aware (see Section 5.4 below), between his "imitation game" and the mathematical modeling of physical phenomena. Of course, by applying our strategy of iteration against ergodic simulation, we would find ourselves with four trajectories all differing from one another and, in some cases, being all as realistic as one another. But we had to renounce simulation as such, as modeling of the deterministic system by a system of equations or of formal rules of inference implemented on a computer, and we have gone towards a weaker notion, that of equivalence as indistinguishably modulo a finite interface, without engaging ourselves upon the identity of the laws of behavior (the machine's program is not supposed to implement the same laws which "determine" the physical system). In fact, that is what the imitation game is and it brings us directly to the high stakes of the "simulation" of a deterministic system by ergodic method: a simulation which is in fact an imitation, to put it – like Turing – in a quite appropriate but uncommon manner.

The precisions we shall add in the next section require somewhat more competence or mathematical attention: the humanist reader who has grasped this first difference between a DSM and a dynamical system may directly jump to Section 5.4.[5]

[5] This reader, while the others read the §.2, could consult the following page http://www∼cse.ucsc.edu/-charlie/3body/ for about ten extraordinary examples of mechanical iteration of perfectly regular orbits, for 3, 6, . . . , 19, 99 bodies (crossed 8s, fantastical flowers . . . absolutely no chaos). Once found, the exact initial conditions that generate these periodical orbits, thanks to very difficult mathematics, the machine, at each click of the observer, starts over with the exact same trajectories, as perfect as unreal. Unreal, because these orbits are critical: the

5.3 Between Randomness and Deterministic Chaos

Two questions are raised at this point. The first is quite general: from a computational viewpoint, may randomness be distinguished, in practice, from chaotic determinism? And if, during our game, in order to trick the observer of the strategy of iteration, we first accepted to simulate the dynamical system (to develop the computation of an equational model), but, at the second turn, the computer added small random perturbations to the initial data or to each step of the discrete evolution?

So we have two phases. During the first (single-turn game), we observe a physical system, of which we know the discrete measurements via a teleprinter (or by screen pixels), and a computer which generates a random trajectory. Now, there exists deterministic systems, maximally unstable, such that no known method allows us to distinguish between their evolutions, reproduced upon a screen, and the generation of a random sequence: these are the "Bernoulli systems".[6] For these systems, knowledge of the past does not allow to determine the future evolution; we then say that the flow is random. Draws at lottery or dice are typical examples of this: these systems are deterministic, yet perfectly chaotic. In the two cases, the number of parameters and of equations may be quite great, yet finite, and sensitivity to the initial conditions is such that it is absolutely not worth it to attempt to write these equations: it is preferable to analyze the phenomenon in terms of laws of probability ("limit laws", for 'large numbers"). On the other hand, there exists very simple Bernoulli systems, described by one or two equations. It is thanks to these systems that we program a computer to generate random series: techniques based upon simple trigonometric properties and the multiplication of angles around 0, for example, will produce random series of $+$ and $-$ signs. Also the logistic equation of Section 5.2, for $k = 4$, generates, and in a quite economic and deterministic fashion, series of which the "global geometry" is (pseudo-) random.[7]

gravitational field of a small comet at 10 billion kilometers would topple these "planets" far away from their periodical trajectories. Some of these images give rise to laughter (and the admiration for the mathematicians who worked on them), so much are they physically absurd: even in physics, some sense of humor can help us distinguish between real world and virtual reality.

[6] For an introduction to the determinism of chaotic systems (see Dahan et al. 1992). For an increasing technicity (see Alligood et al. 2000; Lighthill 1986; Devaney 1989).

[7] In these two last cases of programmable ergodicity, it is the global knowledge of the past which says nothing about the future (the series have the appearance of globally random sequences – they can concentrate for a long time near certain values, change suddenly of attraction zone, topple a group of values very far, with no apparent regularities), but, locally, we perfectly know the next step – we have explicitly described (programmed) the laws of determination, conversely to dice and Lottery. It is the similar geometry of trajectories that allow to call ergodic all these series, physical or programmable: they show no visible regularities.

5.3.1 INTERMEZZO I (Determinism and Knowledge)

The question to which Turing brings us becomes in fact quite delicate and interesting: we do not know of "proper random" systems, in classical! Physics. More precisely, in the discrete realm, we have an excellent concept, or even a mathematical definition, of random sequence (Kolmogorov, Martin-Loef, Chaitin: "the shortest program that generates it is the sequence itself" or... "wait and see"), but all examples of natural or artificial sequences, that we know of, come from a physical deterministic system (chaotic) or from a deterministic computer program, in fact, laplacian. These programs, written in two lines, produce long "random" series: as generated by a DSM, Turing would soundly consider those sequences as being predictable (as a matter of fact, these sequences, called pseudo-random, are periodic, since they are generated by functions f as $x_{n+1} = f(x_n)$: on a concrete DSM, the finite decimal representation on a finite data base forces them to go back, soon or late, to the same number value, thus to the same sub-sequence. And, periodicity is the opposite of randomness, yet... the period may be very very long).

In a note, we have already observed that determinism is essential to the construction of scientific objectivity in classical Physics (it is "objective"); we can now add that the classical randomness is epistemic (it is a matter of "perspective" and of knowledge, it is not inherent to theoretical construction; even a gas obeys deterministic laws of local interaction between particles). Shortly, the classical randomness which we know, is nothing but highly unstable determinism *or* of unstable appearance (the computer which calculates the logistic ergodic sequence, for a fixed x_0, remains, simply and permanently, upon a trajectory which is critical, but dense in the phase space - there is the purely epistemic chaos) *or* with a very great yet finite number of parameters (dice, a gas), these "or" not being exclusive. Once again, the sequences generated by the logistic function or by a game of dice, Bernoulli's fluxes, are deterministic and ergodic. However, there is a great difference between the number of laws and of degree of freedom which will determine them and, moreover, in the logistic equation, once x_n determined, we can compute and determine x_{n+1}, as opposed to dice where a draw in no manner determines the next (see preceding note). In this sense, their common ergodicity is epistemic, for, on one hand, the observer writes the equations (the logistic equation) or knows the pertinent laws of evolution (dice) and, on the other hand, he observes a total lack of regularity in the two evolutions. It is the visible total irregularity, the geometry of the attractors if they exist, which is similar: the logistic series, just like the series of draws at dice, jumps from one end to the other of possible values, with no visible pattern. Through differing modalities, the objective determinism (or in principle) generates epistemic chaos and the phenomenal unpredictability associated to it.

But God, the perfect and infinite being who masters all laws of the Universe and who measures exactly, without approximation, without intervals, knows perfectly well the evolution of dice games and of the lottery – and of the Universe, as rightfully stated by Laplace, in a very famous and often misinterpreted page. By those words, Laplace merely lays the right absolute definition of deterministic system, outside of any construction of knowledge and of scientific objectivity, based upon

strong and well-explicated hypotheses on God, and he is right. In classical Physics, we write the same equations as God, as soon as we are capable of it, so had Galileo already claimed. But we, men (and women), we have a few problems concerning physical measurement and a different on-look than His regarding the geometry of trajectories determined by these equations: and all this becomes very important for dynamical systems, as Poincaré proved, because they may be sensitive to initial (contour) conditions and, thus, to perturbations/ fluctuations below the possible measure interval. Laplace's erroneous conjecture lies elsewhere and consists within the central hypothesis at the origin of the "calculus of perturbations" to which it has greatly contributed: from small perturbations will follow small consequences. The determinism would therefore imply the predictability modulo the inevitable approximation of the physical measurement, of which he is well aware. The invalidation of Laplace's conjecture by Poincaré will then make us understand classical randomness as particular case of deterministic chaos. And all this is very important to grasp Turing's attempt to imitate, and not to model, a continuous system by a laplacian DSM.

Now, if we want non-deterministic randomness, we can but recourse to quantum physics, thus beyond of our rather classical game: the indeterminism then, at least for the Heisenberg-type interpretation, is not epistemic, but becomes "inherent" to the construction of scientific objectivity: the probabilities are "intrinsic" to the theory and… a needle, positioned with care upon its tip, falls, classically, upon a value or another of the green mat upon which it was, after an inherently random quantum fluctuation (God, himself, really knows to play dice, but only beneath Plank's h).

So there are the stakes which are the object of such debate: classical determinism does not know, in fact, proper randomness, but only the more or less chaotic evolutions, according to various modes of determination. On the other hand, for an important trend in physical thought, quantum indeterminism is inherent to the theory. Sometimes; the latter manifests itself to our classical observation, on the tip of a needle.

Let's go back to the first phase of our game (single turn game): without God's help, we would be unable to distinguish a Bernoulli physical system from an ergodic imitation by the machine. However, there exists a continuum of classical dynamical systems which range from stable systems to Bernoulli's fluxes: in intermediary situations, the future may be predicted for the more or less long term and, particularly, the past has a greater or lesser global influence upon future trajectories. Now there are measurements, of which some are based upon the notion of entropy (topological, see (Adler 1979)), which allow to decide a deterministic system's degree of instability: on one hand, systems with nil entropy are predictable: on the other, in very high entropy systems, no observables are predictable. Between the two, numerous physical systems may be finely analyzed and, in certain cases, but there exists no general method, a partition of phase space (a topological covering by small cells), allows to conjecture the dynamics. That is, the experimental observation of a discrete trajectory allows the proposition of a deterministic law for the evolution; in these cases, different trajectories allow to guess different dynamics (in technical terms, the partitions have "generating series"). It therefore suffices to propose one

of these moderately unstable systems for a good mathematician observer to be able to recognize the random imitation made by the computer. We shall further discuss this, below, to make sure that, in this case, the strategy is in fact a winning one.

Second Phase In order to thwart this latest strategy as well as that of iteration (the two-turn game of Section 5.2) the computer implements an equational model of the physical system. However, at the second turn, in order to not fall into the trap of the genesis of an evolution identical to the first, it randomly introduces small perturbations, which may have huge consequences, of course. This second turn thus bases itself on the computation of a new deterministic system, that which adds the first to a random sequence's mechanical generator. The situation becomes delicate. If the system would admit generating series and if we were to fall upon, at the second turn, on two series which allow to guess out two differing dynamics, the distinction between the dynamical system and the DSM would be made: the series engendered by the computer would no longer be derived from the equations that modified the physical system, but a variant due to the addition of a perturbation generator. And the mathematician who knows how to reconstruct equations from generating series, once again recognizes the formal machine. But, however... even if we were to choose a system with the right level of entropy to play this game, it is not certain that we would fall upon generating series nor that we could use the rare applicable techniques to reconstruct the dynamics from these series: the machine, then, by this astute mix of modeling and ergodic imitation, would risk winning. We would then need to play the tough game of turbulence.

As of 1941, Kolmogorov and his school in fact proposed a stochastic approach to turbulence (see, with regards to this and more on turbulence, M. Farge's article in (Dahan et al. 1992)). Kolmogorov's idea was that certain random systems could adequately model turbulent phenomena. This approach, still greatly studied today, bases itself upon a quite strong hypothesis, the ergodic hypothesis. It supposes, among others, the homogeneity, the isotropy and the self-similarity of the system's evolution. Lacking of something better, the ergodic methods represent an important tool for the analysis, but it is increasingly obvious that, in certain cases, the hypotheses upon which they base themselves are not corroborated and that, to the contrary, what is important, with turbulence, is exactly the complex mixture between relatively stable structures and strong instabilities (non homogeneity, non isotropy...). Generally speaking, one does not propose meteorological previsions using ergodic methods; likewise, these methods are strongly unrecommended for the modeling of turbulence generated by a plane's wing; it would be like to trust the lottery as for the conception and the security of flight structures. In mathematical physics and in Computer Science, normally and as early as possible, one would model, meaning that one would propose and program deterministic laws which reproduce at best the natural phenomenon in question. The turingian distinction between imitation and modeling then becomes crucial: stochastic imitation à la Kolmogorv vs. modeling, for example by the Navier-Stokes equations, in our case (see (Cannone 2003) for these classical equations, today).

Now the ergodic hypothesis is invalidated by the presence of movement invariants, a sort of coherent structure, whirlpools for example, where rotation wins over deformation and who remain stable quite beyond what any statistical theory could predict. R. Thom in his work often considers these structures where, despite a highly unstable dynamic, there is a certain bearing of geometrical forms (structural stability); but that does not prevent - as Prigogine would state it - this interplay between locality stable structures and global system, of which the equations determine the range of possible regimes, from being based upon small fluctuations which, amplified, induce the choice of one of these regimes.[8]

So, on one hand, thanks to the very specific geometry of the zones of stability and of fluctuations, we know today that pure ergodicity cannot trick the expert observer (according to (Farge, 1992), Kolmogorov had understood already in 1949 the theoretical shortcomings of the ergodic hypothesis). On the other hand, we already observed that pure modeling is defeated, in the imitation game between a machine and a physical dynamical system (including a turbulent one), by iteration (Section 5.2). Finally, if the programmer mixes both strategies (modeling + ergodicity) in order to play a second turn against a well-chosen turbulent system, the coherent structures, the movement invariants, can be broken in an unnatural way and allow to distinguish the machine: there lies our thesis, based upon an anterior experience of digital techniques, by finite elements methods, for the solution of differential equations. In fact, if we fix equations for turbulence (Navier–Stokes, typically, but others are beginning to be proposed) and we implement them in a machine, the

[8] Thom's and Prigogine's points of view have enormously enriched our knowledge and, despite important differences, they are mathematically and physically compatible: the analysis in (Petitot 1990) shows it quite well. Unfortunately, the trap of ontologizing Platonism gives rise to inescapable quarrels, because it leads to confound the mathematical construction of scientific objectivity that constitutes itself between us and the world, with preexisting ontologies. An objectivity constituted between us and this reality which canalizes and causes friction upon our organisative propositions, propositions that are in no way arbitrary because they are the result of our action in this world and they are embedded in our cognitive practices and structures (Longo 2003a,b). In effect, the mathematical concepts require a conceptor who draws them on the phenomenal veil starting upon regularities that impose themselves upon his/her cognitive structure (those he/she "manages to see"); the mathematical explicitation of these regularities are part of the very process of the construction of mathematical knowledge and objectivity. To put it in husserlian terms, Platonism reduces and confounds transcendental constitution and transcendence. How much damage has this understandable reaction, in foundational reflections, of numerous great mathematicians (Gödel, Thom, Connes ... caused by the dominating formalist philosophies, which are technically difficult, but conceptually poor (those of foundations in meaningless logico-formal calculations, see next intermission). For example, in the quarrel about determinism, we even get to a dualistic separation that gives a different ontological status to fluctuation, a *material cause*, than to the global mathematical structure (the equations of a dynamic), efficient or *formal cause*, in the aristotelian terminology so dear to Thom. This latter would be the "in-itself" or the platonic idea and would precede the phenomenal appearance (Petitot 1990). The revitalization of Aristotle's fine causal analysis is very interesting (but one must not forget the "*final cause*", see (Stewart 2002)); there is, however, no need of an ontological (platonician) distinction among these four different causes. To the contrary, their interplay and temporal and conceptual simultaneity, within physical and biological phenomena, with their 'teleonomy', is the scientific challenge of today.

addition of random perturbations during the computation will not allow to choose a priori (to program) the consequences of the perturbation. Meaning that the perturbation of a step of the digital computation might, in certain instants, not limit itself to the modification of incoherent residual flows (vorticity filaments, for example), nor to redirect the regime towards other possible ones, but may break structures which have all the macroscopic characteristics of coherence and of a long stability. In short, a pebble that is thrown in a whirlpool is visible, as foreign to the turbulence: it breaks it beyond what would be, from an internal view point, the physically (geometrically) plausible. And the physical world wins again against virtual Reality.

By this, we hope to have answered to Turing's remark which proposes to imitate a continuous system, by a random system. In fact, we have taken it in a strong sense, of which he does not talk of explicitly: the possibility of a mix of strategies, modeling and ergodic imitation. Of course, we have not responded to the other great question that bothers Turing: which is the difference between a man and a woman? How to distinguish them if the man tries to imitate the woman? And if we replaced the man by a computer? Can we grasp the difference by the intermediary of a teleprinter, without seeing, without touching? (What a limitation of our material, visual and caressing humanity, but that's what the linguistic turn is).[9]

5.4 Logical, Physical and Biological Machines

In our opinion, Turing is perfectly aware of the difference between imitation and mathematical modeling for a quite simple reason: he is already working upon a remarkable mathematical model of morphogenesis in a field of chemical diffusion (a fundamental article, one of the departing points, with the work of D'Arcy Thòmpson, of the modem analyses of morphogenesis). In fact, the, most interesting property the equations to be found in (Turing 1952), is that a very small variation of the boundary conditions, obviously in a continuous system, can radically change the evolution of the model. And this property is not the laplacian nondeterminism or randomness, but the sensitivity to the contour conditions and situates itself at the heart of the deterministic model of morphogenesis à la Turing. One thing is thus the "imitation game", another mathematical modeling of physical and physico-chemical or biological phenomena: the turingian DSM does not claim to model the brain, in the physico-mathematical sense – the latter is a continuous system for Turing – it can only attempt to trick an observer (for this reason, maybe and quite rightly

[9] "[The game] is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. [. . .] We now ask the question, 'What will happen when a machine takes the part of A in this game?' Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, 'Can machines think?"' (Turing 1950).

so, some mark the beginning of classical Artificial Intelligence with this article by Turing). In the Section 5.3 we have seen that even the imitation can be revealed: in general, imitation of a dynamical system cannot be accomplished in an indistinguishable, read satisfactory manner by ergodic means, in particular if it is somewhat turbulent, but not too much. Second important precision to analyze in Turing's hypotheses. At page 47, he continues: "Even when we consider the actual physical machines instead of the idealized machines… "they are laplacian machines, as any DSM. True and false: true, the real (sequential) computer, as a DSM's realization, is by principle condemned to always make the same computation, from the same pool of discrete data and of programs, that is its logico-formal architecture (its logical gates and its programs, as formal languages). False, because it is also a physical machine, subject to variations below of its digital approximations, due to the possible small defects of its electronic circuits, to the cosmic rays that would befall upon it… It's extremely rare, but it happens. Evidently, these are sensitivities to limit conditions which have nothing to do with those, intrinsic, of continuous systems which happen to be simulated (and enormously more rare, therefore easy to detect by statistic means, by iterating the process a few times).

As a matter of fact, an abstract, mathematical DSM, such as Turing's machine, is not conceived as a physical machine, but a logical machine, a human in "the minimal act of thought" – of formal thought.[10] Consequently, its expressivity is mechanical yet purely logico-formal: typically, its expressive power is independent of spatial dimensions – of the tape, of the read/write head – a property absolutely foreign to the physical processes, which all depend and strongly upon the dimensions of space. However, when we physically bring a DSM into being, it poses new physical problems – from cosmic radiation to the synchronicity, sometimes even relativistic, of modern concurrent systems, distributed in space. Let's forget the comparison between formal DSMs and living machines, which are physical, obviously, but are moreover subject to phenomena of integration-regulation which keep them in an "extended critical state[11]"; this state is unknown by the non-living

[10] "A man provided with paper, pencil, and rubber, and subject to a strict discipline, is in effect a universal machine!… LCMs (logical computing machines, see note 1) can do anything that could be described as 'rule of thumb' or 'purely mechanical' " (Turing 1948). And Wittgenstein continues: "Turing's 'Machines'. These machines are humans who calculate." (Wittgenstein 1980). "No insight or ingenuity on the part of the human being carrying out the computation": the LCM is the breaking down of formal thought into the simplest mechanical gesture, but as a human abstraction, upon a finite sequence of meaningless signs, outside of the world.

[11] Turing refers to the brain as, at least, a dynamical physical system. To stay within his image, take a turbulent system that is at the same time very stable and very unstable, very ordinate and very inordinate; insert it sandwich style between different levels of organization that regulate it and that it integrates. You will then have a very pale physical image of a biological entity. Among these entities, quite material, soulless and without software distinct from the hardware (the modern dualism of the cognitivism of the formal rule and of the program), you will also find bodies with nervous systems that integrate and regulate them (as networks of exchange and communication), within which they integrate themselves (as organs) and by which they are regulated (by hormonal cascades. for example). These systems organize the action of the body by keeping it in a state that is physically critical, yet extended (it subsists in time and following relatively spaced out rails

and its mathematics; mathematics which must therefore be extended and adapted to the new job (dynamical systems are "only" one of the best approximations we have, for the moment). It is exactly this integration of the brain within a body, their reciprocal regulation and by such a rich environment that confers it a quite peculiar structure and functional stability; and when these regulative/integrative linkages by/of/in a body are weakened – in the course of a dream for example – the brain appears to be rather unstable (likewise in case of serious deprivation - artificial, for example – from sensation). A stability in the change (homeorhesis), anchored upon self-organization and being a feature of the living which appears extraordinarily apt to constitute invariants, from the invariants and stabilities of action to the cognitive, indeed conceptual invariants (at the heart of thought). In short, despite that we too never repeat the "same thing", in the sense of a DSM, we stabilize instabilities and critical states in a way still very ill understood, from the mathematical viewpoint. Some will then exchange the brain for a DSM: to the contrary, it is a dynamical system enormously more complex than anybody physical system or turbulent stream (... think that the banks "regulate" a stream and, there the Navier-Stokes equations tell us very little of the turbulence close to the edges; and this is nothing compared to the complexity of a brain's friction with its environment, by way of its interactions with the different levels of organization of the body to which it belongs).[12]

5.4.1 INTERMEZZO II (Machines and Deductions)

Inter 11.1 The equivalence theorems of Turing-Kleene et al. 1936–1937 (see introduction) should be considered as the second great negative result for logical formalisms, after Gödel's incompleteness theorem, 1931. That any formal deductive system, endowed with a notion of decidable proof (so any hilbertian system), can be completely simulated by a machine that goes "right, left, write/erase 0, 1", is

(see Bailly, Longo, 2010); within the limits of this state, we can find both stability and instability, variance and invariance, integration and differentiation, see Bailly, Longo 2003b). And all this in a dynamic ecosystem and in the changing history of a community of bodies-brains that interact by gestures and language (ulterior levels of organization, external to, but generated by the biological objects, this time).

[12] May it be said between us that the winning strategy proposed above for a dynamical system also applies to a man (or a woman): ask a thousand questions that require a few lines of answers each, to the human and to the machine, via a teleprinter as Turing would want. Ask the same questions the next day: you will not obtain the same responses from the human, only a continuity of meaning. In this case, the random mechanical genesis of variants is more of an attempt to trick than a mathematical counter-strategy like those of which we speak above, because there is the vexed question of meaning as well as the dynamic stability of the biological object's identity, which would show the difference. But that goes beyond the modest ambitions of this article: here we are only talking about digital machines and Physics.

a true catastrophe: what a conceptual misery these systems! (The difficulty is concealed within the monstrosity of the encoding). This philosophical shortcoming was already clear to Poincaré:

"Hilbert and Peano think that mathematics is like Chicago's sausage machine: porks and axioms go in, theorems and sausages come out" (and there comes mathematics reduced to the "manipulations of concrete signs" of which some philosophers still talk today, logic conceived as "purely formal" and mathematics – an enormous logico-analytical tautology – ready to be entirely computer generated). In fact, DSMs are generalized sausage machines (and are absolutely tremendous, for their specific uses – but sausage machines too are quite useful!). Let's not forget, however, to appreciate the full half of the glass: what an idea that of Turing who, by inventing the notion of programmable machine, manages to compute all the partial recursive functions (an enormous class of functions on $\{0, 1\}^*$, the integers) by a man/machine which goes "right, left, write/erase 0, 1". Quite obviously, this idea, with its notion of program, is the true beginning of Computer Science.

Inter 11.2 The typed lambda-calculus (Church 1940) is the only system which allows to see with equilibrium the half-full glass: the formal deductions, with all their limits and their expressivity, directly become computations, without coding (this property is called "Curry-Howard isomorphism", see (Howard 1980)). The "human computer" of Peano, Hilbert and Turing, this alienation of human rationality in a laplacian mechanism, instead of going "left, right, 0, 1", applies a little bit more complex basic formal rules – "implication–introduction", "implication–elimination" and a few others, by replacement of a sequence of signs by another and by sequence-matching (identification by mechanical superposition of signs without meaning). With recursion, the system is also a good (or paradigmatic functional) programming language. No miracle, only a very elegant constructive representation of formal proofs as programs, which placed this system at the center of the mathematics - Logic and Category Theory – for sequential calculi and languages (see Girard et al. 1990; Asperti and Longo 1991). Quite recently, it has been proposed to cognitivists to stop searching, in the brain, for a Turing Machine, but for a typed Lambda-machine (at last!): this DSM, at least, applies sequence-matching directly to rules for deduction. The lambda-calculus, "at last", because if, quite beyond of the Turing imitation game's objectives, one would obstinate oneself to seek the implementation of universal-formal rules of thought (the Laws of Thought) in the brain, one must know at least that the encoding of these laws is very important, just as under Unix or Mac-OS. In fact, the choice of the programming style (functional, logical, imperative, object oriented ..., for example) and the conception of a language with its own method for its specific coding-representation of the world and its actual expressivity, are at the heart of Computer Science, as a science, quite difficult and important, of DSMs. The computational equivalence proclaimed by the "Church thesis", is of no interest for Computer Science, since long (see the introduction at (Aceto et al. 2003)): a good share of the work happens to consist of the explicitation and use of the expressiveness of the language proposed or analyzed. Now, the terms-programs of the lambda-calculus, contrarily to the Turing Machines

and to the other formalisms, encode a great part of "the architecture" of deduction in formal systems: and, in general, "a proof has an architecture", Poincaré had already exclaimed against Hilbert and his rather flat arithmetic encodings.

It should be clear, that the limits of lambda-calculus are those of any computational formalism: it proceeds by mechanical replacement of meaningless sequences of signs and by sequence-matching. To the contrary we, when saying "if… then… else…", are not performing sequence-matching: we are displacing mountains of significations. That is the mathematical incompleteness of formalisms and the great, monist, cognitive stake for knowledge, well beyond the software/hardware/meaning distinction, quite convenient for machines and post-turigian functionalistic models of the mind, outside of this world.[13]

Let's return a last time to our game, in order to reflect. How is it possible that a great mathematician such as Turing would believe that a discrete access grid, fixed once and for all (the letters of a teleprinter, the pixels of a screen), could conceal the geometrical difference between a dynamical system (very complex, the brain) and a laplacian mechanical machine? In fact, until the results by Kolmogorov–Arnold–Moser and Ruelle in the 1960s and 1970s, the complexity (geometrical!) of continuous systems was not entirely clear, particularly the idea that the "critical" points can be dense. But the possible philosophy existed. Let's explain ourselves.

Laplace already knew well that there are critical points: the summit of a mountain of potential, for example. It is Poincaré who, thanks to his work in celestial Mechanics, will understand that the problem is "global", that it is proper to dynamical systems and to their geometry and not to a few isolated points. There is the meaning of his famous remark on sensitivity to the initial conditions: these critical points are "a bit everywhere", even though he did not exactly have the theorem which demonstrates it. It is also this attention to the physico-mathematical complexity that makes him also… conjectures the incompleteness of formal set theory, pretended universal sausage machine for mathematics (independence of the Continuum Hypothesis, in a letter to Zermelo: the theorems will come 34 and 60 years later). Just as Weyl conjectures the incompleteness of arithmetic in 1918 (Weyl 1918). Despite logicism, the philosophy of physics and that of mathematics must be profoundly linked, in order to better understand at least, as demonstrated by Poincaré and Weyl. In short, there are those who grasp the "secret darkness of milk" and its importance to knowledge and science and those who see the world through a laplacian DSM. Turing beings to the first group, except that he pushes as far as possible, within the limits of the mathematical knowledge of his times, his genius idea, the modern DSM and its notion of program, last great invention of logic-formal mechanics. Others to the contrary will follow, claiming that a DSM is a model of the brain, or even that the brain is a DSM itself (even stronger). Their motivations are often based upon this article by Turing or upon the formal Set Theory and/or Type Theory: the first is a bad reading and the second is a mathematical error (that follows from the *mathematical*, concrete, incompleteness of formalisms, see, for example (Longo 2002)).

[13] The mathematical incompleteness of formalisms is a theme strongly related to what we discuss here, see (Longo, 1999a and 2002; Bailly, Longo , 2003a) for analyses based upon recent results.

5.5 Predictability and Decidability

In a very brief text ("Laplace", downloadable, author's web) we argue the conceptual equivalence of Laplace's key hypothesis for the analysis of perturbations (the predictability of deterministic systems – as decidability of the evolution) and of the hypothesis of completeness (decidability of deducibility) of hilbertian systems, an analogy also hinted by Girard in his introduction to Turing's article. But with "Laplace" we also observed that the deterministic unpredictability à la Poincaré (the three bodies theorem 1891) is the analog and the precursor of goedelian incompleteness (undecidability) for any Hilbert-like formalism. One must however add a nuance to this analogy between the two great respective limitative results: unpredictability à la Poincaré and Gödel-like incompleteness (which corresponds to the undecidability of the halting problem, demonstrated by Turing in 1936 for his logical machine, see Girard's introduction to (Turing 1950)). The first appears "at a finite level", and very early (cf. the growth of Liapounov's coefficients in the Lindstedt–Fourier series), the latter is a problem "at infinity" (the halting problem or the non-termination of computations… forever). For example, it cannot be decided where a double pendulum will be, after 10 oscillations, nor the evolution of the solar system beyond 1 million years (Laskar 1990), a short astronomical time. So unpredictability is a "stronger" result, within the framework of an essential philosophical equivalence of the two approaches to knowledge (Laplacian in physics and formalist in logic) and of their limitative results (Poincaré and Gödel). The unpredictability of a physical dynamical system is related, in particular, to the impossibility in principle to travel the same path in the phase space, from the same initial conditions (measured by interval), whereas a DSM obstinate itself to do so. It must be observed that also Turing speaks of the unpredictability of a DSM with a large memory and very long programs (p 59), a daily experience for any computer scientist, but he is clear in these regards: we are dealing with a practical unpredictability and not one of principle, mathematical (see Turing 1950; p 47), already quoted above). We should call this unpredictability "by incompetence", like the "unpredictability" of pseudo-random mechanical generators: it has little to do with the epistemic unpredictability of the dice or of the solar system in 100 billion years. By iteration, as for pseudo-random generators, one gets the same evolution or sequence – just iterate, then you may predict. This doesn't work with dice, nor any sufficiently unstable physical systems (and a better definition of classically random process would be: if iterated under the same conditions, in general, it does not follow the same path).

The analysis we are sketching here differs from many writings, in Theory of Mind and Artificial Intelligence, regarding the "Turing test".[14] In fact,

[14] But why change the name given by Turing to the imitation game between a machine and a man/woman? The slip of scientific vision, implicit in this change of name, is very well underlined by Lassègue (1998). But would have these authors failed to grasp the profound and dramatic irony of this improbable game in which to make a computer participate: to play the difference between man and woman? Would have they ignored the evolution and the mathematical stakes of Turing's

our comparison develops itself between predictability and decidability and it is philosophical, in the sense of the theory of knowledge, but it must be reconstructed from mathematics. By this, we could understand why "imitation", such as defined by Turing, is detectable. Its mathematical (geometrical!) limit finds itself exactly in the difference between the unpredictability/undecidability results. DSMs have properties of undecidability at infinity, but are predictable in the finite realm: by looking at the program and the discrete databases one can perfectly predict the next computation step and, above all, they are predictable with regards to the iteration of the process, as described in Section 5.2. In a turingian DSM, all the laws of evolution/behavior of its own universe are explicitly and fully given (programmed) and measurement, as access to a digital database, is perfect; exactly as for God, who perfectly knows the laws and the exact measures in his universe, ours (first Intermezzo). The myth of formal machine and of absolute divinity meet and, both, their ways, detach the analysis of knowledge from its constitutive interface, between us and reality. Their counterparts in the foundations of mathematics have quartered the century between mechanistic formalism and ontologizing Platonism.

Note that Turing is so firmly convinced that his DSM is laplacian that he makes a mistake: he explicitly claims that sensitivity to initial conditions does not apply to DSMs (he stresses "*discrete-state machines*", p. 47), even in the sense that "reasonably accurate knowledge of the state (of the machine) at one moment yields reasonably accurate knowledge any number of steps later" (p 47). That is, DSMs would satisfy also Laplace's erroneous conjecture concerning approximations. Now, this happens to be false, since if the machine starts on very close but different values (reasonably accurate – but not exact – knowledge of the discrete state of the machine) for, say, x_0 in the computation of the logistic sequence, this leads, on a set of measure 1, to very different evolutions and, thus, it suffices to make the trajectory eventually unpredictable for the observer. But digital data bases are exact and the machine is laplacian, since, as for Laplace's God, the access to and use of data base, which are *discrete* and *definite*, is meant to be exact: the machine computes over a precise x_0, and not over an inevitably inexact physical measure. Moreover, the laws, organized as programs, are all given. This minor mistake by Turing is understandable, as there was little computational! experience at the time on discrete sequences engendered by non-linear equations (a rare exception is (von Neumann and Ulam 1947)); the topic came to the limelight only during the 1970s). However, this is the same mistake that lies at the hearth of his attempted undetectable imitation: the idea that a discrete grid of access, would allow to control/predict also an unstable evolution. No, control and prediction, such as made explicit by perfect iteration, are due to the exact nature of digital data bases and of formally programmed dynamics, *within* a DSM.

---

scientific project, at the same time as the tragedy of the "game" lived by this man of genius who first *projected himself* into a machine (human computer), then condemned for his homosexuality and soon to commit suicide; would they have so badly understood his mathematics as much as ignored his suffering between being and imitation: man/woman/machine?

It is modern mathematics then that makes us understand the extent to which logico-computational philosophy in cognition and foundations of mathematics stems from this newtonian-laplacian culture which has endured for too long in science, to the point of even inhibiting physico-mathematical work (and of stimulating the platonic response in philosophy of mathematics). In classical mechanics, after Poincaré, and with the exception of Hadamard and of one or two great russian mathematicians, we needed to wait for the 1960s and 1970s for his philosophies and his mathematics to be taken up. In philosophy, classical! cognitivism, stuck in the "linguistic turn", suffered the consequences of it, since it has lost first of all, in the Boole and Frege mouvance and against the philosophy of Riemann and Poincaré, the "sense of space" and of geometrical complexity. Turing, in 1950, situates himself between the two cultures, as his article in philosophy proves, jointly to his subsequent paper on morphogenesis: one must grasp the mathematical subtleties of his imitation game in order to appreciate it and to not proclaim, against Turing, that the brain is – or can be modeled by – a Turing machine, meaning a "programmable laplacian machine", all while adding… "in the end", the fateful sentence of all simplistic reductions ever promised and never accomplished.

In fact, in cognition (but also in classical Artificial Intelligence and in – formalist – philosophy of mathematics, the loci of the discrete-arithmetic modeling of the world and of thought, along the lines of. Hilbert's laplacian conjectures), we still await for a conscious reflection on paradigms comparable to the one explicitly made by Sir James Lighthill, during his chairman period at the International Association for Mechanics: "Here I have to pause and speak once again on the behalf of the broad global fraternity of practitioners of mechanics. We are deeply conscious today that the enthusiasm of the forebears for the marvelous achievements of Newtonian mechanics led them to make generalizations in this area of predictability which, indeed, we may have generally tended to believe before 1960, but which we now recognize to be false. We collectively whish to apologize for having mislead the general educated public by spreading ideas about the determinism of systems satisfying Newton's laws of motion that after 1960, were to be proved incorrect" (Lightill 1986).

In short, in Physics, Laplacian philosophy has played its part about two centuries ago; in logic, almost a century eater, it suggested an elegant formalism which engendered the Computer Science of sequentiality and its beautiful mathematics (but also a philosophy of knowledge anchored upon the physics of the nineteenth century); yet, all this is over, even in Computer Science. Quite obviously, some of its great concepts remain pillars of the modem analyses of computer programming – the structures of types, polymorphism, for example – just as the notions of hamiltonian and of lagrangian in classical mechanics have diffused into the different branches of the physics of the twentieth century, but the conceptual framework and its philosophy are radically changing. In fact, in Computer Science, the time has come for the cornputability of "data flows", of synchrony and of concurrency in (spatially) distributed systems, as opposed to that of "input–output" calculations, outside of the world – because beyond space and physical time (their time is secreted by the clock, sec (Bailly and Longo 2003a)) – typical of Laplace–Turing

sequential machines. These concurrent machines remain DSMs, so they are quite different from any dynamical system (continuous, said Turing), but they pose physical problems, as any real system, so also of spatio-temporal nature (synchronization, connectivity – as homotopy, for example (Goubault et al. 2000)). Their mathematics are in the process of realization and are about to give us a novel theory of discrete computations which greatly enriches that of Turing, Church, and of the other greats of the 1930s, because it responds to other questions than those of computability à la Turing (see Aceto et al. 2003).

5.6 Conclusion: Irreversible vs Unrepeatable

We have briefly mentioned the essential, constitutive, role of determinism in the classical Physical theories: a role confirmed by the great turning point of Poincaré, who has distinguished, mathematically, determinism from predictability. By this way, he has led us to understand randomness as epistemic, within the framework of deterministic theories (later, we even managed to say that a programmed sequence is random, if we do not know the laplacian program which generates it and if it has a behavior, a geometry, that is ergodic). On the other hand, an important trend in modern physics considers indeterminism as inherent to quantum theories and probabilities as intrinsic to this approach to microphysics.

Dynamical systems (thermodynamical and of critical type) have introduced, in modern fashion, "the arrow of time", following the essential irreversibility of their processes. But there is another concept which Computer Science places at the center of its own scientific construction: that of the repeatability of the process. In fact, it is inherent to the notion of program, the possibility of repeating the unfolding of the computation in time. That is, to start over from the same initials conditions and to follow the exact same evolution: the discrete nature of the system allows to avoid the consequences of a possible sensitivity to initial conditions, even when they are implicit in the equations implemented. There lies an essential, constitutive component of the laplacian nature of DSMs, to which Turing so clearly refers: "It is an essential property of... (DSMs) that this phenomenon does not occur". In summary, if a system is stable *or* if it is a DSM (discrete state machine!), its trajectories are repeatable, because it is not sensitive to the initial conditions *or* the eventual sensitivity does not manage to deploy its "destabilizing" effects, for re-initialization is perfect, and the unpredictability is "pushed to infinity" (the undecidability of the halting problem, Turing-style, see the beginning of Section 5.5). As does a simple pendulum, as does a clock, the computer iterates without difficulty: in fact, iteration is their job. And iteration, in Computability Theory, begins by primitive recursion, characteristic of the functions of Herbrand and Gödel Arithmetic, goes through general recursion of this same formal system and of lambda-calculus, and arrives to a very important global property of programs: the portability of software (would you buy a piece of software if it was not transferable onto any compatible machine and iteratable at will?). In short, the repeatability, along the discrete processes, is inherent to

the Theory of Computability and to its remarkable practical development, Computer Science. Specifically, it tells us that one thing is the physico-mathematical modeling, by equations with their solutions, continuous or analytical for example and if possible; and another, an ulterior step, is the implementation of these on a DSM: the latter will give us an absolutely remarkable imitation (though detectable), which is indispensable to modern science, but essentially different from (our understanding of) the physical process, for it is a discrete realization of the continuous mathematical modeling. It is necessary to grasp this point in order to develop and apply at best this talent for imitation and iteration characteristic of DSMs. Galileo would have enormously envied our possibility to iterate without limit virtual physical experiences: he had to make do with throwing and throwing again his simple pendulum and its weight, in order to propose to us the first great laws of classical physics.

On the other hand, the dynamical processes, just slightly more complex — which interest us today, are not repeatable: a double pendulum or a turbulent river do not manage to follow again and exactly the same evolution. Moreover, for some dynamical systems, recurrence theorems confirm the difference: while a continuous system only goes very close to a previously explored state, its discrete implementation eventually forces identical iterations, when the recurrence interval is below the intended decimal approximation. Thus, sequences which are recurrent or ergodic, thus dense in the phase space, become... periodic and start repeating themselves over and over again. More generally, any sequence generated by an iterated function system ($x_{n+1} = f(x_n)$) is periodic on a concrete DSM, as much as any pseudo-random generator, since they can take only a finite number of values. And, as already observed, periodicity is the opposite of density and ergodicity (but the period may be *very* long).

Unrepeatability is a concept to add to irreversibly: it does not coincide with the latter, because one can iterate the irreversible evolution of a gas, for example, as a *global* statistic, evolution of the system. It is the *local* behaviour of a particle or the series of couplings (fluctuation, bifurcation) which are unrepeatable. Similarly, it is easy to describe a reversible process, which is unrepeatable. Conjointly with determination, the (fluctuation, bifurcation) coupe is constitutive of classical dynamics and even more of biological processes: with structural stability, it participates in morphogenesis à la Turing and in the variability which is at the heart of evolution, phylogenetic and ontogenetic; it contributes to the dynamics of cognitive phenomena.

There are the stakes proposed by our response to Turing, based upon the repeatability of certain "continuous" processes, within the physical framework that he suggests herself for his game. A framework which constitutes a displacement of scientific attention from his behalf: his first works and his formal machine are part of the great ideas in Logic and in the foundations of the mathematics of the 1930s; his reflections, in the 1950 article, enrich themselves with an on-look upon contemporary mathematical physics. He thus goes beyond the limits of laplacian philosophy that had characterized the first years of work in Logic. But how is it possible that a whole branch of scientific reflection, so important technically, Mathematical Logic, could have taken such a backlog, in philosophy of nature and of knowledge, in comparison with other disciplines, Physics particularly?

The weighty, historical, responsibility of the philosophies attached to laicism and to formalism was first to isolate the problem of the mathematical foundations of our relationship to phenomenal space (we discuss this in (Longo 2003a, b)). This choice originally had good motivations, very well explicated by the two great founders, who were soundly worried for the upheaval of non-Euclidean geometries: it was urgent to abandon any reference to physical space and to base the foundational analysis upon pure logic and/or formal coherence (Frege 1884; Hilbert 1891).[15] This theoretical breakage gave us remarkable logico-formal machine, as perfect as out of this world (at least, until the arrival of today's networks and of concurrency). But, at the same time, it separated the analysis of the foundations of mathematics and, worse, of cognition, from that of Physics, because exactly at that time, between the nineteenth and twentieth centuries, new theories emerged strictly related to the problem of the mathematical intelligibility of space and time (geometry of dynamical systems and of relativistic spaces). Consequently, it separated them from our efforts in the construction of modern scientific knowledge, so strongly correlated to the constitution of mathematical concepts and structures, as well as from the major change in the philosophies of Nature proposed by the new physical theories. For example, symmetries and symmetry-breaking, at the heart of modem Physics, appear only in (Weyl (1952) as a component of the foundation (as genesis) of mathematical structures, and, more recently, in Proof Theory, by the work of Girard.

By consequence, the Platonism/formalism scholastic dominant in the philosophy of mathematics (do triangles and real numbers really exist?... "the Scylla of ontologism,... the Charybdi of nominalism... from both sides I see the emergence of the ghost of a new scholastic" (Enriques 1935) missed out on the great foundational debates in Physics, about the structure of space, about determinism, "non-locality" etc. (relativistic, dynamic, quantum systems), which marked the century. And it left us with formalisms, technically marvelous to invent and work on DSMs, but laplacian in their conception of the world – or in the organization of their own universe; a universe subdivided into small discrete boxes, well localized and perfectly stable, such as the bits of computer's memory. Turing was in the process of grasping this point, as pointed out by his imitation game between deterministic systems with differing spatio-temporal evolution ("morphogenesis"), a game between the discrete and the continuum; but he died, at age 42.

[15] This issue of well explicating the hypotheses must be a feature of the Greats (Laplace, Frege, Hilbert, Turing, ...): probably because they understand the novelty of the original conceptual framework they are proposing. If not, one may find, even quite recently, people who say they have "demonstrated" Church Thesis; small implicit hypothesis: the Universe, with all of its subsystems, is an enormous laplacian machine. But, Church Thesis is an implication, which goes from all informal definition, that of potentially mechanizable deductive calculus à la Hilbert, to specific formal systems (Church, Turing, ...). As an implication, today one could say that it is certainly within the limits of truth, in Thom's sense: "the limit of the true is not the false, but the insignificant" (see for a modem appreciation (Aceto et al. 2003)). Quite obviously the ultimate goal of these "proofs" is to talk of the brain, finite sub-systems of the Universe (for a brief history of Church's Thesis – Church–Turing's, more specifically – and of its physical and cognitive caricatures (see Copeland 2002).

Let's try to not reach the same stalemate with Biology, of which cognitive sciences cannot do without, because the living makes even less sense without its space, its action within an ecosystem, its dynamic of forms. A dialogue with these rapidly growing sciences, within which mathematics cannot pretend to any hegemony, nor to ontological priority, and which would be at the same time technical and foundational, is essential to mathematics and to their foundation, because there cannot be a philosophy of mathematics without a philosophy of nature. There lies one of the great teachings of this article by Turing, and, long before, also of Poincaré and of H. Weyl (Weyl 1918 1927); another "lone wolf" – according to his own definition – at a time when it was still being tried to demonstrate the laplacian completude of logico-formal potentially mechanizable systems. Deductive systems of which some seek, even today, the implementation in the brain and, sometimes, claiming to speak in Turing's name; and they go from imitation to model, up to the discreet seduction of the metaphor.[16]

The distinction hinted by Turing, and at the heart of our analysis, between modeling (as mathematical proposal of constitutive principles for a physical process) and imitation (functional imitation, with no commitment on the "nature" of phenomena) is a fundamental idea. It should be taken up today, both from a foundational and practical view point, as discrete-state machines are essential to modern science by their extraordinary modeling/imitation abilities.

A recent project, see the team "Morphological Complexity and Information,"[17] attempts to propose a foundational dialogue with the natural sciences (see Longo 2003a,b; Bailly and Longo 2010) as well as a few alternatives, modest and specific, to the stalemate of the arithmetic encoding of the world - a coding which is changing this very world by the descendants of Turing's DSM and their extraordinary networks, but which, transformed into a philosophy of knowledge, may prevent us of grasping its complexity and... to start thinking to the next machine.

References

Aceto L, Longo G, Victor B (eds) (2003) The difference between sequential and concurrent computations. Special Issue, Mathematical structures in computer science, vol 13, no. 4–5. Cambridge University Press

Adler RL (1979) Topological entropy and equivalence of dynamical systems. American Mathematical Society

[16] "The model simplifies, the metaphor complicates" (Nouvel 2002); it adds information, it refers to a (another) impregnating conceptual framework, a universe of methods and of knowledge that we transform onto the first one. "When a model functions as metaphor, the model becomes an object of seduction for thought if we then use it as a suggestion for the solution of a philosophical question, we will manage, abetted by this confusion, to make this metaphor appear as a 'philosophical consequence"' of mathematical modelling (Nouvel 2002).

[17] Web page: Giuseppe Longo, ENS (the papers quoted below are downloadable).

Alligood K, Sauer T, Yorke J (2000) Chaos: an introduction to dynamical systems. Springer, New York
Amadio R, Curien P-L (1998) Domains and lambda-calculi. Birkhuaser, Berlin
Asperti A, Longo G (1991) Categories, types and structures. MIT, Cambridge, MA
Bailly F, Longo G (2002) "Incomplétude et incertitude en mathématiques et en physique", actes du colloque en mémoire de *Gilles Chatelet*, Paris, Juin 2001, et actes du colloque *Giulio Preti* a trent'anni dalla scomparsa, Castello Pasquini, Castiglioncello (LI), Ottobre 2002
Bailly F, Longo G (2003a) Space, time and cognition: from the standpoint of mathematics and natural sciences. In: Peruzzi (ed) Causality and mind. Kluwer, Dordrecht
Bailly E, Longo G (2003b) Objective and epistemic complexity in biology. Invited lecture, international conference on theoretical neurobiology. New Delhi, February
Bailly F, Longo G (2010) Mathematics and Natural Sciences. Imperial College/World Scientific, to appear
Barendregt H (1984) The lambda–calculus: its syntax, its semantics. North-Holland, rev. edit
Cannone M (2003) Harmonic analysis tools for solving Navier–Stokes equations. In: Friedlander S, Serre D (eds) Handbook of mathematical fluid dynamics, vol 3. Elsevier
Copeland B (2002) The Church–Turing thesis. Stanford Encyclopedia of Philosophy. Web edition in http://p!ato.stanford.edu/entries/church-turing/#Bloopers
Dahan Delmedico A, Chabert J-L, Chemla K (1992) Chaos et déterminisme. Seuil
Devaney RL (1989) An introduction to chaotic dynamical systems. Addison-Wesley
Edelman G, tononi GA (2000) Universe of consciousness. How matter becomes imagination. Basic Books
Enriques E (1935) Philosophie scientifique et empirisme logique. Actes du Congrès international de philosophie scientifique. Hermann, Paris
Farge M (1992) Evolution des théories sur la turbulence développée.In: Dahan et al., Seuil, Paris
Frege G (1884 The Foundations of Arithmetic (English transl. Evanston 1980)
Gandy R (1988) The Confluence of Ideas in 1936. In: The Rolf Herken (ed) Universal Turing machine. Oxford University Press, Oxford, pp 55–111
van Gelder T (1998) The dynamical hypothesis in cognitive sciences, target article and discussion, Behavioral and Brain Sciences, n. 21
Girard JY, Lafont Y, Taylor P (1990) Proof and types. Cambridge University Press
Gödel K, Nagel E, Newman J, Girard JY (1989) Le théoreme de Gödel. Seuil
Goubault E (ed) (2000) Geometry in Concurrency, Special issue, Mathematical structures in computer science, vol 10, n. 4. Cambridge University Press
Hilbert D (1971) Les fondements de la géométrie, 1899 (trad. fran., Dunod, 1971)
Hindley R, Seldin J (1986) Introduction to Combinators and Lambda Calculus. London Mathematical Society
Howard W (1980) The formulas-as-types notion of construction (manuscript written in 1969). In: Seldin, Hindley (eds) To H.B. Curry: essays in combinatory logic, lambda calculus and formalism. Academic, London
Krivine JL (1990) Lambda-calcul: types et modeles. Masson, Paris
Laskar J (1990) The chaotic behavior of the solar system. Icarus 88:266–291
Lassègue J (1998) Turing. Les Belles Lettres, Paris
Lighthill J (1986) The recent recognized failure of predictability in Newtonian dynamics. Proc R Soc Lond A 407:35–50
Longo G (1988) On Church's formal theory of functions and functionals. Invited lecture, Conference on Church's Thesis after 50 years, Zeiss (NL), June 1986, in Annals Pure Appl. Logic 40:93–133
Longo G (1999a) Mathematical Intelligence, Infinity and Machines: beyond the Goedelitis. J Cons Stud special issue on Cognition 6:11–12
Longo G (1999b) The mathematical continuum, from intuition to logic. In: Petitot J et al. (eds) Naturalizing phenomenology: issues in contemporary phenomenology and cognitive sciences. Stanford University Press

Longo G (2002) On the proofs of some formally unprovable propositions and prototype proofs in Type Theory Invited Lecture, Types for Proof and Programs. Durham (GB), Dec. 2000; Lecture Notes in Computer Science, vol. 2277 (Callaghan et al. eds). Springer, pp 160–180

Longo G (2003a) The reasonable effectiveness of mathematics and its cognitive roots. In: Boi L (ed) New interactions of mathematics with natural sciences. Springer

Longo G (2003b) The constructed objectivity of mathematics and the cognitive subject. In: Mugur-Schachter M (ed) Quantum mechanics, mathematics, cognition and action. Kluwer

Nouvel P (ed) (2002) "Modèles et métaphores" dans Enquete sur le concept de modèle. Presses Univ. de France

Petitot J (1990) Note sur la querelle du déterminisme. In: Amsterdamski et al. (eds) La querelle du déterminisme. Gailimard, Paris

Pilyugin S Yu (1999) Shadowing in dynamical systems. Springer, Berlin

Sauer T (2003) Shadowing breakdown and large errors in dynamical simulations of physical systems. preprint, George Mason University

Stewart J (eds) (2002) La modélisation en biologie. In: Nouvel P (ed) Enquête sur le concept de Modèle. Presses Univ. de France

Turing A (1936) On computable numbers with an application to the Entscheidungsproblem. Proc Lond Math Soc 42:230–265

Turing A (1948) Intelligent machinery. National Physical Laboratory Report. In: Meltzer B, Michie D (eds) 1969. Machine intelligence 5. Edinburgh University Press

Turing A (1950) Computing machines and intelligence. Mind LIX (page references to its reprinted version in Boden M (ed). Oxford University Press, 1990; traduction française et introduction dans Turing A, Girard J-Y La machine de Turing. Seuil, 1991)

Turing AM (1952) The chemical basis of morphogenesis. Philos Trans R Soc B237:37–72

Von Neumann J, Ulam S (1947) On combinations of stochastic and deterministic processes. Bull AMS 53:1120

Weyl H (1918) Das Kontinuum (trad. italiana di B. Weit, Bibliopolis, 1977)

Weyl H (1927) Philosophy of mathematics and of natural sciences (English transl. Princeton University Press, 1949)

Weyl H (1952) Symmetry. Princeton University Press

Wittgenstein L (1980) Remarks on the philosophy of psychology, Blackwell, Oxford

Chapter 6
Ω-Incompleteness, Truth, Intentionality

Sergio Galvan

The subject of the paper is the ω-incompleteness of a formal theory which seeks to formalize finitist arithmetic. PRA (i.e. primitive recursive arithmetic) is normally considered to be the theory that formalizes finitist arithmetic.[1] But the arguments which follow also hold if one assumes PA (i.e. Peano arithmetic) as the theory formalizing finitist arithmetic (in a broader sense, of course). I take two points of view: one internal to the theory, and one relative to some suitable non-conservative extension of it. I shall seek to show that: (i) with respect to the first point of view, ω-incompleteness entails an irreducible distinction between truth in finitist arithmetic and provability through methods based on finitist (finitary and concrete) evidence; (ii) with respect to the second point of view, this irreducible distinction can be overcome, but only if one accepts a form of evidence (non-finitary with respect to content, finitary in form but abstract). Abstract evidence is thus the finite expression of an intensional relationship between the subject and an infinite reality.

Point (ii) will be subsequently confirmed by analysis of certain kinds of prototypical proof.

My thesis is developed on the basis of detailed formal analysis of the ω-incompleteness of first-order numerical theories (PRA in particular), and of certain kinds of prototypical proof: (1) the Euclidean proposition concerning the relationship between lowest common multiple and greatest common divisor; (2) the Euclidean algorithm of the remainders; (3) Friedman's finite form of Kruskal's theorem. The analysis of the forms of prototypical proof is conducted in Section 6.2.2.

S. Galvan (✉)
Catholic University of Milan Largo A. Gemelli, 1 20123 Milan, Italy
e-mail: sergio.galvan@unicatt.it

[1] Cfr. Smorynski (1985), pp 16–25, Simpson (1999), pp 373–374 and 381–382 and Galvan (1992), pp 117–126.

A. Carsetti (ed.), *Causality, Meaningful Complexity and Embodied Cognition*,
Theory and Decision Library A 46, DOI 10.1007/978-90-481-3529-5_6,

6.1 Irreducible Distinction Between Truth and Provability Within T

Consider the following three statements (where T can be considered coincident with PRA):

(a) (om n)(T - $\vdash \alpha(\overline{n})$), i.e. T - $\vdash \alpha(0)$ and T - $\vdash \alpha(\overline{1})$ and ....[2]
(b) T - $\vdash \forall x \mathrm{Pr}_T (\ulcorner \alpha(\dot{x}) \urcorner)$.
(c) T - $\vdash \forall x \alpha(x)$.

Firstly, (a) $\Rightarrow$ (c) expresses the usual property of omega-completeness (in short omega-3), and its falsity is well-known. The formalization of (a) $\Rightarrow$ (c) also enables one to show that omega-3 entails the inconsistency of T. In fact, the formalization of (a) $\Rightarrow$ (c) is:

omega-3 $\qquad \forall x \mathrm{Pr}_T(\ulcorner \alpha(\dot{x}) \urcorner) \rightarrow \mathrm{Pr}_T(\ulcorner \forall x \alpha(x) \urcorner)$

Now, for a specific α we have:

$$\forall x \mathrm{Pr}_T(\ulcorner \neg \mathrm{Prov}_T(\dot{x}, \ulcorner \bot \urcorner) \urcorner) \rightarrow \mathrm{Pr}_T(\ulcorner \forall x \neg \mathrm{Prov}_T(x, \ulcorner \bot \urcorner) \urcorner)$$

hence:

$$\forall x \mathrm{Pr}_T(\ulcorner \neg \mathrm{Prov}_T(\dot{x}, \ulcorner \bot \urcorner) \urcorner) \rightarrow \mathrm{Pr}_T(\ulcorner \mathrm{Cons}_T \urcorner) \qquad \text{def. } \mathrm{Cons}_T$$

$$\forall x \mathrm{Pr}_T(\ulcorner \neg \mathrm{Prov}_T(\dot{x}, \ulcorner \bot \urcorner) \urcorner) \urcorner) \rightarrow \neg \mathrm{Cons}_T \qquad \text{by G2}$$

$$\neg \mathrm{Cons}_T \qquad \text{by Feferman's Lemma}^3$$

The non-validity of omega-3 shows immediately that the derivability predicate does not behave like the truth predicate. If $\mathrm{Tr}(\alpha(0))$ and $\mathrm{Tr}(\alpha(1)) \ldots$ then $\mathrm{Tr}(\forall x \alpha(x))$, whilst the derivability of $\alpha(\overline{n})$ for all n, (om n)(T - $\vdash \alpha(\overline{n})$), does not guarantee the derivability of $\forall x \alpha(x)$. If we say: truth consists in derivability in T, then we cannot say that $\mathrm{Tr}(\forall x \alpha(x))$ even if $\mathrm{Tr}(\alpha(0))$, $\mathrm{Tr}(\alpha(1))$, etc.

But why is it not possible to pass from (a) to (c)? The passage from (a) to (c) would require two critical steps which keep the extremes (a) and (c) detached. Each of these critical steps is a reason for omega-incompleteness. I begin with the first, which consists in the fact that it is not generally the case that (a) implies (b) (i.e. omega-1).

Let (a) be the starting-point. We want to see why it is not always possible to reach (b). To show why this is not the case, I consider the moves by which one usually goes from (a) to (b).

[2] "om" is the metatheoretical universal quantifier. It means "for all".

[3] Smorinski (1977), p 847.

(a) i.e. (om n)(T - $\vdash \alpha(\bar{n})$), is an abbreviation for the following metatheoretical infinite conjunction:

$$T - \vdash \alpha(0) \text{ and } T - \vdash \alpha(\bar{1}) \text{ and } T - \vdash \alpha(\bar{2}) \text{ and } \ldots$$

Now, from the usual perspective of proof theory, a theory is constructed in order to obtain all the propositions that are true in the standard interpretation of the theory. But if (om n)$\alpha(\bar{n})$ (i.e., $\alpha(0)$ and $\alpha(\bar{1})$ and $\alpha(\bar{2})$ and ...) is true in the standard interpretation of T, then also $\forall x\alpha(x)$ is true, at least in the standard interpretation of T. It is therefore to be expected that also T - $\vdash \forall x\alpha(x)$, i.e. the theorem relative to $\forall x\alpha(x)$, follows from the infinite conjunction of the theorems relative to each of the numerical examples (om n)(T - $\vdash \alpha(\bar{n})$).

How can this infinitary relation be translated into a finitary relation of derivability? The usual arithmetical practice in cases like this is to find a proof of $\alpha(k)$ (for a certain k) which does not depend on the specific nature of k but only on the fact that k is a numeral. If the proof satisfies this requirement, it coincides with a particular exemplification of a uniform structural scheme which is invariant in the proof of the other cases, with the sole difference that other numerals take the place of k. As well-known, this is the notion of prototype-proof proposed by Herbrand: "... when we say that a theorem is true for all x, we mean that for each x individually it is possible to iterate it as proof, which may just be considered a prototype of each individual proof." How can a prototype-proof be translated within the context of a purely formal standard language? The translation is performed by identifying a term t(n) which describes uniformly for all n the proof of $\alpha(n)$ in T and by proving this in T. For T to be able to do this, however, the proof in T must be carried out with respect to the open code for the closure of $\alpha(x)$ under substitution of numerals. In symbolic terms, this requires establishing the following:

$$T - \vdash \forall x \mathrm{Prov}_T(t(x), {}^{-}\alpha(\dot{x})^{-})$$

Whence:

(b) $T - \vdash \forall x \mathrm{Pr}_T({}^{-}\alpha(\dot{x})^{-})$

Note the essential presence of the variable x in the above formula. This guarantees that t(x) is the description of the invariant proof schema underlying the proof for each single numeral. If this were not a free variable, the empty structure of the schema would not be expressible in T. T would thus express only instances of the schema and this would entail the impossibility of the finitary translation of (a).

Yet the passage from (a) to (b) is not always guaranteed. It is possible that the theory T does not know (om n)$\alpha(\bar{n})$, even though it does know that α holds individually for each numeral: $\alpha(0)$, $\alpha(\bar{1})$, $\alpha(\bar{2})$, ... The non-validity of omega-1 expresses the general non-validity of (a) $\Rightarrow$ (b). The formalization of (a) $\Rightarrow$ (b), in fact, is:

omega-1 $\forall x \mathrm{Pr}_T({}^{-}\alpha(\dot{x})^{-}) \rightarrow \mathrm{Pr}_T({}^{-}\forall x \mathrm{Pr}_T({}^{-}\alpha(\dot{x})^{-})^{-})$

which is incompatible with the scheme of uniform omega-consistency restricted to the PR-formulas (see Galvan 1994).

Hence, with respect to the context fixed by a numerical theory T, it is not always possible to prove in T that a particular property is provable in T for all numerals. Sometimes, the existence of the proof for each individual numeral may not be brought to evidence. To conclude: this is the first reason that explains the phenomenon of omega-incompleteness. It is not possible for (om n)(T- $\vdash \alpha(\bar{n})$) to imply T- $\vdash \forall x\alpha(x)$ (that is, (a) $\Rightarrow$ (c)), because this would also imply T- $\vdash \forall x\mathrm{Pr_T}(\ulcorner\alpha(\dot{x})\urcorner)$ (that is, (a) $\Rightarrow$ (b)).

But let us suppose for a moment that T- $\vdash \forall x\mathrm{Pr_T}(\ulcorner\alpha(\dot{x})\urcorner)$ holds. Does this necessarily guarantee also T - $\vdash \forall x\alpha(x)$? No, it does not, because this is the step when the second reason for omega-incompleteness comes into play, and the second critical juncture arises. To assume that T- $\vdash \forall x\mathrm{Pr_T}(\ulcorner\alpha(\dot{x})\urcorner)$ implies T - $\vdash \forall x\alpha(x)$ is in fact to hold that (b) $\Rightarrow$ (c), which can be formalized as follows:

omega-2 $\mathrm{Pr_T}(\ulcorner\forall x\mathrm{Pr_T}(\ulcorner\alpha(\dot{x})\urcorner)\urcorner) \rightarrow \mathrm{Pr_T}(\ulcorner\forall x\alpha(x)\urcorner)$

And yet, as above, considering a specific α, we have:

$$\mathrm{Pr_T}(\ulcorner\forall x\mathrm{Pr_T}(\ulcorner\neg\mathrm{Prov_T}(\dot{x},\ulcorner\bot\urcorner)\urcorner)\urcorner) \rightarrow \mathrm{Pr_T}(\ulcorner\forall_x\neg\mathrm{Prov_T})(x,\ulcorner\bot\urcorner)\urcorner)$$

then:

$\mathrm{Pr_T}(\ulcorner\forall x\mathrm{Pr_T}(\ulcorner\neg\mathrm{Prov_T}(\dot{x},\ulcorner\bot\urcorner)\urcorner)\urcorner) \rightarrow \mathrm{Pr_T}(\ulcorner\mathrm{Cons_T}\urcorner)$ def. $\mathrm{Cons_T}$

$\mathrm{Pr_T}(\ulcorner\forall x\mathrm{Pr_T}(\ulcorner\neg\mathrm{Prov_T}(\dot{x},\ulcorner\bot\urcorner)\urcorner)\urcorner) \rightarrow \neg\mathrm{Cons_T}$ by G2

$\neg\mathrm{Cons_T}$ by Feferman's Lemma and D1

The result thus obtained is the same one that follows from assuming the validity of omega-3 (the derivation is also the same, with the sole difference that D1 is applied to Feferman's Lemma): the assumption of omega-2 implies the non-consistency of T.

To conclude: the non-validity of omega-2 tells us that even if T-$\vdash \forall x\mathrm{Pr_T}(\ulcorner\alpha(\dot{x})\urcorner)$, that is, even if it is *provable* in T that all the numerical instances of $\alpha(x)$ are derivable in T (and it is not only *true* that all the numerical instances of $\alpha(x)$ are derivable in T), $\forall x\alpha(x)$ is not provable in T. In other words, the non-validity of omega-2 tells us that – at least regarding formulae $\alpha(x)$ like $\neg\mathrm{Prov_T}(x,\ulcorner\bot\urcorner)$ – the fact that the truth that all the numerical instances of $\alpha(x)$ are derivable in T does not guarantee the derivability of $\forall x\alpha(x)$ in T, depends not on the underivability of that truth in T but on the fact that the derivability of that truth is not sufficient to guarantee the derivability of $\forall x\alpha(x)$ as well.

We may now ask why also omega-2 fails. If, at least in certain cases, incompleteness is due not to the fact that T is unaware that all the numerical cases of $\alpha(x)$ are demonstrable, but to the fact that this does not enable the theory to be aware of the truth of $\forall x\alpha(x)$, what is the reason for this inability? The fact is that the theory's knowledge is closed only under the *formal* relation of logical consequence. However, the truth of (om n)$\alpha(\bar{n})$ necessarily implies the truth of $\forall x\alpha(x)$ only in the standard model, and it is well known that a first-order numerical theory is unable to characterize the standard model of natural numbers. For this reason, the theory knows (om n)$\alpha(\bar{n})$ without knowing $\forall x\alpha(x)$.

All three forms of omega-incompleteness express the distance between truth and provability As said, omega-3-incompleteness immediately shows that the *derivability* predicate does not behave like the *truth* predicate. If $Tr(\alpha(0))$ and $Tr(\alpha(1))\ldots$ then $Tr(\forall x\alpha(x))$, while the derivability of $\alpha(\bar{n})$ for all n, (om n)$(T\vdash \alpha(\bar{n}))$, does not guarantee the derivability of $\forall x\alpha(x)$.

Omega-1-incompleteness confirms the distance and increases it by extending it or deepening it. Here the difference resides not in the fact that although $T \vdash \alpha(0)$ and $T \vdash \alpha(\bar{1})$ and $T \vdash \alpha(\bar{2})$ and $\ldots$, $T \nvdash \forall x\alpha(x)$, but in the fact that although $T \vdash \alpha(0)$ and $T \vdash \alpha(\bar{1})$ and $T \vdash \alpha(\bar{2})$ and $\ldots$, $T \nvdash \forall x Pr_T(\ulcorner\alpha(\dot{x})\urcorner)$, which means that the *truth* of the infinite conjunction of the derivability assertions for each of the numerical instances is not substitutable by the *derivability* in T of the finitary assertion which expresses that conjunction. The gap between truth and derivability is increased even further in this case by the fact that the truth – which cannot be replaced with derivability – has a *syntactic* content (it concerns, that is to say, facts of derivability). Moreover, particular forms of formulae $\alpha(x)$ determining omega-1-incompleteness have the complexity of PR-formulae – that is, they are decidable formulae. The truth of the infinite conjunction therefore cannot be disputed even from a constructivist point of view. It is already established by the way in which the n-th case must be decided, although the time of the decision may be distant. (In other words, the proofs potentially already exist although they have not yet been actualized). Yet, although the series has already been determined, it is not possible to prove the statement that describes it in finitary terms.

Omega-2-incompleteness manifests another aspect of the irreducibility of truth to derivability. This consists in the fact that certain truths – for example, the truth of $\forall x\alpha(x)$ with respect to the standard model of arithmetic – are not derivable from the axioms of the theory because they are not their logical consequences. Moreover, there are no axioms able to restrict the structures to the standard ones – like the above-mentioned standard model of natural numbers – so that the said truths could be transformed into logical consequences.

6.2 External Point of View and Non-finitist Evidence

6.2.1 Platonism versus Constructivism

The above observations on the relationship between truth and derivability are conclusive only if one remains within the framework of a particular theory. What changes if the point of view is extended to the perspective of some higher theory? It is well known that truth relative to the language of a theory T can be considered as derivability in a higher theory T′, so that truths which are non-derivable in T become derivable if the theory considered is the stronger theory T′. This may induce the belief that the separation between truth and derivability is only a question of point of view. If the point of view is internal to the theory, then the split between truth and derivability exists; if instead the point of view is external, the split disappears.

However, this conclusion is based on the naive idea that the passage to the higher level is unproblematic, that is to say, that it can be accomplished without paying a price. In other words, non-derivable truths in T can indeed be derived in a higher theory, but unless one wishes to assume a pragmatist view which simply eliminates the question of justification, it is necessary to provide the reasons that underpin the higher theory. Now, what is relevant in justifying the passage from T to T′ is the justification of axioms of T′. But this justification cannot be assured by the finitist evidence because T′ is essentially more powerful than T. Alternative forms of non-finitist evidence will therefore be necessary. But what distinguishes between finitist and infinitary evidence? The answer requires one to reflect on the fact that T formalizes finitist procedures. This means that the *syntactic* notion of *derivability* in PRA is synonymous with that of finitist *evidentiability*, that is to say, with the evidentiability of contentual features of the finite concrete linguistic objects of which a formal theory is constituted. On the contrary, the non-finitist evidence is differently characterized. The difference lies maily in content: the content of finitary evidence consists of concrete and finite objects[4]; the content of non-finitist evidence consists of non-finite and non-concrete objects.[5] In respect to the first aspect (non-finite), infinitary objects may be *actually infinite* objects. In respect to the second (non-concrete) they are *abstract* objects. Those who accept an infinitary objectuality in the first sense are inclined to accept a Platonist position on the mathematical universe. Those who instead insist on the second characteristic to the exclusion of all others are clearly inclined towards a constructivist view of mathematical entities. In this sense, evidence concerns *constructive possibilities* which extend beyond those implicit in finitist procedures of construction and calculation, and which therefore require more complex forms of semanticization and conceptual explicitation. There are obviously profound differences between the realistic and constructivist options. Nevertheless, they have one point in common: the role performed by abstract non-concrete notions in both approaches and which can be briefly described as follows. The abstract is the category which allows one to contemplate the infinite (according to the Platonist approach) and perhaps to constitute it (according to the constructivist approach).

The role of the abstract in constructivism is obvious. However, the category of the abstract plays an essential role in the Platonist approach as well. Although in the latter approach (formalism + non-finitist semantics) it makes sense, for example, to view truth in the standard model of the natural numbers as the infinite conjunction of every numerical instance, knowledge of this truth consists in the derivability of the general formula. As we have seen, this is not generally possible within the theory itself, with the consequence that it is necessary to resort to an adequate higher theory. In the case of (the formal expression) of the consistency of PRA, for example, although it is possible to demonstrate in PRA of every single numeral that it is not (the

[4] Cfr. Tieszen (2005), p. 152: "Objects or concepts that can be completely represented in space-time as finitary, concrete, real, and immediately intuitable".

[5] Cfr. Tieszen (2005), p 152: "Objects or concepts that are in some sense infinitary, ideal or abstract, and not immediately intuitable".

code of) a proof of contradiction, the consistency of PRA cannot be obtained within PRA itself, but in some non-conservative extension like PA or ZF. Yet, that this latter is the derivation of the formula that, semantically understood, describes the general fact of the non-derivation for each single numerical instance (i.e. that it corresponds to the truth in the standard model of the infinite conjunction of all numerical cases) follows only on condition that the axioms of PA or ZF are taken to be *true axioms* on their corresponding domains, and that the theory as a whole is correct.[6] But if there are reasons evidencing that this is so (i.e. it is not a mere assumption), they cannot but be reasons based on evidence concerning abstract concepts. Indeed, the still partial grasp of the infinite is granted to the finite human mind only in the guise of the abstract. Hence, the Platonist approach is the infinitary semantic counterpart of a formalistic apparatus whose finite signs convey abstract meanings exemplified in the structures of the semantic dimension.

The constructivist approach, by contrast, eliminates the distance between semantic object and linguistic instrument. Consequently, it does not give rise to the phenomenon just described, in which the abstract mediates between the finite-linguistic and the set-theoretical infinite. Constructivism identifies the two moments, relinquishing on the one hand the requirement of *pure formality* of the linguistic guise, and on the other the set-theoretical universe characterized by the notion of actual infinity. How is this possible? By contentually extending the concept of derivation and by eliminating the sharp separation between the formal plane and the semantic one. The essential difference consists in the rejection of the plane of pure formality and the assumption of precise meanings in the conceptual construction and in the proof-theoretic practice. And the central category in all this is always abstractness.[7]

---

[6] Of course, the formalists claim that the truth of the axioms of ZF, or PA about the correspondent *abstract structure* of sets is not required. Indeed, considering only the case of ZF, it follows from Kreisel's conservation theorem that if α is a Π_1-formula and ZF- $\vdash \alpha$ then PRA- $\text{Cons}_{\text{ZF}} \vdash \alpha$. As a consequence, in order to obtain the consistency of PRA we just need to assume the truth of Cons_{ZF}, i.e. of a sentence regarding a *concrete* syntactical fact, and we must not committ ourselves to assume the truth of the ZF axioms about the *abstract* structure of sets. Actually, it is (trivially) true that the consistency of PRA may be obtained from the consistency of ZF, but the problem is precisely to justify the latter assertion. To require the existence of an *abstract model* is a way of solution. Another way consists in elaborating a *constructive* proof (that, of course, can be carried out with less difficulty for the included theory). In this case, however, to prove the truth of consistency means to show the truth of the syntactical fact of consistency by means of the *abstract structural* properties of the proof itself. This is typical of the constructivistic approch, about which we are going to speak below.

[7] Cfr. Gödel (1972), pp 271–273: "P. Bernays has pointed out on several occasions that, in view of the fact that the consistency of a formal system cannot be proved by any deduction procedures available in the system itself, it is necessary to go beyond the framework of finitary mathematics in Hilbert's sense in order to prove the consistency of classical mathematics or even of classical number theory. Since finitary mathematics is defined as the mathematics of *concrete intuition*, this seem to imply that *abstract concepts* are needed for the proof of consistency of number theory ... By abstract concepts, in this context, are meant concepts wich are essentially of the second or higher level, i.e. which do not have as their content properties or relations of *concrete objects* (such as combinations of symbols), but rather of *thought structures* or *thought contents* (e.g., proofs,

6.2.2 *Non-finitary Evidence and Prototypical Proofs*

This subsection furnishes some examples of mathematical arguments based on abstract and infinitary forms of evidence. These are prototypical proofs. The examples are taken from arithmetical mathematics. I shall seek to show their close connection with the phenomenon of ω-incompleteness.

If α(x) is evident because of the *general and abstract* structure of x as a standard natural number (its *intension*), then it is evident that every individual standard number x is α. The evidence of the generality rests on the evidence of an abstract intension. This is the important aspect of prototypical arguments (in Herbrand's sense as illustrated above). There are numerous examples of *universalizing* arguments of this type in mathematical practice. Let us look at some of them.

1. The Euclidean theorem on the relationship between the greatest common divisor and the lowest common multiple (proposition 34 in Book VII of Euclid's Element*s*). According to this theorem, the product of two numbers a and b is equal to the product of their greatest common divisor (MCD(a,b)) and their lowest common multiple (mcm(a,b)). Its intuitive proof proceeds in the following 'concrete' manner.

Let M be a common multiple of the integers a and b. That is, let $M = a \cdot k$ (for a certain integer k). But M is also a multiple of b, so that

$$\frac{a \cdot k}{b} = h \text{ (for another integer h).}$$

Now setting $a = a_1 \cdot \text{MCD(a,b)}$ and $b = b_1 \cdot \text{MCD(a,b)}$ one obtains:

$$h = \frac{a_1 \cdot \text{MCD}(a, b) \cdot k}{b_1 \cdot \text{MCD}(a, b)} = \frac{a_1 \cdot k}{b_1}.$$

On the other hand, $\text{MCD}(a_1, b_1) = 1$, so that k must be divisible by b_1. Thus one has:

$$k = b_1 \cdot t = \frac{b}{\text{MCD}(a, b)} \cdot t \text{ (where t is an integer)}$$

whence

$$M = a \cdot \frac{b}{\text{MCD}(a, b)} \cdot t.$$

However, the argument holds for any multiple of a and b, so that all common multiples of a and b can be represented in the above standard form. What, therefore,

---

meaningful propositions, and so on), where in the proofs of propositions about these mental objects insights are needed which are not derived from a reflection upon the combinatorial (space time) properties of the symbols representing them, but rather from a reflection upon the *meanings* involved."

is the lowest common multiple? It is the one that results from setting $t = 1$ in the standard formula. Thus

$$\mathrm{mcm}(a, b) = \frac{a \cdot b}{\mathrm{MCD}(a, b)}$$

and therefore

$$a{\cdot}b = \mathrm{mcm}(a, b){\cdot}\mathrm{MCD}(a, b).$$

Now consider the type of argument used to obtain the Euclidean result. This is an argument conducted for any numeral. In fact, the proof starts from a certain common multiple and shows that it can be transformed into a standard form. The transformation procedure is general, with the consequence that the result holds for all numerals. The result is therefore generalized.

2. Theorem according to which the greatest common divisor of the two positive integers coincides with the last remainder different from O in Euclid's algorithm. The theorem is usually proved in the following manner.

Let a and b be two positive integers. Given that any integer a is univocally representable in the form $a = b \cdot q + r$ (for $0 \leq r < b$), it is possible to construct the following sequence of equations:

$$\begin{aligned}
&a = b \cdot q_1 + r_1 && 0 < r_1 < b.\\
&b = r_1 \cdot q_2 + r_2 && 0 < r_2 < r_1.\\
&r_1 = r_2 \cdot q_3 + r_3 && 0 < r_3 < r_2.\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&r_{n-2} = r_{n-1} \cdot q_n + r_n && 0 < r_n < r_{n-1}.\\
&r_{n-1} = r_n \cdot q_{n+1}.
\end{aligned}$$

Which necessarily terminates with $r_{n+1} = 0$, given that the sequence $b, r_1, r_2, \ldots$ is a decreasing succession of integers starting from b and consequently cannot contain more than positive b integers. Now, we know that if $a = b \cdot q + c$, then MCD(a,b) = MCD(b,c).

Therefore

$$\mathrm{MCD}(a, b) = \mathrm{MCD}(b, r_1) = \ldots = \mathrm{MCD}(r_{n-1}, r_n) = r_n,$$

whence

$$\mathrm{MCD}(a, b) = r_n.$$

That is to say, the greatest common divisor of a and b coincides with the last remainder different from zero in Euclid's theorem.

Note the use made of Euclid's algorithm in the above proof. One deduces from it that the greatest common divisor is the final remainder different from zero generated by the Euclidean computation. The algorithm is therefore used as a computational scheme applicable to any two numbers a and b and able to show, after a finite number of steps in calculation of the remainders, that the final remainder different from

zero is the greatest common divisor of a and b. The computation is performed on 'concrete' values of a and b. Otherwise it would not make sense to say that the sequence b, r_1, r_2, ... cannot contain more than positive b integers. And yet the result is generalized owing to the fact that the algorithm is executable for each of these values. Here too, therefore, we have a deductive procedure structurally analogous to the one illustrated by the above example: the proof is performed on concrete example, but in principle it can be conducted for each of them. Therefore the result is generalizable.

Let us now make explicit the argument common to the two above examples. Let the expression $\alpha(x)$ denote the (open) statement corresponding to any one of the above results, for example $a \cdot b = mcm(a, b) \cdot MCD(a, b)$ (one parameter for this statement has been omitted for the sake of simplicity). What is the content of the result obtained? It has been proved for any natural n that $\alpha(n)$. This has been done by identifying the prototypical scheme transferable to each of the natural numbers. In other words, it has been obtained for a certain t(x) $Prov_T(t(x), {}^{-}\alpha(\dot{x})^{-})$, from which it is possible to derive also $\forall x Pr_T({}^{-}\alpha(\dot{x})^{-})$. Now, considering that in the standard interpretation the arithmetical predicate Pr_T represents the syntactic predicate of provability in T, and considering that in T are formalized certain evidence contents, the expression $\forall x Pr_T({}^{-}\alpha(\dot{x})^{-})$ can be translated into the expression $\forall x E\alpha(x)$ (where E is the evidence operator relative to (certain contents of) the standard model). $\forall x E\alpha(x)$ therefore states that it is evident for every standard natural x, owing to the relative abstract intension mediated by the syntactic concept of any numeral, that $\alpha(x)$. But with respect to the standard model of natural numbers, if $\alpha(x)$ is evident for every natural x image of some numeral, then also $\forall x\alpha(x)$. In other words, with respect to the standard model, the evidence operator satisfies the following principle of ω-completeness: $\forall x E\alpha(x) \rightarrow E\forall x\alpha(x)$. It is for this reason that a prototypical argument is a rigorous and secure form of proof. However, it is not a purely formal proof. It is formal until the derivation of $\forall x Pr_T({}^{-}\alpha(\dot{x})^{-})$, i.e. as a transferable procedure restricted to numerals. Thereafter, because it relies on the restriction of arithmetical models to the standard model alone, it is no longer formal but requires acceptance of a form of *abstract intensional* evidence.

To conclude, a prototypical procedure comprises a formalizable part but does not consist solely in this. It is constituted by the formal procedure restricted to numerals (entirely realizable in a system) plus the passage *ad omnes* based on the (non-formal) consideration that the procedure can be performed for every standard element (image of a numeral). This passage presupposes a capacity to grasp the abstract meaning (the intension) of a standard natural number.

3. A final observation concerns the proof of Friedman's theorem (the finitary form of Kruskal's theorem).

As said, prototypical proofs do not guarantee the entirely formal proof of statements in generalized form. However, this does not signify that theorems are not provable in an entirely formal manner if use is made of principles different from, or more powerful than, those employed for the proof restricted to numerals. This is the case of both the theorems set out above. These can be derived within the same

system – PA for example[8] – but by arguing in a broader manner (that is, following a longer route) and using complex instances of induction axiom. However, there are elementary arithmetical theorems provable with prototypical procedures that are not formally obtainable within PA. This is the case of Friedman's theorem (the finitary form of Kruskal's theorem), which can be obtained in PA for any standard integer n but cannot be formally generalized within PA. The fact is that the proof depends on the type of n, in that it expressly uses n as a standard integer in the finite ramification of Koenig's tree.[9]

The fact that Friedman's theorem is not derivable within PA, whereas derivable in PA is every example of it restricted to numerals, clearly confirms the infinitary and intensional nature of the shift to generalization implicit in the prototypical procedure. Just as recourse to principles extending beyond PA is an appeal to forms of evidence more complex and abstract than those formalizable in PA, so generalization from every numeral example is only justified by intensional evidence showing that the argument restricted to numbers holds for each of them, and that the set of objects to which the universally quantified expression refers is constituted exclusively by the domain of standard natural numbers.

6.3 Concluding Remarks on Intentionality

What is the connection between the conclusions just reached and intentionality? In other words, what does the abstractness of non-finitist evidence have to do with intentionality? In this final part of the paper I shall argue that forms of non-finitist evidence have a distinctive intentional character in the classical sense. But what does intentionality in the classical sense mean? In the contemporary theory of knowledge, by 'intentionality' is normally meant the relation, inherent in some activity by a subject, of being oriented to an objectual content. Of course there are very different opinions on whether some activity or other is oriented to an object and it is therefore intentional. However, intentionality consists in directedness to an objectual content. Yet the classical notion of intentionality, as rigorously yet innovatively propounded by Brentano, contains a component that goes beyond simple directedness to an object. It consists in the relation whereby an objectual content *appears* to the subject or is *present* to the subject. In Scholastic philosophy, this relation is termed an 'identity' – an identity, that is, which is intentional. What matters in this relation is not so much the identity (which simply expresses the fact that the subject enters into 'contact' with the object) as the fact that the object is grasped (received) by the subject as *something else* (aliquid aliud). The grasping by the subject of the object as something else signifies that the object appears or is present to the subject.

[8] Cfr. Galvan (1983), section 2, pp 255–375.

[9] On this see Longo (2002).

Now, my concluding thesis is that an abstract content can only be apprehended by a subject in the form of intentional presentness. Intentionality as simple directedness at the object can in fact be interpreted as a causal relation on the part of the object which exerts a stimulus on the subject and which is then processed by the subject himself (herself). In this case, directedness is determined by the fact that not all stimuli are processed, but only those which match the structures responsible for stimuli apprehension and processing. An abstract concept, not being reducible to concrete contents, cannot exert this influence. It can only exert an influence if it appears to the subject, i.e. is *present to his or her consciousness*. Non-finitist evidence therefore requires the activating of this capacity for intentioning the mathematical object which is realized in the multiple forms (visual, geometric, combinatorial, set-theoretic, etc.) of the being present, of the being seen, in a word, of the appearing. This capacity, in conclusion, is to be understood in terms of intentionality of consciousness, and intentionality – in as much as it is the place where the object is present to consciousness – is just what mechanical minds lack.

References

Galvan S (1983) Teoria formale dei numeri naturali. Angeli, Milano

Galvan S (1992) Introduzione ai teoremi di incompletezza. Angeli, Milano

Galvan S (1994) A note on the ω-incompleteness formalization. Studia Logica 53:389–396

Gödel K (1972) On an extension of finitary mathematics which has not yet been used. In: Feferman S et al. (eds) The collected works of Kurt Gödel, vol II: Publications: 1938–1974. Oxford 1990, pp 271–280

Longo G (2002) On the proofs of some formally unprovable propositions and prototype proofs in type theory. In: Callaghan et al. (a cura di), Lecture Notes in Computer Science, vol 2277, pp. 160–180. http://www.dmi.ens.fr/users/longo

Smorynski C (1977) The incompleteness theorems. In: Barwise J (ed) Handbook of mathematical logic. North Holland, Amsterdam

Smorynski C (1985) Self-reference and modal logic. Springer, Berlin, Heidelberg, New York

Simpson SG (1999) Subsystems of second order arithmetic. Springer, Berlin, Heidelberg, New York

Tieszen R (2005) Phenomenology, logic, and the philosophy of mathematics. Cambridge University Press, Cambridge

Part III
Complexity, Incomputability and Emergence

Chapter 7
Leibniz, Complexity and Incompleteness

Gregory Chaitin

I want to tell you about the ideas of Leibniz, Weyl and Popper on conceptual complexity or complexity of ideas, and then about more recent developments deriving from that. But first of all I'd like to thank Professor Carsetti for inviting me here to give this talk after I failed to appear when he invited me to give a talk 25 years ago. It's not often that life gives one a second chance.

I'd also like to say that I'm particularly pleased to be here in Italy because at the moment two books of mine are available in Italian: *Alla ricerca di omega* published by Adelphi in Milan, and available in the UK as *Meta Maths* and in the US as *Meta Math!*, and *Teoria algoritmica della complessità* published by Giappichelli in Turin. The Giappichelli volume is a collection of essays translated into Italian by Professor Ugo Pagallo of the University of Turin, who himself has two other volumes on complexity published by Giappichelli.

Also in May 2006 I had an article in *Le Scienze* on "I limiti della ragione," which is an Italian translation of my Gödel centenary piece on "The limits of reason" in the March 2006 issue of *Scientific American.*

Furthermore, I lived for many years in Buenos Aires, where many waves of immigration were Italian and where Spanish is spoken as if it were Italian, so I feel very much at home in Italy. Also, I get the impression that on the whole my work is regarded with sympathy in Italy, which perhaps reflects an anarchic element deep within the Italian soul.

Actually, let me start with Hermann Weyl, who was a fine mathematician and mathematical physicist. He wrote books on quantum mechanics and general relativity. He also wrote two books on philosophy: *The Open World: Three Lectures on the Metaphysical Implications of Science* (1932), a small book with three lectures that Weyl gave at Yale University in New Haven, and *Philosophy of Mathematics and Natural Science*, published by Princeton University Press in 1949, an expanded version of a book he originally published in German.

G. Chaitin (✉)
G. J. Chaitin, IBM T. J. Watson Research Center, P. O. Box 218, Yorktown Heights, NY 10598, USA
e-mail: gjchaitin@gmail.com

A. Carsetti (ed.), *Causality, Meaningful Complexity and Embodied Cognition*, Theory and Decision Library A 46, DOI 10.1007/978-90-481-3529-5_7,

In these two books Weyl emphasizes the importance for the philosophy of science of an idea that Leibniz had about complexity, a very fundamental idea. The question is what is a law of nature, what does it mean to say that nature follows laws? Here is how Weyl explains Leibniz's idea in *The Open World*, pp 40–41: The concept of a law becomes vacuous if arbitrarily complicated laws are permitted, for then there is always a law. In other words, given any set of experimental data, there is always a complicated *ad hoc* law. That is valueless; simplicity is an intrinsic part of the concept of a law of nature.

What did Leibniz actually say about complexity? Well, I have been able to find three or perhaps four places where Leibniz says something important about complexity. Let me run through them before I return to Weyl and Popper and more modern developments.

First of all, Leibniz refers to complexity in Sections V and VI of his 1686 *Discours de métaphysique*, notes he wrote when his attempt to improve the pumps removing water from the silver mines in the Harz mountains was interrupted by a snow storm. These notes were not published until more than a century after Leibniz's death. In fact, most of Leibniz's best ideas were expressed in letters to the leading European intellectuals of his time, or were found many years after Leibniz's death in his private papers. You must remember that at that time there were not many scientific journals. Instead European intellectuals were joined in what was referred to as the Republic of Letters. Indeed, publishing could be risky. Leibniz sent a summary of the *Discours de métaphysique* to the *philosophe* Arnauld, himself a Jansenist fugitive from Louis XIV, who was so horrified at the possible heretical implications, that Leibniz never sent the *Discours* to anyone else. Also, the title of the *Discours* was supplied by the editor who found it among Leibniz's papers, not by Leibniz.

I should add that Leibniz's papers were preserved by chance, because most of them dealt with affairs of state. When Leibniz died, his patron, the Duke of Hanover, by then the King of England, ordered that they be preserved, sealed, in the Hanover royal archives, not given to Leibniz's relatives. Furthermore, Leibniz produced no definitive summary of his views. His ideas are always in a constant state of development, and he flies like a butterfly from subject to subject, throwing out fundamental ideas, but rarely, except in the case of the calculus, pausing to develop them.

In Section V of the *Discours*, Leibniz states that God has created the best of all possible worlds, in that all the richness and diversity that we observe in the universe is the product of a simple, elegant, beautiful set of ideas. God simultaneously maximizes the richness of the world, and minimizes the complexity of the laws which determine this world. In modern terminology, the world is understandable, comprehensible, science is possible. You see, the *Discours* was written in 1686, the year before Leibniz's nemesis Newton published his *Principia*, when medieval theology and modern science, then called mechanical philosophy, still coexisted. At that time the question of why science is possible was still a serious one. Modern science was still young and had not yet obliterated all opposition.

The deeper idea, the one that so impressed Weyl, is in Section VI of the *Discours*. There Leibniz considers "experimental data" obtained by scattering spots of ink on a piece of paper by shaking a quill pen. Consider the finite set of data points thus

obtained, and let us ask what it means to say that they obey a law of nature. Well, says Leibniz, that cannot just mean that there is a mathematical equation passing through that set of points, because there is always such an equation! The set of points obey a law only if there is a **simple** equation passing through them, not if the equation is "fort composée" = very complex, because then there is always an equation.

Another place where Leibniz refers to complexity is in Section 7 of his *Principles of Nature and Grace* (1714), where he asks why is there something rather than nothing, why is the world non-empty, because "nothing is simpler and easier than something!" In modern terms, where does the complexity in the world come from? In Leibniz's view, from God; in modern terminology, from the choice of the laws of nature and the initial conditions that determine the world. Here I should mention a remarkable contemporary development: Max Tegmark's amazing idea that the ensemble of all possible laws, all possible universes, is simpler than picking any individual universe. In other words, the multiverse is more fundamental than the question of the laws of our particular universe, which merely happens to be our postal address in the multiverse of all possible worlds! To illustrate this idea, the set of **all** positive integers 1, 2, 3, ... is very simple, even though **particular** positive integers such as 9859436643312312 can be arbitrarily complex.

A third place where Leibniz refers to complexity is in Sections 33–35 of his *Monadology* (1714), where he discusses what it means to provide a mathematical proof. He observes that to prove a complicated statement we break it up into simpler statements, until we reach statements that are so simple that they are self-evident and don't need to be proved. In other words, a proof reduces something complicated to a consequence of simpler statements, with an infinite regress avoided by stopping when our analysis reduces things to a consequence of principles that are so simple that no proof is required.

There may be yet another interesting remark by Leibniz on complexity, but I have not been able to discover the original source and verify this. It seems that Leibniz was once asked why he had avoided crushing a spider, whereupon he replied that is was a shame to destroy such an intricate mechanism. If we take "intricate" to be a synonym for "complex," then this perhaps shows that Leibniz appreciated that biological organisms are extremely complex.

These are the four most interesting texts by Leibniz on complexity that I've discovered. As my friend Stephen Wolfram has remarked, the vast Leibniz *Nachlass* may well conceal other treasures, because editors publish only what they can understand. This happens only when an age has independently developed an idea to the point that they can appreciate its value plus the fact that Leibniz captured the essential concept.

Having told you about what I think are the most interesting observations that Leibniz makes about simplicity and complexity, let me get back to Weyl and Popper. Weyl observes that this crucial idea of complexity, the fundamental role of which has been identified by Leibniz, is unfortunately very hard to pin down. How can we measure the complexity of an equation? Well, roughly speaking, by its size, but that is highly time-dependent, as mathematical notation changes over the years and it is highly arbitrary which mathematical functions one takes as given, as primitive

operations. Should one accept Bessel functions, for instance, as part of standard mathematical notation?

This train of thought is finally taken up by Karl Popper in his book *The Logic of Scientific Discovery* (1959), which was also originally published in German, and which has an entire chapter on simplicity, Chapter VII. In that chapter Popper reviews Weyl's remarks, and adds that if Weyl cannot provide a stable definition of complexity, then this must be very hard to do.

At this point these ideas temporarily disappear from the scene, only to be taken up again, to reappear, metamorphised, in a field that I call **algorithmic information theory**. AIT provides, I believe, an answer to the question of how to give a precise definition of the complexity of a law. It does this by changing the context. Instead of considering the experimental data to be points, and a law to be an equation, AIT makes everything digital, everything becomes 0s and 1s. In AIT, a law of nature is a piece of software, a computer algorithm, and instead of trying to measure the complexity of a law via the size of an equation, we now consider the size of programs, the number of bits in the software that implements our theory:

Law: Equation → Software,

Complexity: Size of equation → Size of program, Bits of software.

The following diagram illustrates the central idea of AIT, which is a very simple toy model of the scientific enterprise:

Theory (01100...11) → **COMPUTER** → Experimental Data (110...0).

In this model, both the theory and the data are finite strings of bits. A theory is software for explaining the data, and in the AIT model this means the software produces or calculates the data exactly, without any mistakes. In other words, in our model a scientific theory is a program whose output is the data, self-contained software, without any input.

And what becomes of Leibniz's fundamental observation about the meaning of "law?" Before there was always a complicated equation that passes through the data points. Now there is always a theory with the same number of bits as the data it explains, because the software can always contain the data it is trying to calculate as a constant, thus avoiding any calculation. Here we do not have a law; there is no real theory. Data follows a law, can be understood, only if the program for calculating it is much smaller than the data it explains.

In other words, understanding is compression, comprehension is compression, a scientific theory unifies many seemingly disparate phenomena and shows that they reflect a common underlying mechanism.

To repeat, we consider a computer program to be a theory for its output, that is the essential idea, and both theory and output are finite strings of bits whose size can be compared. And the best theory is the smallest program that produces that data, that precise output. That's our version of what some people call Occam's razor. This approach enables us to proceed mathematically, to define complexity precisely and to prove things about it. And once you start down this road, the first thing you

discover is that most finite strings of bits are lawless, algorithmically irreducible, algorithmically random, because there is no theory substantially smaller than the data itself. In other words, the smallest program that produces that output has about the same size as the output. The second thing you discover is that you can never be sure you have the best theory.

Before I discuss this, perhaps I should mention that AIT was originally proposed, independently, by three people, Ray Solomonoff, A.N. Kolmogorov, and myself, in the 1960s. But the original theory was not quite right. A decade later, in the mid 1970s, what I believe to be the definitive version of the theory emerged, this time independently due to me and to Leonid Levin, although Levin did not get the definition of relative complexity precisely right. I will say more about the 1970s version of AIT, which employs what I call "self-delimiting programs," later, when I discuss the halting probability Ω.

But for now, let me get back to the question of proving that you have the best theory, that you have the smallest program that produces the output it does. Is this easy to do? It turns out this is extremely difficult to do, and this provides a new complexity-based view of incompleteness that is very different from the classical incompleteness results of Gödel (1931) and Turing (1936). Let me show you why.

First of all, I'll call a program "elegant" if it's the best theory for its output, if it is the smallest program in your programming language that produces the output it does. We fix the programming language under discussion, and we consider the problem of using a formal axiomatic theory, a mathematical theory with a finite number of axioms written in an artificial formal language and employing the rules of mathematical logic, to prove that individual programs are elegant. Let's show that this is hard to do by considering the following program **P**:

P produces the output of the first provably

elegant program that is larger than **P**.

In other words, **P** systematically searches through the tree of all possible proofs in the formal theory until it finds a proof that a program **Q**, that is larger than **P**, is elegant, then **P** runs this program **Q** and produces the same output that **Q** does. But this is impossible, because **P** is too small to produce that output! **P** cannot produce the same output as a provably elegant program **Q** that is larger than **P**, not by the definition of elegant, not if we assume that all provably elegant programs are in fact actually elegant. Hence, if our formal theory only proves that elegant programs are elegant, then it can only prove that finitely many individual programs are elegant.

This is a rather different way to get incompleteness, not at all like Gödel's "This statement is unprovable" or Turing's observation that no formal theory can enable you to always solve individual instances of the halting problem. It's different because it involves complexity. It shows that the world of mathematical ideas is infinitely complex, while our formal theories necessarily have finite complexity. Indeed, just proving that individual programs are elegant requires infinite complexity. And what precisely do I mean by the complexity of a formal mathematical theory? Well, if you take a close look at the paradoxical program **P** above, whose

size gives an upper bound on what can be proved, that upper bound is essentially just the size in bits of a program for running through the tree of all possible proofs using mathematical logic to produce all the theorems, all the consequences of our axioms. In other words, in AIT the complexity of a math theory is just the size of the smallest program for generating all the theorems of the theory.

And what we just proved is that if a program **Q** is more complicated than your theory **T**, **T** can't enable you to prove that **Q** is elegant. In other words, it takes an **N**-bit theory to prove that an **N**-bit program is elegant. The Platonic world of mathematical ideas is infinitely complex, but what we can know is only a finite part of this infinite complexity, depending on the complexity of our theories.

Let's now compare math with biology. Biology deals with very complicated systems. There are no simple equations for your spouse, or for a human society. But math is even more complicated than biology. The human genome consists of 3×10^9 bases, which is 6×10^9 bits, which is large, but which is only finite. Math, however, is infinitely complicated, provably so.

An even more dramatic illustration of these ideas is provided by the halting probability Ω, which is defined to be the probability that a program generated by coin tossing eventually halts. In other words, each **K**-bit program that halts contributes 1 over $2^{\mathbf{K}}$ to the halting probability Ω. To show that Ω is a well-defined probability between zero and one it is essential to use the 1970s version of AIT with self-delimiting programs. With the 1960s version of AIT, the halting probability cannot be defined, because the sum of the relevant probabilities diverges, which is one of the reasons it was necessary to change AIT.

Anyway, Ω is a kind of DNA for pure math, because it tells you the answer to every individual instance of the halting problem. Furthermore, if you write Ω's numerical value out in binary, in base-two, what you get is an infinite string of irreducible mathematical facts:

$$\Omega = .11011\ldots$$

Each of these bits, each bit of Ω, has to be a 0 or a 1, but it's so delicately balanced, that we will never know. More precisely, it takes an **N**-bit theory to be able to determine **N** bits of Ω.

Employing Leibnizian terminology, we can restate this as follows: The bits of Ω are mathematical facts that refute the principle of sufficient reason, because there is no reason they have the values they do, no reason simpler than themselves. The bits of Ω are in the Platonic world of ideas and therefore **necessary** truths, but they look very much like **contingent** truths, like accidents. And that's the surprising place where Leibniz's ideas on complexity lead, to a place where math seems to have no structure, none that we will ever be able to perceive. How would Leibniz react to this?

First of all, I think that he would instantly be able to understand everything. He knew all about 0s and 1s, and had even proposed that the Duke of Hanover cast a silver medal in honor of base-two arithmetic, in honor of the fact that everything can be represented by 0s and 1s. Several designs for this medal were found

among Leibniz's papers, but they were never cast, until Stephen Wolfram took one and had it made in silver and gave it to me last year as a 60th birthday present. And Leibniz also understood very well the idea of a formal theory as one in which we can mechanically deduce all the consequences. In fact, the calculus was just one case of this. Christian Huygens, who taught Leibniz mathematics in Paris, hated the calculus, because it was mechanical and automatically gave answers, merely with formal manipulations, without any understanding of what the formulas meant. But that was precisely the idea, and how Leibniz's version of the calculus differed from Newton's. Leibniz invented a notation which led you automatically, mechanically, to the answer, just by following certain formal rules.

And the idea of computing by machine was certainly not foreign to Leibniz. He was elected to the London Royal Society, before the priority dispute with Newton soured everything, on the basis of his design for a machine to multiply. (Pascal's original calculating machine could only add.)

So I do not think that Leibniz would have been shocked; I think that he would have liked Ω and its paradoxical properties. Leibniz was open to all *systèmes du monde*, he found good in every philosophy, ancient, scholastic, mechanical, Kabbalah, alchemy, Chinese, Catholic, Protestant. He delighted in showing that apparently contradictory philosophical systems were in fact compatible. This was at the heart of his effort to reunify Catholicism and Protestantism. And I believe it explains the fantastic character of his *Monadology*, which complicated as it was, showed that certain apparently contradictory ideas were in fact not totally irreconcilable.

I think we need ideas to inspire us. And one way to do this is to pick heroes who exemplify the best that mankind can produce. We could do much worse than pick Leibniz as one of these exemplifying heroes.

For more on such themes, please see the collection of my philosophical papers, Chaitin, *Thinking about Gödel and Turing: Essays on Complexity, 1970–2007*, just published by World Scientific in Singapore. World Scientific also published my 60th birthday *festschrift* volume, Calude, *Randomness and Complexity, from Leibniz to Chaitin.* See also Pagallo, *Introduzione alla filosofia digitale, da Leibniz a Chaitin,* published by Giappichelli.

Chapter 8
Incomputability, Emergence and the Turing Universe

S. Barry Cooper*

The theme of this article concerns the way in which mathematics can structure everyday discussions around a range of important issues – and can also reinforce intuitions about theoretical links between different aspects of the real world. This fits with the widespread sense of excitement and expectation felt in many fields – and of a corresponding confusion – and of a tension characteristic of a Kuhnian paradigm shift. What we have below can be seen as tentative steps towards the sort of mathematical modelling needed for such a shift to be completed.

In Section 8.1, we outline the decisive role mathematics played in the birth of modern science; and how, more recently, it has helped us towards a better understanding of the nature and limitations of the scientific enterprise. In Section 8.2, we review how the mathematics brings out inherent contradictions in the Laplacian model of scientific activity. And we look at some of the approaches to dealing with these contradictions. All this leads us back in Section 8.3 to a closer examination of those aspects of the real world which most obviously test the Laplacian model. In particular, we take a close look at the phenomenon of emergence, and learn from attempts to extract the mathematical content of emergent phenomena. Most important here is the exploration of the close relationship between emergence, definability, and invariance.

Section 8.4 involves a step back from placing too much explanatory burden on emergence and its mathematics. The need for this becomes particularly clear from an excursion into the philosophy of mind, and from some complementary input from neuroscience. In Section 8.5, we finally introduce and exercise our mathematical model, and in Section 8.6, give it what we call a 'physics road test'.

S.B. Cooper (✉)
University of Leeds, Leeds LS2 9JT, U.K.
e-mail: pmt6sbc@leeds.ac.uk

*Research supported by a Royal Society International Joint Project Grant, and by EPSRC Research Grant No. EP/G000212, *The computational structure of partial information: Definability in the local structure of the enumeration degrees*.

A. Carsetti (ed.), *Causality, Meaningful Complexity and Embodied Cognition*, Theory and Decision Library A 46, DOI 10.1007/978-90-481-3529-5_8,

8.1 The Laplacian Model Becomes More of a Model

Newton's successful prediction of planetary motions assembled the important ingredients that have become the features of scientific achievement over more than 300 years. To the powerful combination of theoretical speculation and real-world observational data, he added the computational facilitation of mathematics. As White (1997, p. 93) describes in *Isaac Newton – The Last Sorcerer*:

> If the mathematics had not been developed during the 1660s, Newton's intuitive grasp of the nature of planetary motion would have remained little more than a good idea. Without his in-depth knowledge of alchemy (which he practised during the 1670s and'80s), he would almost certainly never have expanded the limited notion of planetary motion as he saw it in 1665/6 into the grand concepts of universal gravitation, of attraction and repulsion, and of action at a distance. Finally, if the experimental evidence had not been gathered, then Newton's theories, even if substantiated by mathematics, would not have carried the weight they did in his *Principia*, nor would they have so readily inspired the practical application of mechanics and the laws of motion which led, a century later, to the Industrial Revolution.

And the essential underlying product of this coming together was the emergence into the light of day – the conscious recognition – of the computational content of the world and its amenability to capture in mathematical predictions.

Looking more closely at the nature of Newton's scientific revolution, one sees how computable prediction became part of the subsequent scientific benchmark. Going back to Aristotle and before, observation and speculation had had a close relationship. But the modern empiricism associated with Bacon, and Galileo before him, further emphasised the role of data, and of measurement, with its mathematical focus on real numbers. While Bacon's view of the inductive establishment of form in nature tied theory and observation even closer: Quoting from Francis Bacon's *Novum Organum* (Bacon 1901, p. 50):

> There are and can be only two ways of searching into and discovering truth. The one flies from the senses and particulars to the most general axioms, and from these principles, the truth of which it takes for settled and immovable, proceeds to judgment and to the discovery of middle axioms. And this way is now in fashion. The other derives axioms from the senses and particulars, rising by a gradual and unbroken ascent, so that it arrives at the most general axioms at last. This is the true way, but as yet untried.

It was in this context that Newton's work laid the basis for a model of scientific practice and theory which was to fit well with the Baconian agenda, and set constraints on science which, in retrospect, would be impossible to respect in the longer term. Twentieth century science would both expose a glaring philosophical gap in the Newtonian picture – it is a *background dependent* theory, which gives no explanation of the structure of space-time it incorporated – and demand new kinds of theory from which computable predictions would be harder to extract and verify. Relativity and quantum theory are successful theories even by Newtonian standards, allowing the extraction of computable content of a high order of predictive usefulness. But collectively these have deficiencies necessitating a bizarre range of conjectural proposals, string theoretical ones being best known (of which more later).

With the benefit of our better understanding of the nature of the relationship of theory, computation and observation, one does not need to be a philosopher to recognise the inevitability of this. There is no rigid division between theories concerned with making computable predictions, and ones which are pure metaphysics. Logical analysis of the language in which theories are framed leads us to a detailed analysis of definability in the real world, connecting with well-known hierarchies, and what is known about their computational content – more of this in Section 8.3. The point is that one cannot be surprised that reality needs a richer language than that which delivers purely computable predictions. Or, for that matter, that some mathematics capturing so-called 'causal' relationships might not be reducible to the mechanistic models sought by the good Newtonian.

Anyway, the overarching aim of science, since the time of Galileo, Bacon and Newton, became the extraction of the computational content of the world, at whatever level this might occur. The process of discovery might not have a simple model, but the outcomes should have computational content with predictive utility, and scientific experiments and mathematical proofs should be reproducible and communicable to fellow scientists. This is what Einstein (1969, p. 54) is referring to when he says:

> When we say that we understand a group of natural phenomena, we mean that we have found a constructive theory which embraces them.

Why not apply this approach to science itself? Just as Quine, Hilbert, Gödel and others provided us with a model of mathematical proof, and a better understanding of the constraints on the working mathematician, can one similarly model science and its deliverables? In a sense Laplace provide scientists with an aspirational model with his 'predictive demon' (de Laplace 1951):

> Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situations of the beings who compose it – an intelligence sufficiently vast to submit these data to analysis – it would embrace in the same formula the movements of the greatest bodies and those of the lightest atom; for it, nothing would be uncertain and the future, as the past, would be present to its eyes.

And over the centuries many have duly internalised this model in a relatively simple form. The aim for them was to emulate the Newtonian successes on which Laplace's conception was based in ever broader contexts. For Europeans, the late nineteenth century expansionism, such as the 'scramble for Africa', gave an appropriate social backdrop to the final throes of this 'onwards and upwards' view of science. In mathematics, Hilbert looked for a mathematical counterpart of the predictive demon. Here is the celebrated declaration from his opening address to the Society of German Scientists and Physicians, in Königsberg, September 1930:

> For the mathematician there is no Ignorabimus, and, in my opinion, not at all for natural science either ... The true reason why [no one] has succeeded in finding an unsolvable problem is, in my opinion, that there is no unsolvable problem. In contrast to the foolish Ignorabimus, our credo avers:
>
> We must know,
> We shall know.

Of course, as described in John Dawson's biography of Gödel, on the same day in another part of the same city, Gödel was announcing his incompleteness theorem which was to do severe damage to the programme of Hilbert, and usher in a new world of unsolvability.

The so-called Laplacian model was not a model in the sense that the Hilbertian model of mathematical proof was, and so was open to various interpretations. One had to wait until the 1930s for something more mathematical and vaguely relevant.

In 1936 Turing's machines appeared (Turing 1936). This was not the first model of computability, but the one closest to the mechanistic spirit of Newton's science, and certainly the one which is reputed to have persuaded Gödel that it did achieve its modelling aim. In the first instance, the Turing machine gave a model of computability of functions over the natural numbers. But given the existence of simple codings, it essentially provided a model of algorithmic natural processes within structures which are countably presented. The mathematics of this needs qualifying, but the wide applicability of the model is generally recognised.

But the Turing's coding techniques for presenting machines gave a Universal Turing Machine – and with this came via the simplest of additions to the language used to describe machines – incomputable objects. Our model of computability arrived, like Sinbad the Sailor bearing the Old Man of the Sea, with a mathematically simple avatar of incomputability on its back. The Universal Turing Machine now has a secure place in the history of the computer – see Davis (2000) *The Universal Computer: The Road from Leibniz to Turing*. In contrast, incomputability is an irrelevance to most people beyond the confines of mathematics, and to many of those within. Teuscher's (2004) comprehensive collection *Alan Turing: Life and Legacy of a Great Thinker* contains not one article on the mathematical theory of Turing incomputability.

8.2 Some Uncomfortable Consequences

Since 1936 there has grown up a rich theory of incomputability, complete with hierarchies, fine structure theory, and an analysis of incomputable objects very close to being computable. The latter include *computably enumerable* sets, which have roughly the same relationship to computable sets that computably simulable events in the real world have to ones in which can be computably predicted. There are other kinds of sets which while not being computable, have approximations with computable characteristics, such as the Δ^0_2 sets of the arithmetical hierarchy, which have computable approximations to its members in which finitely many mistakes are allowed before the approximation settles down. Such sets will be dear to the hearts of those, such as Turing, who recognised the limitations of monotonic reasoning – here is Turing talking to the London Mathematical Society on February 20, 1947 (quoted by Hodges 1992, p. 361):

> ... if a machine is expected to be infallible, it cannot also be intelligent. There are several theorems which say almost exactly that.

Back in the real world, there was a huge investment in the Laplacian model. And any evidence to the contrary was seen more as a discipline problem than glimpse of a new world; a challenge to be soberly put down with reductionist authority. At times this was timely, such as David Deutsch's influential 1985 Royal Society paper (Deutsch 1985) bringing the standard model of quantum computation within the Turing fold. More generally, Davis (2004) writing on *The myth of hypercomputation* has argued that:

> The great success of modern computers as all-purpose algorithm-executing engines embodying Turing's universal computer in physical form, makes it extremely plausible that the abstract theory of computability gives the correct answer to the question 'What is a computation?', and, by itself, makes the existence of any more general form of computation extremely doubtful.

This should be read as a response to what Davis sees as the inflated hypercomputationalist claims of Jack Copeland and others. Copeland coined the term 'hypercomputation' to describe what an oracle Turing machine might perform. In his article (Copeland 1998) on *Turing's O-Machines, Penrose, Searle, and the Brain*, Copeland explains what oracle machines are capable of:

> Let *first-order* O-machines be those whose (only) oracle returns the values of Turing's halting function $H(x, y)$ ... Similarly, the second-order O-machines are those that possess an oracle which can say whether or not any given first-order O-machine eventually halts if set in motion with such-and-such input; and so on for third-order, and in general α-order ...
>
> It is natural to think of the functions, or problems, that are solvable by a first-order oracle machine as being *harder* than those solvable by Turing machine, and those solvable by second-order oracle machine as being harder still, and so forth.

It is the 'might be' that so annoys Davis. It is only 'natural' in a real-world sense if one can tell us where these oracles are coming from, otherwise there is no evidence such machines have any physical existence. In his article, Copeland has already by-passed this question:

> Speculation as to whether there may actually be physical processes that cannot be simulated by Turing machine stretches back over at least four decades (for example Da Costa and Doria 1991; Doyle 1982; Geroch and Hartle 1986; Komar 1964; Kreisel 1967, 1974; Penrose 1989, 1994; Pour-El 1974; Pour-El and Richards 1979, 1981; Scarpellini 1963; Stannett 1990; Vergis et al. 1986). If such processes do exist then perhaps future engineers will use them to implement the non-classical part of some O-machine.

Of course, Copeland and Davis are applying the perspectives of different disciplines, and neither managing to say very much new relating to the nature of physical computation. Of course, the more speculative proposals for computational models transcending the so-called 'Turing barrier', some of which Davis discusses in his paper, are a mixed bag. The impression one gets from the debate is that one still needs to understand more about how the real world computes.

Despite huge advances in our computational capabilities, there persist problems of predictability in the real world – at the quantum level, in the relationship between emergence and chaos, regarding relativistic phenomena (see Németi and Andréka 2006), and, of course, with mental phenomena. And increasingly the computational

capabilities of the physical are seen as relevant to the computing machines we build. There is renewed interest in analog and hybrid computing machines leading van Leeuwen and Wiedermann (2000) to consider that:

> . . . the classical Turing paradigm may no longer be fully appropriate to capture all features of present-day computing.

Despite his 1985 paper (Deutsch 1985) mentioned earlier, Deutsch did not show that quantum computation *cannot* transcend the Turing barrier, just that the current model does not do it. As Andrew Hodges remarks in *What would Alan Turing have done after 1954?* (Hodges 2004):

> Von Neumann's axioms distinguished the **U** (unitary evolution) and **R** (reduction) rules of quantum mechanics. Now, quantum computing so far (in the work of Feynman, Deutsch, Shor, etc.) is based on the **U** process and so computable. It has not made serious use of the **R** process: the unpredictable element that comes in with reduction, measurement, or collapse of the wave function.

Although measurement does play a role in quantum computation, and the probabilities of a particular outcome of a measurement are computable, there are still aspects of the physics which are not used which are thought to be in some sense 'random'. Recently, under reasonable assumptions about the basic character of quantum randomness, Calude and Svozil (2008) have shown that quantum uncertainty does entail incomputability – though just *how* random quantum randomness really is still very much open to question. It may be that despite all the assumptions of physicists, nature is full of incomputability, but does not exhibit any significant level of mathematical randomness. The challenge is to integrate quantum phenomena into a general picture of physical computation. This might not entail a useable unified theory of physics, but would hopefully present quantum uncertainty as a feature of mathematical constraints operative throughout science.

There are clearly features of the classical world which challenge the Davis disciplinary regime. As observed by Copeland, some of the earliest (and deepest) thinking on the question of physical incomputability comes from another distinguished source – back in 1970 the mathematician Georg Kreisel was proposing a collision problem related to the 3-body problem, which might result in "an analog computation of a non-recursive function".

One can find detailed accounts of Kreisel's thinking on extensions to the Church-Turing thesis in the section of Odifreddi's first volume of *Classical Recursion Theory* (Odifreddi 1989), and in Odifreddi's article on the topic in his edited volume *Around and About Georg Kreisel*.

Another challenge arises from the growth of chaos theory, dealing with the generation of informational complexity via very simple rules. Features of chaotic situations include the iteration of simple rules, nonlinearity, and the sort of sensitivity to initial conditions that Lorenz (1963) observed in the development of weather systems. Another feature is the *emergence* of systemic formations, such as the Lorenz attractor, or the strange attractor discovered by Shaw (1981, 1984) in studying the ostensibly very simple chaos of a dripping tap – by varying the

flow to the dripping tap, unpredictable irregularities of intervals between drips were observed, while appropriate plotting of the unpredictable data revealed an interesting 3-dimensional strange attractor.

The special interest of chaos arises not so much from its undeniable novelty of computational character – there are all sorts of explanations of the apparent undeterminacy of outcomes, not usually enlisting incomputability – but from the availability of informative mathematical analogues. It is the link between such structures in nature, and mathematical objects, such as the Mandelbrot and Julia sets, which presents an opportunity of getting closer to a mathematical characterisation of what is happening. At the same time, the mathematical interest and approachability of fractals, with their grounding in the iteration of simple rules paralleling those in nature, makes their computability-theoretic character accessible to serious investigation.

The Mandelbrot set has attracted particular attention from high-profile scientists such as Roger Penrose and Stephen Smale. Its popular appeal is matched by its mathematical interest. As Penrose (1994) puts it in *The Emperor's New Mind*:

> Now we witnessed ... a certain extraordinarily complicated looking set, namely the Mandelbrot set. Although the rules which provide its definition are surprisingly simple, the set itself exhibits an endless variety of highly elaborate structures.

And it is not just the observed patterns which are hard to predict. The computability of the actual point-set is still very much an open problem (despite its incomputability in the Blum–Shub–Smale model Blum et al. (1997) of real computation – see Hertling's (2005) review article).

We saw earlier that it is just the addition of a quantifier to the language used to describe a Turing machine which opens the door to the emergence of incomputability. Looking at the definition of the Mandelbrot set in terms of limiting behaviour of applications of the polynomial rule $z \to z^2 + c$, we immediately get a two-quantifier form for the set. But a little extra work gives the complement of the Mandelbrot set using just one existential quantifier. We need to pursue further the general phenomenon of emergence observed in the above examples, and to relate it to the complexity of language needed to describe it.

8.3 What Is Emergence? – Definability, Nonlocality

Emergence is a much over-worked concept. For example, its perceived potential for undermining determinism makes it specially appealing to those trying to create room for religion in a scientific world. Here is Kauffman (2008, p. 281) making some very grand claims in his recent book *Reinventing the Sacred: A New View of Science, Reason and Religion*:

> We are beyond reductionism: life, agency, meaning, value, and even consciousness and morality almost certainly arose naturally, and the evolution of the biosphere, economy, and human culture are stunningly creative often in ways that cannot be foretold, indeed in ways

> that appear to be partially lawless. The latter challenge to current science is radical. It runs starkly counter to almost 400 years of belief that natural laws will be sufficient to explain what is real anywhere in the universe, a view I have called the Galilean spell. The new view of emergence and ceaseless creativity partially beyond natural law is a truly new scientific worldview in which science itself has limits.

Without saying Kauffman is wrong – his world-view has a lot of appeal – one cannot help but be nervous at such ambitious conclusions based on such a modest grasp of what emergence really is. This is Arkin's (1998, p. 105) comment:

> Emergence is often invoked in an almost mystical sense regarding the capabilities of behavior-based systems. Emergent behavior implies a holistic capability where the sum is considerably greater than its parts. It is true that what occurs in a behavior-based system is often a surprise to the system's designer, but does the surprise come because of a shortcoming of the analysis of the constituent behavioral building blocks and their coordination, or because of something else?

Ronald et al. (1999) have devised a 'Test for Convergence' which usefully clarifies what we expect of an emergent phenomenon. It follows the example of the Turing Test for intelligence machinery in being observer dependent, which solves some problems even if it is not so obviously appropriate. The three criteria they list are (slightly paraphrased):

(1) **Design** The system has been constructed by the designer, by describing local elementary interactions between components (e.g., artificial creatures and elements of the environment) in a language $\mathfrak{L}_1$.
(2) **Observation** The observer is fully aware of the design, but describes global behaviors and properties of the running system, over a period of time, using a language $\mathfrak{L}_2$.
(3) **Surprise** The language of design $\mathfrak{L}_1$ and the language of observation $\mathfrak{L}_2$ are distinct, and the causal link between the elementary interactions programmed in $\mathfrak{L}_1$ and the behaviors observed in $\mathfrak{L}_2$ is non-obvious to the observer – who therefore experiences surprise. In other words, there is a cognitive dissonance between the observer's mental image of the system's design stated in $\mathfrak{L}_1$ and his contemporaneous observation of the system's behavior stated in $\mathfrak{L}_2$.

A useful part of the test is the bringing out of the qualitative difference between the 'design' and the observed 'global behaviours' via the distinction between the languages $\mathfrak{L}_1$ and $\mathfrak{L}_2$ used to describe them.

On the other hand, the parallel with the observer-based Turing Test is weak, with condition (3) of the Emergence Test lacking robustness; how do we evaluate the origin of the observer's 'surprise'? For the Turing Test, the observer's inability to discriminate between the intelligence of machine and human comes with far more weight and relevance. We need to look more closely at the computational content of emergence, with the aim of extracting a clearer "surprise" criterion.

The view we want to pursue is that emergent phenomena not only yield up descriptions, using different language to that used in describing the underlying design; they are actually determined, constrained, *captured* by that which is describable in terms of the basic causal structure.

The intuition that entities exist because of, and according to, mathematical laws, is not new, of course. One can detect it in the words of Leibniz (1999) from 1714 in the *The Monadology*, Section 32:

> . . . there can be found no fact that is true or existent, or any true proposition, without there being a sufficient reason for its being so and not otherwise, although we cannot know these reasons in most cases.

So natural phenomena not only generate descriptions, but arise and derive form from them. So connecting with a useful abstraction, that of mathematical definability – or, more generally, invariance (under the automorphisms of the appropriate structure). And this gives precision to our experience of emergence as a potentially non-algorithmic determinant of events. On the one hand one can attempt to frame criteria for emergence in terms of the complexity of the language used to describe it, and one can also use the known associations between informational and computational complexity to constrain the computability-theoretic character of physical phenomena.

For instance, taking this approach, one might identify the halting set of the Universal Turing Machine as an emergent phenomenon; although it does not have the visual immediacy of the Mandelbrot set, it is incomputable, and that in itself qualifies it as a sufficiently surprising global attribute.

What one would expect from this very clear connection between the underlying basic causal structure (the 'design') and the emergent phenomenon would be a certain level of robustness of the emergence. What one is suggesting, via the association with mathematical definability, is a direct causal relationship between 'design' and emergent phenomenon – and one which is unlike the usual fundamental laws of nature, in that it is more global in respect of the causes it works with – and potentially, with respect to the effects. This is not so surprising from the point of view of carefully delineated experimental contexts, such as that presented by Robert Shaw's dripping tap. More so with the higher-order emergence being called up by Stuart Kauffman. If one goes back to Alexander's (1927) magnum opus from the nineteen-twenties (another theologically inclined writer, one of the British emergentists described by McLaughlin [1992]) one finds the mystery of connection an integral part of the argument.

Anyway, it is just this expected robustness that Martin Nowak identified (as Director of the Program for Evolutionary Dynamics at Harvard University) in emergent aspects of evolution. This is from the interesting of collection of papers from leading scientists brought together in John Brockman's *What We Believe But Cannot Prove* (Brockman 2006):

> I believe the following aspects of evolution to be true, without knowing how to turn them into (respectable) research topics.
> Important steps in evolution are robust. Multicellularity evolved at least ten times. There are several independent origins of eusociality. There were a number of lineages leading from primates to humans. If our ancestors had not evolved language, somebody else would have.

8.4 Is That All There Is? – Turing and the Human Brain

We have kept back our third challenge to Davis Discipline until we were clearer on what we wanted to summon up from it; we are now ready for the complexities of the human mind as case study. It comes with a number of strengths:

- The human mind is very *familiar*, at least to the more self-aware. Experience of its workings is easily got through solving everyday problems, and observing others.
- And the *mechanics of the brain are well-documented*.
- The mind *does not feel, or appear to compute, like a Turing machine* – given the role of creativity, consciousness, intuition.
- The case study is *relevant* – given the importance of copying how humans think for achieving AI, etc. . . . and the intuition that a physical brain reflects processes in the wider universe, so can help with the modelling new aspects of physical computation.

So how do the mind and emergence match up? The surprise criterion is certainly there. Here is a well-known example from Jacques Hadamard's celebrated 1945 study (Hadamard 1945) of *The Psychology of Invention in the Mathematical Field*, based on conversation with Henri Poincaré:

> At first Poincaré attacked [a problem] vainly for a fortnight, attempting to prove there could not be any such function . . . [quoting Poincaré]:
> "Having reached Coutances, we entered an omnibus to go some place or other. At the moment when I put my foot on the step, the idea came to me, without anything in my former thoughts seeming to have paved the way for it . . . I did not verify the idea . . . I went on with a conversation already commenced, but I felt a perfect certainty. On my return to Caen, for conscience sake, I verified the result at my leisure."

Apart from the surprise element, the unexpected arrival of a crucial idea by some unconscious process, there is another important aspect of this story – the *robustness* of the surprise solution to the problem that Poincaré had been stuck on. He could feel enough confidence in his ability to recreate the solution at some later time to be able to carry on a completely unrelated conversation. The idea, it appears, had a memetic quality consistent with the existence of a *representation* of the solution, such as one might expect from an association of emergence with definability.

So much for part (3) of the Emergence Test. But what about the design? One needs to bridge the gap between higher mental functionality and . . . what algorithmic context? One might hope to derive this from existing models of neural functionality. But this is more difficult than one might expect. According to Rodney Brooks in *Nature* in 2001:

> . . . neither AI nor Alife has produced artifacts that could be confused with a living organism for more than an instant.

Another creative participant in the field of AI, Daniel Hillis, Chief Technology Officer of Applied Minds, Inc. (and ex-Vice President, Research and Development at Walt Disney Imagineering), was quoted in April 2001 as doubting whether

design was even sufficient for the building of intelligent machines. Perhaps getting intelligent machines themselves would be via emergence:

> I used to think we'd do it by engineering. Now I believe we'll evolve them. We're likely to make thinking machines before we understand how the mind works, which is kind of backwards.

This is not to say that paradigm-stretching features of connectionist models of computation are lacking. As Smolensky (1988) (recipient of the 2005 David E. Rumelhart Prize) wrote in 1988:

> There is a reasonable chance that connectionist models will lead to the development of new somewhat-general-purpose self-programming, massively parallel analog computers, and a new theory of analog parallel computation: they may possibly even challenge the strong construal of Church's Thesis as the claim that the class of well-defined computations is exhausted by those of Turing machines.

And it is certainly true that connectionist models have come a long way since Turing's (1948) discussion of 'unorganised machines', and McCulloch and Pitts' (1943) early paper on neural nets.

But for Pinker (1997) "... neural networks alone cannot do the job". And focussing on our elusive higher functionality, he points to a "kind of mental fecundity called recursion":

> We humans can take an entire proposition and give it a role in some larger proposition. Then we can take the larger proposition and embed it in a still-larger one. Not only did the baby eat the slug, but the father saw the baby eat the slug, and I wonder whether the father saw the baby eat the slug, the father knows that I wonder whether he saw the baby eat the slug, and I can guess that the father knows that I wonder whether he saw the baby eat the slug, and so on.

Less amusingly, but bringing out even more clearly the role of *recycled* emergence, the neuroscientist Damasio makes a similar point. Here is his nice description of the hierarchical development of a particular instance of consciousness within the brain (or 'organism'), interacting with some external 'object' (Damasio 1999):

> ... both organism and object are mapped as neural patterns, in first-order maps; all of these neural patterns can become images ... The sensorimotor maps pertaining to the object cause changes in the maps pertaining to the organism ... [These] changes ... can be re-represented in yet other maps (second-order maps) which thus represent the relationship of object and organism ... The neural patterns transiently formed in second-order maps can become mental images, no less so than the neural patterns in first-order maps.

The picture is one of *re-representation* of neural patterns formed across some region of the brain, in such a way that they can have a *computational relevance in forming new patterns*. There is a key conception of computational loops incorporating, in a controlled way, these 'second-order' aspects of the computation itself. The exact mechanism for the creation and recycling of emergent outputs is not completely clear. But the actuality of this is substantiated via our mathematical model of the definability of emergent phenomena, whereby new entities are created and defined along with a role in the original structure. It is worth noting in this context that the basic logic underlying natural language, upon which descriptions/definitions are

based, does not have an irreducible, and mysterious special status in our scientific ontology; it arises from the most basic of material algorithms, ones which appear unavoidable in any viable causal context, and derive their position in human discourse via the close relationship (for us) between matter and data.

We are now ready to try and make more explicit our basic computational model. We have talked a lot about the roles of definability and invariance, without placing these notions in a specific setting. Key ingredients to be sought in such a model are those we have been talking about: imaging, parallelism, interconnectivity, and a counterpart to the second-order recursions pointed to above. And the computational content familiar from the material universe should appear explicitly in the model.

Connectionist models are strong on parallelism, interconnectivity, imaging, and can even accommodate recursions – but not in re-integrating the sort of recursions Pinker is describing into the computational process. And echoing Danny Hillis' comment above about the role of design, one may have to look for a model of the fundamental computational structure of the world, without being able to fully model the functionality. Such a model may not provide the design of an artificial brain, but it may help us understand the obstacles to doing that.

8.5 The Extended Turing Model

The theme of computation versus description runs through most of Alan Turing's work, and never more explicitly than in his long, hard-to-read, and immensely influential 1939 article (Turing 1939). An important thread, begun in this paper and running through much of the subsequent history of computability theory, concerns how the computational content of descriptions can be captured hierarchically – but in unpredictable ways.

Turing's approach is largely proof-theoretic, growing out of his interest in Gödel's incompleteness theorem, and what it tells us about the extent of the boundaries of the computable world. Turing shows that despite Gödel's (1931) proof that no consistent first-order theory captures arithmetic, we can hierarchically transcend this barrier, in a quite constructive way – one just iterates the Gödel argument, computably generating new unprovable theorems which are then used to enlarge the theory. One uses computable ordinal notations to iterate this process into the transfinite in a constructive way, thus giving the appearance of computably transcending Gödel's theorem. But a little thought reveals the snag – identifying the route to a new theorem involves using an incomputable oracle, so we avoid the reductionist paradox.

This is how Turing explains what he had done:

> Mathematical reasoning may be regarded . . . as the exercise of a combination of . . . intuition and ingenuity . . . In pre-Gödel times it was thought by some that all the intuitive judgements of mathematics could be replaced by a finite number of . . . rules. The necessity for intuition would then be entirely eliminated. In our discussions, however, we have gone to the opposite extreme and eliminated not intuition but ingenuity, and this in spite of the fact that our aim has been in much the same direction.

So here we have an explanation of why written proofs do not tell us how the proof was discovered. The 'intuition' involved was needed to identify the path to a proof – in the way Poincaré needed it – but having done that by some incomputable process, one immediately has a purely algorithmic demonstration (that is the proof) of why the theorem is true. The result of this process is that one delivers an emergent result into a developing body of mathematics which has a deceptively algorithmic structural appearance.

Having tried unsuccessfully to 'compute the incomputable', Turing introduced a model of natural causality between real data, which could be incomputable. The model – now called an *oracle Turing machine* – was essentially just a Turing machine which could ask questions of an external 'oracle' (usually a set of natural numbers). The number of questions during a particular computation was finite, of course. The result was that instead of getting computable real numbers via the collating of computational outputs of a machine, one now got real numbers computable *relative* to an oracle. Considering the oracles to be inputs, a given machine might capture a particular computable function over the reals, notated as a *Turing functional* from reals to reals. Given the natural form of this quite general notion, it turns out to be sufficient to capture most of the functions one extracts from basic laws of science. For instance, one can easily represent the progress of two given point masses (whose relative states at a given time are represented as a real) according to Newtonian dynamics via a Turing functional. This is not surprising, since such simple basic transformations are routinely captured via functions over the reals which can be computed up to any practicable level of approximation by a real-world computer. Given more point masses, one can still describe the motion *in terms* of that functional, but this does not allow one to extract a new Turing functional to completely express the new causal relationship. Here we have again basic computability leading very quickly via descriptions to a situation with computational content, but not necessarily computable.

But the bottom line is that in 1939 Turing's oracle machines appeared, and that these provided a model of computable content of structures, based on *partial computable* (p.c.) functionals over the reals. This model – the *Turing universe* – was capable of capturing basic computable causal structure in the real world, with the expectation, based on experience, that any incomputable causality would be definable in some natural way from this basic structure.

This extended model of Turing's had a very interesting history. Some of this is described in *The Incomputable Alan Turing* (Cooper 2008). Around 1948 Post (1948) tidied up the model by gathering together computably equivalent reals into equivalence classes called *degrees of unsolvability*, with an ordering induced by that of relative Turing computability. This gave a classification of reals in terms of their relative computability, so giving an informational landscape with a rich structure.

Back in the real world again, we know that we can often describe global relations in terms of well-understood local structure – so capturing the emergence of large-scale formations. We can now formalise this mathematically in terms of definability over structure based on Turing functionals, insofar as we understand the basic causal structure. Again, if one is concerned about the language dependency of the notion

of definability – language is a human construct, and not obviously applicable to the way the universe 'defines' its large scale structure and laws – then one can express things in terms of invariance under automorphisms.

This brings us to *Hartley Rogers' Programme* which (see Rogers 1965) addresses the:

Fundamental Problem *Characterise the Turing invariant relations.*

The intuition is that these relations are key to pinning down how basic laws and entities emerge as mathematical constraints on causal structure.

At one time, it was thought that the structural pathology exhibited by the Turing universe, and the disproportionate technical difficulty of proofs in the area, was evidence of mathematical ugliness, disqualifying the field from serious attention of non-specialists. It is now understood that the richness of Turing structure discovered so far provides the raw material for non-trivially defining a multitude of relations. And that the complexity and pathology of the structure is only what one would expect of something aiming to model global aspects of the real world.

8.6 And a Physics Road Test

The Turing model has considerable explanatory power. In Cooper (to appear) we apply this to the problem of clarifying the connection between the mental and the physical. Here, we focus on very different problems affecting the standard model of particle physics. Concern about the current state of physics is comes from a number of sources. Woit (2006), in the introduction to his 2006 book *Not Even Wrong – The Failure of String Theory and the Continuing Challenge to Unify the Laws of Physics*, describes the situation so:

> By 1973, physicists had in place what was to become a fantastically successful theory of fundamental particles and their interactions, a theory that was soon to acquire the name of the 'standard model'. Since that time, the overwhelming triumph of the standard model has been matched by a similarly overwhelming failure to find any way to make further progress on fundamental questions.

The success he refers to is in terms of practical prediction. The failure in relation to fundamental questions relates to lack of recent progress – the problems themselves have been around in some form or other for a long time. Einstein himself says in his *Autobiographical Notes* (Einstein 1950, p. 63):

> . . . I would like to state a theorem which at present can not be based upon anything more than upon a faith in the simplicity, i.e., intelligibility, of nature . . . nature is so constituted that it is possible logically to lay down such strongly determined laws that within these laws only rationally completely determined constants occur (not constants, therefore, whose numerical value could be changed without destroying the theory) . . .

These may not be quite the same undetermined constants that Peter Woit is pointing to (there are more of them now):

> One way of thinking about what is unsatisfactory about the standard model is that it leaves seventeen non-trivial numbers still to be explained, . . .

But the substance of the complaint is the same; one should not need to adjust elements of the standard model in a seemingly arbitrary way just to get the right answers delivered. The theory should give a complete explanation of the values of constants, etc.

This is what it was hoped string theory would do. In a sense string theory was a departure from the Baconian paradigm, which Einstein himself had initiated, and demonstrated the power of. But things have not worked out well, and as the family of string theories and their offshoots expands, along with the arbitrary choices needed, the argument is that string theory is "the only game in town". One-time string theorist Friedan (2003) is dismissive:

> The longstanding crisis of string theory is its complete failure to explain or predict any large distance physics . . . String theory is incapable of determining the dimension, geometry, particle spectrum and coupling constants of macroscopic spacetime . . . The reliability of string theory cannot be evaluated, much less established. String theory has no credibility as a candidate theory of physics.

Lee Smolin's (2006) 2006 book on *The Trouble With Physics* is another source of dissent. In it he lists "Five Great Problems in Theoretical Physics". What is relevant for us is that each one can be framed as a *problem of definability*:

1. Combine general relativity and quantum theory into a single theory that can claim to be the complete theory of nature.
2. Resolve the problems in the foundations of quantum mechanics
3. The unification of particles and forces problem: Determine whether or not the various particles and forces can be unified in a theory that explains them all as manifestations of a single, fundamental entity.
4. Explain how the values of the free constants in the standard model of physics are chosen in nature.
5. Explain dark matter and dark energy. Or, if they dont exist, determine how and why gravity is modified on large scales.

An indication of the widespread concern about such problems was the 2005 statement from no less than David Gross (co-discoverer of the asymptotic freedom affecting the strong nuclear force), quoted in the Dec. 10, 2005, *New Scientist*, under the heading *Nobel Laureate Admits String Theory Is In Trouble*:

> The state of physics today is like it was when we were mystified by radioactivity . . . They were missing something absolutely fundamental. We are missing perhaps something as profound as they were back then.

So what is it that is 'absolutely fundamental' that is missing? It is worth noting that Smolin's thinking is consistent with our own emphasis on the modelling of basic causal structure. He proclaims that "causality is fundamental". And while pointing to early champions of the role of causality, such as Roger Penrose, Rafael Sorkin, Fay Dowker, and Fotini Markopoulou, he says (Smolin 2006, p. 241):

> It is not only the case that the spacetime geometry determines what the causal relations are. This can be turned around: Causal relations can determine the spacetime geometry . . . Its easy to talk about space or spacetime emerging from something more fundamental, but those who have tried to develop the idea have found it difficult to realize in practice . . . We

> now believe they failed because they ignored the role that causality plays in spacetime. These days, many of us working on quantum gravity believe that causality itself is fundamental – and is thus meaningful even at a level where the notion of space has disappeared.

And we even detect here an implicit searching for a structure in which the definable set of relations on it is rich enough to take in something corresponding to the spacetime geometry we observe.

Of course, from the point of view of Smolin's Great Problem number 2, one might also benefit from a *failure of definability* corresponding to the quantum ambiguity we encounter, and which disappears with the collapse of the wave function during a measurement. Earlier, having noted that quantum uncertainty presented a particularly strong challenge to Davis' reductionist programme, we went on to focus almost entirely on emergence. It is now time to bring quantum phenomena back into the picture. According to our picture, emergence coincides with an assertion of definability in some underlying causal structure. The complexity of the definition gives rise to a related level of surprise and unpredictability.

What we have at the quantum level is something rather different. What is being defined (or not being defined, as the case may be) is *attributes of the basic design*. Following Leibniz, lacking a definition of aspects of a given quantum state, the state has to exhibit whatever it is allowed to. But an intervention involving a measurement or whatever may enrich the context sufficiently to remove these various possibilities, and leave us with a well-defined classical reality. And the process involves a mathematically enforced non-locality, quite in keeping with what is observed. Anyway, the classical level may not so be so much of a surprise to those of us who spend all our time there, but it is nevertheless emergent. What is surprising to us is that there is a level at which not all is unambiguously defined, and the transition between the two. One would also notice that this is a realistic interpretation, achieved without anthropic principles, many-worlds interpretations, or any other level of Max Tegmark's multiverse hierarchy.

Smolin's Great Problem number 1 also raises interesting features. Notice that when we are presented with emergent entities, described in a different language to the underlying design, they may well determine a whole new level of behaviour, complete with their own emergent causal relations. This is a picture familiar which was familiar to the British Emergentists, dealt with in Brian McLaughlin's book McLaughlin (1992) mentioned earlier. They used it to explain the irreducibility of the 'special sciences', postulating a hierarchy with physics at the bottom, followed by chemistry, biology, social science, etc. The emergence, as our model confirms, is irreversible, imposing the irreducibility of say biology to quantum theory – although the British emergentists experienced their heyday before the great quantum discoveries of the late 1920s, and as described in McLaughlin (1992), this was in a sense their undoing.

Now, what would we think of someone who asked for a unified theory of chemistry and biology? It may be that it is equally senseless to be looking for a unified theory of quantum and relativity theory. On the other hand, with the example of the British emergentists who held that the coming together of hydrogen and oxygen to form water was an example of emergence, one can never be quite sure about the extent of application of useful models.

Smolin's Great Problem number 3 is perhaps a little too specific to be obviously within the scope of such a schematic model as we are applying. It may be that there is something basic about the automorphism group of the Turing universe and its corresponding invariant relations which tell us something very relevant about the fundamental structure of the entities making up the universe; we conjectured something of the sort in Haifa back in 1995 (see Cooper 1998). On the other hand, the answer may depend on much more specific considerations arising from physics.

Problem 4 is obviously a question of definability. And so may Problem 5 be, involving levels of failure of definability beyond our observational reach.

What we would look for is solutions to a range of fundamental problems, within a radically deconstructed universe:

- Described in terms of reals ...
- With emergent natural laws based on algorithmic relations between reals
- With emergence described in terms of definability/invariance
- ... with failures of definability modelling quantum ambiguity
- ... which gives rise to new levels of algorithmic structure
- ... and a fragmented scientific enterprise.

What the mathematics can deliver is a causality which is different in nature from that which Newton gave us back at the beginning of the modern scientific era. Alan Guth (the inventor of cosmic inflation) asks in his book Guth (1997) *The Inflationary Universe – The Quest for a New Theory of Cosmic Origins*:

> If the creation of the universe can be described as a quantum process, we would be left with one deep mystery of existence: What is it that determined the laws of physics?

It is important to bring such questions firmly into the scientific domain

References

Alexander S (1927) Space, Time, and Deity. Macmillan, London

Arkin RC (1998) Behaviour-based Robotics. MIT, Cambridge, MA

Bacon F (1901) In: Spedding J, Ellis RL, Heath DD (eds) The Works, vol IV, Longmans, London

Blum L, Cucker F, Shub M, Smale S (1997) Complexity and Real Computation. Springer, New York

Brockman J (ed) (2006) What we believe but cannot prove: Today's leading thinkers on science in the age of certainty. Harper Perennial, New York

Calude CS, Svozil K (2008) Quantum Randomness and Value Indefiniteness. Adv Sci Lett 1:165–168

Cooper SB (1998) Beyond Gödel's Theorem: Turing Nonrigidity Revisited. In: Makowsky JA, Ravve EV (eds) Logic Colloquium '95. In: Proceedings of the Annual European Summer Meeting of the Association of Symbolic Logic, held in Haifa, Israel, August 9–18, 1995, Lecture Notes in Logic, vol 11. Springer, Berlin, pp 44–50

Cooper SB (2008) The Incomputable Alan Turing. In: Proceedings of Turing 2004: A Celebration of His Life and Achievements, electronically published (2008) by the British Computer Society: http://www.bcs.org/server.php?show=nav.9917

Cooper SB From Descartes to Turing: The Computational Content of Supervenience. To appear in *Information and Computation* (editors Mark Burgin and Gordana Dodig-Crnkovic), World Scientific Publishing Co.

Copeland J (1998) Turing's O-machines, Penrose, Searle, and the Brain. Analysis 58:128–138

Damasio AR (1999) The Feeling of What Happens: Body and Emotion in the Making of Consciousness. Harcourt Brace, New York

Davis M (2000) The Universal Computer: The Road from Leibniz to Turing. W.W. Norton, New York

Davis M (2004) The Myth of Hypercomputation. In: Teuscher C (ed) Alan Turing: Life and Legacy of a Great Thinker. Springer, Berlin, pp 195–212

Deutsch D (1985) Quantum Theory, the Church Turing Principle, and the Universal Quantum Computer. Proc Roy Soc A 400:97–117

Einstein A (1950) Out of My Later Years. Philosophical Library, New York

Einstein A (1969) Autobiographical Notes. In: P Schilpp (ed) In: Albert Einstein: Philosopher-scientist. Open Court, La Salle, IL

Friedan D (2003) A Tentative Theory of Large Distance Physics. J High Energy Phys JHEP10:063

Gödel K (1931) Über formalunentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. Monatsh Math Phys 38:173–198; translated in Davis M (ed) In: The Undecidable. Basic Papers on Undecidable Propositions, Unsolvable Problems, and Computable Functions. Raven Press, Hewlitt

Guth AH (1997) The Inflationary Universe – The Quest for a New Theory of Cosmic Origins. Addison-Wesley, New York

Hadamard J (1945) The Psychology of Invention in the Mathematical Field. Princeton University Press, Princeton

Hertling P (2005) Is the Mandelbrot Set Computable? Math Logic Quart 51:5–18

Hodges A (1992) Alan Turing: The Enigma. Vintage, London

Hodges A (2004) What would Alan Turing have done after 1954? In: Teuscher C (ed) Alan Turing: Life and Legacy of a Great Thinker. Springer, Berlin

Kauffman SA (2008) Reinventing the Sacred: A New View of Science, Reason and Religion. Basic Books, New York

de Laplace PS (1951) Essai Philosophique sur les Probabilités. English trans. by F.W. Truscott and F.L. Emory. Dover, New York

Leibniz GW (1999) La Monadologie (1714). English translation by G.M. Ross

Lorenz E (1963) Deterministic Nonperiodic Flow. J Atmos Sci 20:130–141

McCulloch WS, Pitts WH (1943) A Logical Calculus of the Ideas Immanent in Neural Nets. Bulletin of Mathematical Biophysics 5:115–133

McLaughlin BP (1992) The Rise and Fall of British Emergentism. In: Beckermann A, Flohr H, Kim J (eds) Emergence or Reduction? – Essays on the Prospects of Nonreductive Physicalism. de Gruyter, Berlin, pp 49–93. Reprinted in Bedau MA, Humphreys P (eds) Emergence: Contemporary Readings in Philosophy and Science. MIT, Cambridge, MA pp 19–59

Németi I, Andréka H (2006) Can General Relativistic Computers Break the Turing Barrier? In: Beckmann A, Berger U, Löwe B, Tucker JV (eds) Logical Approaches to Computational Barriers, Second Conference on Computability in Europe, CiE 2006, Swansea, UK, June 30–July 5, 2006, Proceedings. Lecture Notes in Computer Science 3988, Springer, Berlin, pp 398–412

Odifreddi P (1989) Classical Recursion Theory, vols I and II. North-Holland/Elsevier, Amsterdam

Pinker S (1997) How the Mind Works. W.W. Norton, New York

Post EL (1948) Degrees of Recursive Unsolvability: Preliminary Report (abstract). Bull Amer Math Soc 54:641–642

Penrose R (1994) The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics. Oxford University Press, Oxford

Rogers H Jr (1965) Some Problems of Definability in Recursive Function Theory. In: Crossley JN (ed) Sets, Models and Recursion Theory. Proceedings of the Summer School in Mathematical Logic and Tenth Logic Colloquium, Leicester, August–September, North-Holland, Amsterdam, pp 183–201

Ronald EMA, Sipper M, Capcarrère MS (1999) Design, Observation, Surprise! A Test of Emergence. Artif Life 5:225–239. Reprinted in (Bedau MA, Humphreys P (eds) Emergence: Contemporary Readings in Philosophy and Science. MIT, Cambridge, MA, pp 287–304

Shaw R (1981) Strange Attractors, Chaotic Behaviour, and Information Flow. Z. Naturforsch, 36A:80–112

Shaw R (1984) The Dripping Faucet as a Model Chaotic System. The Science Frontier Express Series, Aerial Press, Santa Cruz, CA

Smolensky P (1988) On the Proper Treatment of Connectionism. Behavior Brain Sci 11:1–74

Smolin L (2006) The Trouble with Physics: The Rise of String Theory, the Fall of a Science and What Comes Next. Houghton Mifflin, New York

Teuscher C (2004) (ed) Alan Turing: Life and Legacy of a Great Thinker. Springer, Berlin

Turing AM (1936) On Computable Numbers, with an Application to the Entscheidungsproblem. Proc London Math Soc 422:230–265. Reprinted in Turing AM (ed) Collected Works: Mathematical Logic, 18–53

Turing AM (1939) Systems of Logic Based on Ordinals. Proc London Math Soc 45(2):161–228. Reprinted in Turing AM (ed) Collected Works: Mathematical Logic, pp 81–148

Turing AM (1948) Intelligent Machinery. National Physical Laboratory Report. In: Meltzer B, Michie D (eds) Machine Intelligence 5. Edinburgh University Press, Edinburgh, 1969, pp 3–23. Reprinted in Ince DC (ed) A.M. Turing, Collected Works: Mechanical Intelligence. North-Holland, Amsterdam, 1992, pp 107–127

van Leeuwen J, Wiedermann J (2000) The Turing Machine Paradigm in Contemporary Computing. In: Enquist B, Schmidt W (eds) Mathematics Unlimited – 2001 and Beyond, LNCS. Springer, Berlin

White M (1997) Isaac Newton – The Last Sorcerer. Fourth Estate, London

Woit P (2006) Not Even Wrong – The Failure of String Theory and the Continuing Challenge to Unify the Laws of Physics. Jonathan Cape, London

Chapter 9
Computational Models of Measurement and Hempel's Axiomatization

Edwin Beggs, José Félix Costa, and John V. Tucker

9.1 Introduction

We are developing a methodology and mathematical theory to examine how data is represented and computations are performed by physical systems. The research programme is shaped by questions about what can be computed by (i) physical systems in isolation and (ii) physical systems combined with algorithms. The methodology is formulated using five principles that focus on the role of a physical theory in formalising experiments. Our theory for isolated physical systems begins in Beggs and Tucker (2006, 2007, 2008, 2009) and that for physical systems and algorithms begins in Beggs et al. (2008a,b, 2009, submitted). A central technical idea is to use a physical experiment as an oracle to a Turing machine. This changes the nature of oracle queries and introduces new and subtle *protocols* to manage the time taken by queries and tolerances in data exchanges. Typically, we use an experiment $E(x)$ designed to measure a physical quantity represented by a real number x. The oracle is expected to extend the computing power of the Turing machines. For specific experiments, we have characterised the class of sets decidable by these machines

E. Beggs (✉), J.F. Costa, and J.V. Tucker
School of Physical Sciences, Swansea University,
Singleton Park, Swansea, SA2 8PP
Wales, United Kingdom
e-mail: e.j.beggs@swansea.ac.uk;j.v.tucker@swansea.ac.uk

J.F. Costa
Department of Mathematics, Instituto Superior Técnico
Universidade Técnica de Lisboa
Lisboa, Portugal
e-mail: fgc@math.ist.utl.pt
and
Centro de Matemática e Aplicações Fundamentais do Complexo Interdisciplinar
Universidade de Lisboa
and
Centro de Filosofia das Ciências da Universidade de Lisboa
Lisboa, Portugal

A. Carsetti (ed.), *Causality, Meaningful Complexity and Embodied Cognition*,
Theory and Decision Library A 46, DOI 10.1007/978-90-481-3529-5_9,

using non-uniform complexity classes and we have shown that the oracles extend the power of Turing computability substantially.

However, recently in Beggs et al. (submitted, 2009b), we have added a new, sixth principle which changes the perspective of the mathematical theory of Turing machines with physical oracles. Instead of viewing the experiment as an oracle boosting the power of Turing machines, we view the Turing machine as controlling and, indeed, performing the experiment. Specifically, Principle 6 leads us to suppose that:

The Turing machine models a human experimenter conducting the experiment.

The relationship between experimenter and experiment is modelled by the protocols that apply to the oracle queries. In Beggs et al. (submitted) we study in some detail a Newtonian experiment to measure mass, which reveals concepts and properties of wide applicability.

Thus, with Principle 6 of Beggs et al. (submitted, 2009b), we find we are in possession of a fledgling computational model of the process of doing physical experiments and making measurements. The model accommodates

(i) Logical properties of the process of following an experimental procedure, made up of instructions specified by a physical theory;
(ii) Quantitative constraints of precision and error margins and of the cost in time and other resources needed to perform experiments

We have looked at several experiments and the questions arise:

To what extent is our computational model of experimentation general? What is measurement?

In this paper we begin to explore these questions with the help of the philosophy of physics. We relate our computational model to the desiderata of Geroch and Hartle (1986) for an investigation into computable aspects to measurement. We consider the axiomatic theory of measurement established by Carl G. Hempel (1952), and elaborated by Rudolf Carnap (1966), and apply it to our computational models of measurement. *Do our models satisfy Hempel's axioms?* Yes. *Do they reveal new general properties of measurement?* Yes. Indeed, we show that the models uncover some shortcomings in Hempel's characterisation, which we repair with new axioms.

Hempel's theory is based on two predicates intended to make *comparisons between some physical attribute*: think of an equivalence and ordering applied to some attribute of a set of objects. On measuring the attribute using real numbers, the comparison predicates are mirrored by the standard predicates $=$ and $<$, which are undecidable on computable real numbers. This is more than an inconvenience for an axiomatic theory of measurement, where tolerances and accuracy are central concepts. This undecidability can be ameliorated in different ways. We introduce the operational concept of computational resources, specifically *time*, into Hempel's axioms; the resulting axiomatisation we believe to be new. The idea of considering time as a cost in deciding the equality of measurements is suggested by our previous technical work on the model (e.g., see Beggs et al. (2008a, 2009a)).

Let us consider the impact of adding time to Hempel's view of measurement. Hempel uses the experience of measuring mass with a balance scale to introduce his

axioms. The notions of two objects weighing the same, or one weighing less than the other, are quite intuitive. However, as the masses of the two objects approach one another, the measurement becomes more and more troublesome, due to friction and nature of the balance: *two objects in the pans may be in equilibrium one day but are found no longer to be in equilibrium the next*. Hempel (1952), end of Chapter 10 and middle of Chapter 11, develops the following argument:

Hempel 1. The most important — and perhaps the only — type of fundamental measurement used in the physical sciences is illustrated by the fundamental measurement of mass, length, temporal duration, and a number of other quantities. It consists of two steps: first, the specification of a comparative concept, which determines a nonmetrical order; and, second, the metrization of that order by the introduction of numerical values [. . .] Now we return to our illustration [of measuring mass]. In formulating specific criteria for this case, we will use abbreviatory phrases: of any two objects, x and y [. . .] we will say that x *outweights* y if, when the objects are placed into opposite pans of a balance in a vacuum, x sinks and y rises; and we will say that x *balances* y if under the conditions described the balance remains in equilibrium.

Hempel is aware of the need of improving accuracy to define metrical properties for the mass concept (hence the vacuum[1]). However, there is no awareness, either in Hempel's or in Carnap's theories, that the *time to run an experiment* is actually a fundamental concept when allocating numerical values to attributes in a consistent way. Hempel is conscious of this limitation of his axiomatization of measurement of quantities that take real values, or even rational values. In a footnote, he declares the following:

Hempel 2. This account of the fundamental measurement of mass is necessarily schematized with a view to exhibiting the basic logical structure of the process. We have to disregard such considerations as that the equilibrium of a balance carrying a load in each pan may not be disturbed by placing into one of the pans an additional object which is relatively light but whose mass is ascertainable by fundamental measurement. This means that fundamental measurement does not assign exactly one number to every object [. . .]

Measurement is a mapping from objects to numbers. By introducing time in Hempel's axiomatization, we establish a more accurate semantical basis for these maps.

The structure of the paper is this. In Section 9.2, we review the Hempel–Carnap theory of measurement. In Section 9.3, we recall the computational model of an experiment to measure mass from Beggs et al. (submitted) Such computational models are *gedankenexperimente*. We review the ideas of Geroch and Hartle (1986) in Section 9.4. In Section 9.5 we look at mass in Newtonian dynamics. In Section 9.6, we present a new axiomatization of measurement by generalising Hempel's axioms in order to introduce the *time taken by a measurement process*. This is, indeed, a generalisation, from which we can recover the old axiomatization. Finally, in Section 9.7, we show how the computational perspective implies that not all quantities are measurable.

[1] Why should the balance be in a vacuum? It is not because of friction. It is because there are substances in the atmosphere that have "negative weight" such as hydrogen and helium.

9.2 Theory of Measurement

9.2.1 The Three Concepts of Measurement

According to Hempel (1952) and Carnap (1966), the construction of a quantitative concept, based on measurement, involves three phases. For illustration, we use the quantitative concept of mass as measured by the balance.

The Classificatory Phase Classification is based upon some primitive method of sorting concepts into groups according to similarities. What aspect is chosen is termed an *attribute*. Classification is essentially subjective. To make finer classifications, attention must be paid to details of the objects being classified, which demands more time of the taxonomist.

The Comparative Phase The attributes that define the classification need to be compared. A comparative concept is something observable of attributes and what is observed is termed an *event*. It constitutes the basis for a quantitative concept; although the comparative concept seems to be unique, the quantitative one can be understood and axiomatized in different ways.

For the concept of weight, we introduce the comparative concepts of *lighter, heavier*, and *equal* in weight. These concepts have an empirical procedure by which we can take any pair of objects and observe.

If the two objects balance, they are of equal weight. If the objects do not balance, the object on the pan that rises is lighter than the object on the pan that sinks.

Let these observable events define the relations of "equality" $\mathcal{E}$ and "less than" $\mathcal{L}$, respectively.

The Quantitative Phase The attributes we wish to compare are assigned numerical values by a map M from objects to numbers. Carnap (1966), says:

Carnap 1. The qualitative language is restricted to predicates (for example, "grass is green"), while the quantitative language introduces what are called functor symbols, that is, symbols for functions that have numerical values. This is important, because the view is widespread, especially among philosophers, that there are two kinds of features in nature, the qualitative and the quantitative. Some philosophers maintain that modern science, because it restricts its attention more and more to quantitative features, neglects the qualitative aspects of nature and so gives an entirely distorted picture of the world. This view is entirely wrong, and we can see that it is wrong if we introduce the distinction at the proper place. When we look at nature, we cannot ask: "Are these phenomena that I see here qualitative phenomena or quantitative?" That is not the right question. If someone describes these phenomena in certain terms, defining those terms and giving us rules for their use, then we can ask: "Are these the terms of a quantitative language, or are they the terms of a prequantitative, qualitative language?".

The measurements must preserve the comparisons. For mass, we need to define the relations between the events associated with the balance scale and the map M: for any objects a and b, (i) if $a\mathcal{E}b$ then $M(a) = M(b)$ and (ii) if $a\mathcal{L}b$ then $M(a) < M(b)$.

9.2.2 The Axiomatization of Measurement

In Hempel's book (1952), Part III, Chapters 9 to 13, we find an axiomatization of measurement in Physics and other empirical sciences; a discussion of Hempel's axiomatization is Carnap (1966).

Consider a class $\mathcal{O}$ of physical objects endowed with some attribute (such as *mass, electric charge*, or *temperature*, etc.). A measurement of an attribute in the sense of Hempel is a map $M : \mathcal{O} \to N$, where N is a number system such as the *integers* $\mathbb{Z}$, *rationals* $\mathbb{Q}$, or *reals* $\mathbb{R}$. For definiteness, we will choose $M : \mathcal{O} \to \mathbb{R}$.

Hempel's axiomatization of measurement establishes an ordering of the objects of $\mathcal{O}$. To have a measurement, we need an *instrument* or *experimental apparatus*, and observations defining events that implement physically the two special comparative predicates $\mathcal{E}$ and $\mathcal{L}$ over the set $\mathcal{O}$:

1. If objects a and b are identical in the observed attribute, then $a\mathcal{E}b$ is the case.
2. If object a is less than object b in the observed attribute, then $a\mathcal{L}b$ is the case.

The experimental apparatus works with the objects from $\mathcal{O}$, allowing the experimenter to establish a comparison of values of a given attribute.

Definition 1. *Given two binary relations $\mathcal{E}$ and $\mathcal{L}$, $\mathcal{L}$ is* $\mathcal{E}$-irreflexive *if, for all objects a and b in $\mathcal{O}$, if $a\mathcal{E}b$ is the case, then $a\mathcal{L}b$ does not hold.*

Definition 2. *Given two binary relations $\mathcal{E}$ and $\mathcal{L}$, $\mathcal{L}$ is* $\mathcal{E}$-connected *if, for all objects a and b in $\mathcal{O}$, if $a\mathcal{E}b$ does not hold, then $a\mathcal{L}b$ or $b\mathcal{L}a$ is the case.*

Definition 3. *Two binary relations $\mathcal{E}$ and $\mathcal{L}$ determine a* comparative concept, *or a* quasi-series, *for the elements of $\mathcal{O}$, if $\mathcal{E}$ is an equivalence relation and $\mathcal{L}$ is transitive, $\mathcal{E}$-irreflexive, and $\mathcal{E}$-connected.*

Let $\mathbb{E}$ be the set of observable events. Let $\mathcal{I} : \mathcal{O} \times \mathcal{O} \to \mathbb{E}$ be an *abstract implementation map*. In Hempel's examples in Hempel (1952), the set $\mathbb{E}$ of events can be reduced to the bipolar set $\{-1, 0, +1\}$: the outcome of each experiment with objects a and b will tell us that either $a\mathcal{L}b$ (the event denoted by -1), or $a\mathcal{E}b$ (the event denoted by 0), or $b\mathcal{L}a$ (the event denoted by $+1$). The experimenter has to identify which physical events are to be denoted by $-1, 0, +1$.

In the example of the balance, if we put objects a and b in the left pan and the right pan, respectively. Event -1: the left pan rises and the right pan sinks – $a\mathcal{L}b$ is the case. Event $+1$: the left pan sinks and the right pan rises – $b\mathcal{L}a$ is the case. Event 0 (or the non-event): the balance remains in equilibrium – $a\mathcal{E}b$ is the case.

A careful reading of Chapter 12 of Hempel (1952), on the notion of *fundamental measurement*, introduced by Campbell (1928), we find that a detailed sub-structure of $\mathcal{O}$ can be identified, consisting of a standard object, called the *unit mass*, together with its multiples and submultiples: this substructure we call the *toolbox of standards*.[2] By reducing the number of axioms in Hempel's theory (namely, removing

[2] This is done by considering a semigroup of objects $\mathcal{O} = \langle \mathcal{O}, \circ; 1 \rangle$, with the distinguished element 1 called the unit, and some internal structure to generate fractions and multiples of the unit.

the axioms of extensivity, developed by Suppes (1951)), we can provide a first workable definition of *measurement map* for a set of objects:

Definition 4. *Let* $\mathcal{E}$ *and* $\mathcal{L}$ *be comparative relations on the set* $\mathcal{O}$ *of objects* (*Definition 3*). *Suppose there exists an* experimental apparatus *to witness these relations and let* $\mathbb{E}$ *be a set of elements denoting* physical events.

Suppose $\{-1, 0, +1\} \subseteq \mathbb{E}$ *and whenever the experiment is done with arbitrary objects* $a, b \in \mathcal{O}$, *if the outcome is event* -1, *then* $a\mathcal{L}b$ *is the case, if the outcome event is* $+1$, *then* $b\mathcal{L}a$ *is the case, and if the outcome is* 0, *then* $a\mathcal{E}b$ *is the case.*

Then the map $M : \mathcal{O} \to \mathbb{R}$ *is a* measurement map *if*

Axiom 1. *If* $a\mathcal{E}b$, *then* $M(a) = M(b)$.

Axiom 2. *If* $a\mathcal{L}b$, *then* $M(a) < M(b)$.

We think this is a good definition capturing Hempel's construction of a quantitative concept from a comparative concept, as Hempel (1952) suggests:

Hempel 3. Any function M which assigns to every element x of $\mathcal{O}$ exactly one real-number value, $M(x)$, will be said to constitute a *quantitative* or *metrical concept*, or briefly a *quantity* (with the domain of application $\mathcal{O}$); and if M meets the conditions just specified, we will say that it *accords with* the given quasi-series.

The axiomatization allows to prove simple results such as

Proposition 1. *For all a, b in* $\mathcal{O}$, *one, and only one, of the following statements holds:* (*a*) $a\mathcal{E}b$, (*b*) $a\mathcal{L}b$, *or* (*c*) $b\mathcal{L}a$.

Proof. First, we show that at least one of the three conditions hold. Suppose $a\mathcal{E}b$. Then we are done. Suppose that $a\mathcal{E}b$ is not the case. Since $\mathcal{L}$ is $\mathcal{E}$-connected, either $a\mathcal{L}b$ or $b\mathcal{L}a$. Thus, one of the three relations holds. We show that only one can hold.

a. Suppose that $a\mathcal{E}b$. Since $\mathcal{L}$ is $\mathcal{E}$-irreflexive, $a\mathcal{L}b$ is not the case. Since $\mathcal{E}$ is an equivalence, $b\mathcal{E}a$ is also the case. Again, since $\mathcal{L}$ is $\mathcal{E}$-irreflexive, $b\mathcal{L}a$ is not the case.
b. Suppose that $a\mathcal{L}b$. Since $a\mathcal{E}a$, we can not have $b\mathcal{L}a$, because by transitivity we would get $a\mathcal{L}a$ and $\mathcal{L}$ is $\mathcal{E}$-irreflexive. We can not also have $a\mathcal{E}b$, since $\mathcal{E}$-irreflexivity implies that $a\mathcal{L}b$, a contradiction.
c. The argument is the same as b. □

The converse of the axioms in Definition 4 hold.

Proposition 2.

$$\textit{If } M(a) = M(b), \textit{ then } a\mathcal{E}b. \tag{9.1}$$

$$\textit{If } M(a) < M(b), \textit{ then } a\mathcal{L}b. \tag{9.2}$$

Proof. We argue by contraposition. (1) Suppose that $a\mathcal{E}b$ is not the case. Then we have either $a\mathcal{L}b$ or $b\mathcal{L}a$, that is either $M(a) < M(b)$ or $M(b) < M(a)$, by

definition. It follows that $M(a) \neq M(b)$. (2) Suppose now that $a \mathcal{L} b$ is not the case. Then either $a \mathcal{E} b$ or $b \mathcal{L} a$, that is either $M(a) = M(b)$ or $M(b) < M(a)$. □

Proposition 3.

$$\forall x \, \forall y \, (x \mathcal{E} y \Leftrightarrow \forall u \, ((x \mathcal{L} u \Leftrightarrow y \mathcal{L} u) \wedge (u \mathcal{L} x \Leftrightarrow u \mathcal{L} y))) \, . \tag{9.3}$$

$$\forall x \, \forall y \, \forall z \, ((x \mathcal{E} y \wedge y \mathcal{L} z) \Rightarrow x \mathcal{L} z) \, . \tag{9.4}$$

Axioms 1 and 2 in Definition 4, are not far from Hempel's own theory as stated in Hempel (1952):

Hempel 4. Let $\mathcal{E}$ and $\mathcal{L}$ be two relations which determine a quasi-serial order for a class $\mathcal{O}$. We will say that this order has been metricized if criteria have been specified which assign to each element x of $\mathcal{O}$ exactly one real number, $M(x)$, in such a manner that the following conditions are satisfied for all elements x, y of $\mathcal{O}$: [follows Axioms 1 and 2].

This (first) axiomatization of measurement[3] is troubled by the undecidability of $=$ for quantities ranging over the real numbers. In Section 9.6, we will show how to generalize Hempel's axioms in order to have decidable comparison relations, by the introduction of time complexity to an experiment.

9.3 The Collider Experiment

In this section we describe an example of an experiment about elastic collision for the purpose of measuring the unknown (inertial) mass of a particle. The experiment is conducted exactly as described in Beggs et al. (submitted). This type of experiment to measure mass was and still is at the heart of mechanics. A generalization of the collision experiment can be used to measure the mass of a star or of a planet, measures that cannot be done with the balance scale.

9.3.1 Theory

As a *gedankenexperiment*, we consider a very simple situation at the limit of physical reality: a one dimensional elastic collision of two particles. The elastic collision between two particles on a line is dictated by two basic laws of Physics: the conservation of *linear momentum* and the conservation of *kinetic energy*, both of which can be derived from Newtonian laws of dynamics (see Section 9.5).

[3] There can be further structure for the map M, e.g., depending on the fact that the attribute considered is either *extensive* (e.g., mass) or *intensive* (e.g., temperature).

9.3.2 Experiment

In the one dimensional collision the center of mass of the two particles are in the same line of motion. Let m and μ be the masses of the two particles. We will assume that the particle of "unknown" mass μ is always at rest before the collision, and that the "proof" particle of mass m is projected along the line towards the particle of unknown mass μ with speed $u = 1.0\ (\pm\varepsilon)\ \mathrm{ms}^{-1}$, e.g. with $0 \leq \varepsilon \leq 0.1$.[4] After the collision the particle of mass m acquires the speed v_m and the particle of mass μ is projected forward with speed v_μ.

By the conservation of momentum and kinetic energy, the collision is described by the equations:

$$mu = mv_m + \mu v_\mu, \tag{9.5}$$

$$\frac{1}{2}mu^2 = \frac{1}{2}mv_m^2 + \frac{1}{2}\mu v_\mu^2, \tag{9.6}$$

that can be solved for v_m and v_μ:

$$v_m = \frac{m-\mu}{m+\mu}u, \tag{9.7}$$

$$v_\mu = \frac{2m}{m+\mu}u. \tag{9.8}$$

From these formulae we see that after a collision:

a. if $m < \mu$, then the proof particle move backwards after the collision.
b. if $m > \mu$, then the proof particle will move forward.
c. if $m = \mu$, then the proof particle of mass m comes to rest and the particle of unknown mass μ is projected forward with the previous value of the speed of the proof particle.

This experiment can be designed to measure the unknown mass μ, using proof particles of known mass m projected at the same speed u.

We establish the convention that the particle of unknown mass is placed at the origin of coordinates and points $P^- \equiv -1m$ and $P^+ \equiv +1m$ are the *flags* of the experimenter's observations: when the proof particle is seen crossing the points P^- or P^+ the experiment terminates. If the proof mass crosses the flag P^- then we have $m < \mu$ (as depicted in Fig. 9.1), and if it crosses the flag P^+, we have $m > \mu$.

For this experiment there are various facts that are largely irrelevant, or where errors can be tolerated. These include the (finite) distance between the two flags, the precision of the placement of the flags, the error in placing the particle of the unknown mass at the origin (let us say approximately $0\,m$), and the initial speed of the

[4] This error margin in the initial speed of the proof particle of mass m means that precision in speed does not matter for this experiment.

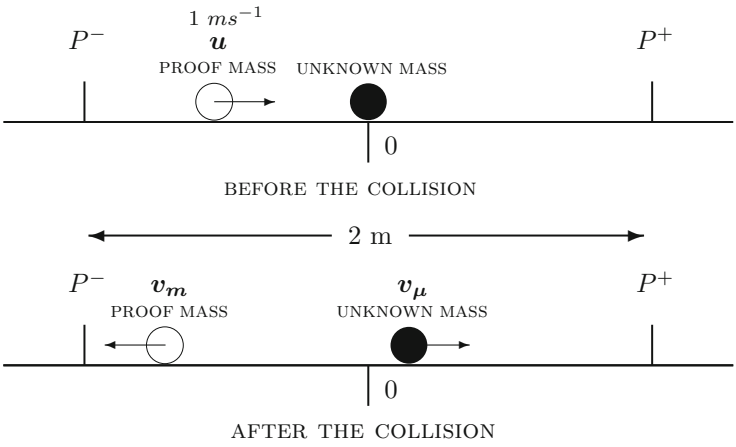


Fig. 9.1 Collider machine experiment

proof particle (let us say approximately 1 ms^{-1}). Note that the observed velocities of the particles after the collision, after crossing one or both the flags, are irrelevant.

However quantities and facts that are relevant include: the one dimensional character; that the masses of the unknown particles are continuous variable in the range (0,1); that the particle of unknown mass μ is at rest; and that the collisions are elastic.

Looking closer to the experiment, we however find an experimental barrier: the time for the proof particle crossing the distance of 1 m after the collision is given by

$$t_{exp} = \frac{1}{u}\left|\frac{m+\mu}{m-\mu}\right|, \tag{9.9}$$

that, for the values we will take of the masses and initial speed, is of the order of

$$\frac{A}{|m-\mu|} \leq t_{exp} \leq \frac{B}{|m-\mu|}, \tag{9.10}$$

for some constants A and B.

9.3.3 CME as Oracle

In the *shooting state* the machine prepares and fires a proof particle of mass m as detailed above. The experiment continues until the proof particle crosses one of the flags $P^{\pm}$, and then returns a state $m < \mu$ or $m > \mu$ to the Turing machine.

The Turing machine is connected to the collider experiment CME in the same way as it would be connected to an oracle: we replace the query state with a *shooting*

state (q_s), the "yes" state with a *lesser state* (q_l), and the "no" state with a *greater state* (q_g). The resulting computational device is called the (*analogue-digital*) *collider machine experiment*.

In order to carry out an experiment, the machine will write a word z in the query tape and enter the shooting state. The word z codes for a dyadic rational mass m of the "proof" particle. In the *shooting state* the machine prepares and fires a proof particle of mass m as detailed above. The experiment continues until the proof particle crosses one of the flags $P^{\pm}$, and then returns a state $m < \mu$ or $m > \mu$ to the Turing machine.

Technically, this word z will either be "1", or a binary word beginning with 0. We will use y ambiguously to denote both a word $y_1 \cdots y_n \in \{1\} \cup \{0s \colon s \in \{0,1\}^*\}$ and the corresponding dyadic rational $\sum_{i=1}^{n} 2^{-i+1} y_i \in [0,1]$. In this case, we write $|y|$ to denote n, i.e., the size of $y_1 \cdots y_n$.

Consider the precision of the experiment. When measuring the output state the situation is simple: either the proof particle of mass m crosses P^- or it crosses P^+ (or, after some timeout, no proof particle is detected). Errors in observation do not arise. There are different postulates for the precision of the experiment, and we list some in order of decreasing strength:

Definition 5. *The CME is error free if the mass of proof particle can be set exactly to any given dyadic rational number. The CME is error prone with arbitrary precision if the mass of proof particle can be set only to within a non-zero, but arbitrarily small, dyadic precision. The CME is error prone with fixed precision if there is a value $\varepsilon > 0$ such that the mass of proof particle can be set only to within a given precision ε.*

9.3.4 Bisection Algorithm

Now we can describe the algorithm in full detail. Let $T : \mathbb{N} \to \mathbb{N}$ be the time given for the experiment to take place as a function (total map) of the size of the sequence of bits setting the value of the mass of the proof particle. The function T can be seen as a *schedule*, i.e., in each experiment, in order to read the $|m|$-th bit of the mass μ, $T(|m|)$ gives *the amount of time steps that the experimenter is prepared to wait* until resuming the experimental conditions. The function T can either be a computable function or a non-computable function of its argument.

After setting the mass m, the CME will fire a proof particle of mass m, wait $T(|m|)$ time units, and then check if the particle crossed one of the flags. If the particle crossed the flag P^-, then the Turing machine computation will be resumed in the state q_l. If the particle crossed the flag P^+, then the Turing machine computation will be resumed in the state q_g. Perhaps, after time $T(|m|)$, no proof particle is detected.

Bisection(t) – THE BISECTION ALGORITHM: A PROCEDURE TO READ THE FIRST n BITS OF A UNKNOWN MASS μ

1. **input** n – required precision coded by the number of places to the right of the left leading 0;
2. $m_1 := 0,\ m_2 := 1,\ m := 0$ – initial values with no physical significance; note $|m_1| = 0,\ |m_2| = 1$, and $|m| = 0$;
3. **while** $|m| \leq n$ **do**
   (a) $m := \frac{m_1+m_2}{2}$;
   (b) place the particle of unknown mass $\mu \in [0, 1]$ at the origin;
   (c) project proof particle of mass m to collide with particle of unknown mass;
   (d) **if** proof particle crosses the flag P^- in time $T(|m|)$ **then** $m_1 := m$; append 1; – it is known that $\mu \in]m, m_2[$;
   (e) **if** proof particle crosses the flag P^+ in time $T(|m|)$ **then** $m_2 := m$; append 0; – it is known that $\mu \in]m_1, m[$;
   (f) **if** no particle crosses the flags in time $T(|m|)$ **then** return time out;
4. **end while**;
5. **output** dyadic rational denoted by m.

The bisection method applies to each type of precision.

9.3.5 Notions of Measurable

Definition 6. *A mass μ is said to be* measurable *if there exists a schedule T such that the digits of μ can be computed by performing the collision experiment repeatedly. Otherwise, the mass is said to be* non-measurable.

Definition 7. *A mass μ is said to be* effectively measurable *if there exists a computable schedule T such that the digits of μ can be computed by performing the collision experiment repeatedly. Otherwise, the mass is said to be* effectively non-measurable.

To measure time we need to make step counting and time explicit inside the machine. To introduce a *system clock* as part of the Turing machine we can employ the concept of a *time constructible function*, introduced by Hartmanis in 1965.

Definition 8. *A total function $f : \mathbb{N} \to \mathbb{N}$ is said to be time constructible if there is a Turing machine $\mathcal{M}$ such that, for all $n \in \mathbb{N}$ and all inputs of size n, $\mathcal{M}$ halts in exactly $f(n)$ steps.*

Definition 9. *A mass μ is said to be* feasible *if there exists a time constructible computable schedule T such that the digits of μ can be computed by performing the collision experiment repeatedly. Otherwise, the mass is said to be* non-feasible.

9.3.6 Notions of Computation

Definition 10. *An* error free analogue-digital collider machine *is a Turing machine connected to an error prone CME. In a similar way, we define an* error prone analogue-digital collider machine with arbitrary precision, *and an* error prone analogue-digital collider machine with fixed precision.

If an error prone analogue-digital collider machine, with unknown mass $\mu \in (0, 1)$, is triggered by the proof particle with dyadic rational mass $z \in [0, 1]$, then we are certain that the computation will be resumed in the state ql if $m < \mu$, and that it will be resumed in the state q_g when $m > \mu$. We define the following decision criteria:

Definition 11. *Let $A \subseteq \Sigma^*$ be a set of words over Σ. We say that an error free analogue-digital collider machine $\mathcal{M}$* decides *A if there exists a time constructible schedule t to operate the coupled CME and an oracle μ such that, for every input $w \in \Sigma^*$, w is accepted if $w \in A$ and rejected when $w \notin A$. We say that $\mathcal{M}$* decides *A* in polynomial time, *if $\mathcal{M}$ decides A, and there is a polynomial p such that, for every $w \in \Sigma^*$, the number of steps of the computation is bounded by $p(|w|)$.*

Definition 12. *Let $A \subseteq \Sigma^*$ be a set of words over Σ. We say that an error prone analogue-digital collider machine $\mathcal{M}$* decides *A if there exists a time constructible schedule t to operate the coupled CME with a given oracle μ and a number $\gamma < \frac{1}{2}$, such that the error probability of $\mathcal{M}$ for any input w is smaller than γ. We call* correct *to those computations which correctly accept or reject the input. We say that $\mathcal{M}$* decides *A* in polynomial time, *if $\mathcal{M}$ decides A, and there is a polynomial p such that, for every input $w \in \Sigma^*$, the number of steps in every computation of $\mathcal{M}$ on w is bounded by $p(|w|)$.*

We can end this section with some results about questions that are experimentally undecidable:

Proposition 4. *That the proof mass coincides with the given unknown mass cannot be established experimentally in finite time by the CME.*

Proof. According to Eq. 9.10, as $m \to \mu$ through the bisection method, the time the experimenter has to wait goes to infinity, $t_{exp} \to +\infty$. If the two masses coincide, then the experimenter will never know. □

As a trivial consequence of this statement we have the folowing theorem.

Proposition 5. *To know if the unknown mass is a dyadic rational cannot be established experimentally in finite time by the CME.*

And, finally, one important statement to keep in memory for the sections to follow, and its fundamental consequence.

Proposition 6. *At each stage of the bisection algorithm, the lower bounds on the time of a single experiment with the CME are exponential in the size of the mass of the proof particle.*

Proof. We know that the time taken by a single experiment is given by Eq. 9.10 at step n with $|m| = n$. Thus μ has a pattern of the form $\mu = m \pm m' \times 2^{-n'-1}$, with $m' \in [0, 1]$ and $n' > n$, and t_{exp} has a pattern of the form

$$t_{exp} \sim \frac{K}{\left|m - \left(m \pm m' \times 2^{-n'-1}\right)\right|},$$

that is,[5]

$$t_{exp} \sim \frac{K}{\left|\pm m' \times 2^{-n'-1}\right|} \in \Omega\left(2^n\right),$$

Thus, we have the following consequence: □

Proposition 7. *The protocol that processes queries between a Turing machine and the collider takes time that is at least exponential in the size of the mass of the proof particle specified by the queries.*

9.4 Geroch–Hartle on Computability and Measurement

Let us consider the reflections of physicists Geroch and Hartle on computability and measurement (Geroch and Hartle 1986). Several of their speculations and questions are analysed formally in our theory.

Geroch and Hartle start by considering the concept of *measurable number* in contrast to the concept of *computable number*:

Geroch–Hartle 1. We propose, in parallel with the notion of a computable number in mathematics, that of a measurable number in a physical theory. The question of whether there exists an algorithm for implementing a theory may then be formulated more precisely as the question of whether the measurable numbers of the theory are computable.

Then they add some considerations on numbers being measurable and/or computable:

Geroch–Hartle 2. We argue that the measurable numbers are in fact computable in the familiar theories of physics, but there is no reason why this need be the case in order that a theory have predictive power. Indeed, in some recent formulations of quantum gravity as a sum over histories, there are candidates for numbers that are measurable but not computable.

They introduce the notion of a technician measuring physical variables:

Geroch–Hartle 3. Regard number w as measurable if there exists a finite set of instructions for performing an experiment such that a technician, given an abundance of unprepared raw materials and an allowed error ε, is able by following those instructions to perform the experiment, yielding ultimately a rational number within ε of w.

[5] Let f and g be total maps with signature $\mathbb{N} \to \mathbb{N}$. We say that $f \in \Omega(g)$ if there exists a constant $k \in \mathbb{R}$ such that, for an infinite number of values of $n \in \mathbb{N}$, $f(n) > kg(n)$.

The accuracy ε is to be understood as arbitrarily small. The technician and set of instructions, together with some memory to take account of intermediate calculations, we replace by a *Turing machine*. In our model of measurement embodied in Principle 6, the Turing machine *represents formally* the physicist or the experimenter. Thus, we propose the assumption:

Thesis 1. *The experimenter following his or her instructions is modelled by a Turing machine. The measuring process is controlled by an algorithm that runs on the machine, generating the atomic instructions, specified by theory $\mathcal{T}$, to be performed at each step of the experimental procedure.*

This postulate says that the experimenter cannot escape *the logic of following a set of rules* as formalised by computability theory; and, of course, that the logic of experimental procedures can be captured completely by a Turing machine.

A point not considered in Geroch and Hartle (1986) is that not all measurements are possible. Assuming the physicist to be a Turing machine, then the limits of Turing machine computation can determine limits on measurements and, therefore, on the nature of physical experiments.

As we will see in Section 9.6, our work makes the concept of *measurable* as precise as the concept of *computable*. Now this was *not* the intention of Geroch and Hartle (1986):

Geroch–Hartle 4. "Measurable" is analogous to, although of course much less precise than, "computable". The technician is analogous to the computer, the instructions to the computer program, the "abundance of unprepared raw materials" to the infinite number of memory locations, initially blank. Indeed, one can think of the measurable numbers as those that are "computable" using an analog, rather than digital, computer.

Geroch and Hartle stress need for a theory to specify a *gedanken experiment* as follows:

Geroch–Hartle 5. The notion "measurable" involves a mix of natural phenomena and the theory by which we describe those phenomena. Imagine that one had access to experiments in the physical world, but lacked any physical theory whatsoever. Then *no* number w could be shown to be measurable, for, to demonstrate experimentally that a given instruction set shows w measurable would require repeating the experiment an infinite number of times, for a succession of εs approaching zero. One could not even demonstrate that a given instruction set shows measurability of any number at all, for it could turn out that, as ε is made smaller, the resulting sequence of experimentally determined rationals simply fails to converge. It is only a *theory* that can guarantee otherwise. The situation is analogous to that of trying to demonstrate that a given Fortran program shows some number to be computable. There is no general algorithm for deciding this. In particular, it would not do merely to run the program for a few selected values of ε.

Now, how does the Turing machine communicate with Nature? We believe that this interaction is captured by the concept of the continuing evolution of a physical experiment acting as an oracle.

Thesis 2. *The measurement apparatus is taken to be an oracle to a Turing machine. The interaction is achieved through a protocol which counts time. After each consultation, the oracle may provide one bit of the measurement. This bit also provides the necessary information to the machine to proceed with the experimental procedure.*

Geroch and Hartle argue that *every computable number is measurable*. A few paragraphs further on, Geroch and Hartle provide the flavour of a *proof*. This proof is given to the reader by the following:

Geroch–Hartle 6. This is easy to see: Let the instructions direct that the raw materials be assembled into a computer, and that a certain Fortran program – one specified in the instructions – be run on that computer. That is, every digital computer is at heart an analog computer.

Then the authors ask the following question:

Geroch–Hartle 7. We now ask whether, conversely, every measurable number is computable – or, in more detail, whether current physical theories are such that their measurable numbers are computable. This question must be asked with care.

Actually, the question received a very careful answer in our Beggs and Tucker (2007): *the experiment SME demonstrates that there are numbers that are measurable in Newtonian dynamics but that are not computable*.

9.5 The Laws of Dynamics

In this section we explain how the collider experiment lies at the heart of measuring masses in Classical Mechanics. Our aim is to define formally the measurement function for (inertial) mass from Newtonian dynamics.

First Law The first law of Newton establishes that *a particle not subjected to a net force will move in a uniform motion in a straight line*. Since the motion of a particle has to be specified with respect to a particular reference frame, the content of the first law can only be understood if such a reference frame is provided. Also, looking at the statement of the first law, we see that the concept of *force* was not yet defined. The first law should be regarded in the following way: in a region of space containing the particle, far away from all other matter, we can always define a reference frame with respect to which that particle will move in a uniform motion in a straight line. Such a reference frame is the *inertial reference frame*; an example is that of the stars – Kepler's reference frame.[6]

Second Law Having found a inertial reference frame, the departure from a uniform motion in a straight line is "measured" by the kinematic concept of acceleration. The departure from a constant speed in a straight line should be due to a *force* that is impressed on the particle by some physical process. If $\boldsymbol{v}$ is the velocity of a particle in that reference frame, in an arbitrary instant of time t, its acceleration $\boldsymbol{a} = \frac{d\boldsymbol{v}}{dt}$ will be nonzero, and this quantity will be a convenient measure of the force $\boldsymbol{f}$ being applied.

[6] The reference frame of the stars is a good inertial frame for experiments carried out on Earth.

In accordance with the Aristotelian principle that causes should be proportional to their effects, Newton assumed that $\boldsymbol{f}$ is proportional to $\boldsymbol{a}$, or $\boldsymbol{f} = m\boldsymbol{a}$, where m is the coefficient that will depend on the particle under consideration and that we will call *(inertial) mass*.[7]

Third Law According to Newton's third law, when two particles P and Q interact, the force applied on P by virtue of Q is equal to the force applied on Q by virtue of P, but of opposite direction.

Newton defined *momentum* $\boldsymbol{p}$ of a particle as the product of its inertial mass m by its velocity $\boldsymbol{v}$.[8] Taken together, the second and the third laws give rise to the law of *conservation of momentum* that implies that the sum of momenta of two particles before a collision is equal to the sum of momenta of the same particles after that collision. If μ and 1 are the masses of the two particles a and b, respectively, and $\boldsymbol{u}_a$ and $\mathbf{0}$ are their respective velocities immediately before the collision, and $\boldsymbol{v}_a$ and $\boldsymbol{v}_1$ are their velocities immediately after the collision, then

$$\mu \boldsymbol{u}_a = \mu \boldsymbol{v}_a + \boldsymbol{v}_1 \tag{9.11}$$

that is

$$\mu = \frac{\|\boldsymbol{v}_1\|}{\|\boldsymbol{u}_a - \boldsymbol{v}_a\|} \tag{9.12}$$

and

$$(\boldsymbol{u}_a - \boldsymbol{v}_a)\,\mu = \boldsymbol{v}_1. \tag{9.13}$$

This last equation implies that the vectors $\boldsymbol{u}_a - \boldsymbol{v}_a$ and $\boldsymbol{v}_1$ are colinear, a result that constitutes the essence of the third law of Newton. For the unidimensional collider, Eq. 9.12 can be rewritten with the velocity scalars:

$$\mu = \frac{v_1}{u_a - v_a} \tag{9.14}$$

where u_a and v_1 are always positive and v_a, speed of the particle of proof mass, can be either negative or positive depending on its behaviour after the collision – bouncing back or going forward.

The Determination of Mass These equations show that the third law is also the way to ascertain the value of the coefficient called *mass*. Eq. 9.12 gives the mass of an arbitrary particle using a standard particle (of mass 1 kg): this value can be measured in a collision experiment. Thus, if one of the particles is chosen as unit, then the masses of all other particles can be determined by making them collide with the standard particle. Consider a possible measurement map M for mass.

[7] To Aristotle the *force* applied is the cause and in some way the velocity is the *effect*. Since uniform motion in a straight line does not need any explanation, Newton searched for the variation of uniform motion in a straight line as the required effect.

[8] In the *Principia*, Newton defined force as change of momentum, i.e., $\boldsymbol{f} = \frac{d\boldsymbol{p}}{dt}$.

The *inertial mass* $M(a)$ of a particle a, as determined by the collider and velocity measurements only, is defined by Eq. 9.14 rewritten in the form:

$$M(a) = \frac{v_1}{u_a - v_a}, \tag{9.15}$$

where $\boldsymbol{u}_a$ and $\boldsymbol{v}_a$ are the velocities of particle a before and after the collision, and $\boldsymbol{v}_1$ is the velocity after the collision of the standard reference particle. Here are some simple consistency theorems:

Proposition 8. *$M(a) < M(b)$ if, and only if, the particle a of mass μ bounces back when projected towards the particle b of mass μ' at rest.*

Proof. By Eq. 9.7, we have that

$$v_a = \frac{\mu - \mu'}{\mu + \mu'} u_a,$$

where the sign of v_a is decided by the difference $\mu - \mu'$. Thus, we only have to prove that $\mu < \mu'$. But, since $M(a) < M(b)$, we conclude

$$\frac{v_1}{u_a - v_a} < \frac{v_1'}{u_b - v_b},$$

if, and only if,

$$\frac{\mu v_1}{\mu u_a - \mu v_a} < \frac{\mu' v_1'}{\mu' u_b - \mu' v_b},$$

and, by conservation of momentum, if, and only if,

$$\frac{\mu v_1}{v_1} < \frac{\mu' v_1'}{v_1'},$$

and, finally, if, and only if, $\mu < \mu'$. □

In a similar way, it is straighforward to prove that:

Proposition 9. *$M(a) = M(b)$ if, and only if, the particle a of mass μ becomes at rest when projected towards the particle b with the same mass at rest.*

The basic question is: Does the CME implement a *comparative concept* supporting a *formal measurement* M in the sense of Hempel? Does M qualify as a measurement function? We will see that, indeed, we have both a comparative concept and a measurement.

9.6 Refinement of the Theory of Measurement

9.6.1 Measuring Quantities

Suppose that we wish to measure an attribute of an object of $\mathcal{O}$ using real numbers. We need a map $M : \mathcal{O} \to \mathbb{R}$ assigning to each object $a \in \mathcal{O}$ an attribute value $M(a)$. Such a map cannot be chosen arbitrarily. To qualify as a measurement in an empirical science, an experiment must be conceived that "validates" or "witnesses" the definition. The experimental apparatus works with the objects from $\mathcal{O}$, allowing the experimenter to compare different objects with respect to a given attribute. The outcome of each experiment is an event that tells us whether or not the attribute of object a is less than the attribute value of object b. Observing the equipment, there will be an event for "yes", an event for "no", and an event for "don't know". As we will see shortly, in our theory, "don't know" is an event "experiment timed out". With time in mind, we adapt the notation in Section 9.2.2: in the bipolar subset of events we replace 0 with $\perp$ ("undefined") to mark that the binary equivalence $\mathcal{E}$ is true.

Let us assume there is a time $t \in \mathbb{N}$ associated to each experiment. A collection of such times constitute the schedule of the collider protocol. In all measurement procedures in this paper, the experimenter – the Turing machine – generates a possibly infinite sequence of binary words $\{z_i\}_{i\in\mathbb{N}}$. If the time schedule of oracle consultation allows, then this sequence converges into the unknown real ζ being measured (in its binary expansion).

For the purpose of what follows, every number ζ can be seen as an infinite binary string. We don't accept infinite suffixes of 1s to denote dyadic rationals. If a sequence is finite, then we consider an infinite number of 0s padded to its right. The concept of limit induces a topology over the set of finite and infinite binary sequences $\{0, 1\}^\omega$.

Definition 13. *We say that the sequence of binary words $\{z_i\}_{i\in\mathbb{N}}$ converges to ζ if (a) for all $i \in \mathbb{N}$, z_i is a finite sequence, (b) for all $i \in \mathbb{N}$, z_i is a prefix of ζ, and (c) for each prefix z of ζ, there is a $i \in \mathbb{N}$ such that z is a prefix of z_i.*

Each experimental apparatus $\mathcal{A}$ we have explored so far is specified by a physical theory $\mathcal{T}$ and is designed to measure a real number ζ. Let $\mathcal{A}(\mathcal{T}, \zeta)$ denote the experimental apparatus together with the quantity. We are able to define precisely the notion of a measurable number:[9]

Definition 14. *Let $\mathcal{A}(\mathcal{T}, \zeta)$ be an experimental apparatus for physical theory $\mathcal{T}$ and physical quantity ζ. The number ζ is* measurable *if the Turing machine equipped with the physical oracle $\mathcal{O}(\mathcal{T}, \zeta)$ and a time schedule can produce an infinite sequence of prefixes of ζ, $\{z_i\}_{i\in\mathbb{N}}$, without timing out in any query, such that*

[9] Compare the context of Geroch and Hartle (1986) and Beggs et al. (2008a, c, 2009a).

$$\lim_{i\to\infty} z_i = \zeta. \tag{9.16}$$

In the bisection method, the infinite sequence of queries is almost such a sequence $\{z_i\}_{i\in\mathbb{N}}$, but *not quite* since each query may differ in the last bit from a prefix of the unknown number being measured. We define the meet operation, which allows us to identify the largest common prefix to two given words over the same alphabet Σ:

Definition 15. *Let α and β be two finite or infinite words over the same alphabet Σ. We define the meet $\alpha \sqcap \beta$ as the finite word γ over Σ, if it exists, such that (a) γ is prefix of both α and β and (b) if δ over Σ is prefix of both α and β, then δ is a prefix of γ. It such a prefix does not exist we say that the meet is undefined.*

Thus, according with our previous analysis of experimental situations, the sequence of queries involved in the bisection procedure has the following property: if ζ is measurable, then the sequence $\{z_i \sqcap \zeta\}_{i\in\mathbb{N}}$ converges to ζ. Notice that, whenever one of the words over Σ is finite, the meet is always defined. If the meet is undefined, we say that its size is infinite. The following proposition is straightforward to prove:

Proposition 10. *Let $\mathcal{A}(\mathcal{T},\zeta)$ be an experimental apparatus for physical theory $\mathcal{T}$ and physical quantity ζ. The number ζ is measurable if, and only if, a Turing machine with physical oracle $\mathcal{O}(\mathcal{T},\zeta)$ and a time schedule can produce an infinite sequence of queries $\{z_i\}_{i\in\mathbb{N}}$ such that*

$$\lim_{i\to\infty} z_i \sqcap \zeta = \zeta. \tag{9.17}$$

9.6.2 Measurement Axioms with Time

We begin with some properties of abstract binary relations indexed by a real parameter "time" $t > 0$ on a set $\mathcal{O}$.

Definition 16. *A relation $\mathcal{E}_t$ in $\mathcal{O}\times\mathcal{O}$, for the time bound $t > 0$, is said to be a* timed equivalence relation *if there is a $K \geq 1$ so that*

(a) *$\mathcal{E}_t$ is reflexive,*
(b) *$\mathcal{E}_t$ is* timed symmetric*: for every a, b in $\mathcal{O}$, if $a\mathcal{E}_t b$, then $b\mathcal{E}_{t/K}a$,*
(c) *$\mathcal{E}_t$ is* timed transitive*: for every a, b, and c in $\mathcal{O}$, if $a\mathcal{E}_t b$ and $b\mathcal{E}_t c$, then $a\mathcal{E}_{t/K}c$,*
(d) *if $t < t'$, then $a\mathcal{E}_{t'}b \Rightarrow a\mathcal{E}_t b$.*

Definition 17. *Two binary relations $\mathcal{E}_t$ and $\mathcal{L}_t$ $(t > 0)$ determine a* timed comparative concept *for the elements of $\mathcal{O}$, if*

(a) *$\mathcal{E}_t$ is a timed equivalence relation,*
(b) *there is a $K \geq 1$ so that for every a, b, c in $\mathcal{O}$, if $a\mathcal{L}_t b$ and $b\mathcal{L}_t c$, then $a\mathcal{L}_{t/K}c$,*
(c) *for all $t > 0$ and $a, b \in \mathcal{O}$, exactly one of $a\mathcal{E}_t b$, $a\mathcal{L}_t b$, $b\mathcal{L}_t a$ holds,*
(d) *if $t < t'$, then $a\mathcal{L}_t b \Rightarrow a\mathcal{L}_{t'}b$.*

Note that Definition 17(c) summarises the ideas of *irreflexivity* and *connectedness*.

Note also that, although property 16(d) is kept explicitly, it can be omitted, since it is derivable from the other properties listed in Definition 16 and those listed in Definition 17.

Proposition 11. *If $t < t'$, then $a\mathcal{E}_{t'}b \Rightarrow a\mathcal{E}_t b$.*

Proof. Suppose that $a\mathcal{E}_{t'}b$ holds. Then $a\mathcal{L}_{t'}b$ does not hold, due to property Definition 17(c). We conclude, by Definition 17(d), that $a\mathcal{L}_t b$ does not hold. Then, either $b\mathcal{L}_t a$ or $a\mathcal{E}_t b$ holds. If $b\mathcal{L}_t a$ holds, then $b\mathcal{L}_{t'}a$ holds and $a\mathcal{E}_{t'}b$ cannot hold, by Definition 17(c), which is against the hypothesis. Thus $a\mathcal{E}_t b$ is the case. □

Now suppose we have an experimental apparatus for making measurements. This takes the form of some form of comparison of two objects in $\mathcal{O}$ taking place in a given time $t > 0$. (The time t is allowed to vary over real values for convenience, but there would be no problem in restricting it to rational values, or with slight modification to some formulae, integer values.) The possible outcomes for the experiment are labelled $\{-1, \perp, +1\}$, where $\perp$ should be thought of as "no answer". We will now define, for all $t > 0$, binary relations $\mathcal{E}_t$ and $\mathcal{L}_t$ on $\mathcal{O}$ by using this experiment. Later we shall discuss when these relations obey Definition 17.

Definition 18. *Whenever the experiment is done with arbitrary objects $a, b \in \mathcal{O}$, if the outcome in time t is event -1, then $a\mathcal{L}_t b$ is the case, if the outcome in time t is event is $+1$, then $b\mathcal{L}_t a$ is the case, and if the outcome in time t is "no answer" ($\perp$), then $a\mathcal{E}_t b$ is the case.*

Definition 19. *Let $\mathcal{E}_t$ and $\mathcal{L}_t$ be timed comparative relations on the set $\mathcal{O}$ of objects (Definition 17). Suppose there exists an* experimental apparatus *to witness these relations, as in Definition 18. Then the map $M : \mathcal{O} \to \mathbb{R}$ is a* measurement map *if*

1. *For all time $t > 0$, if $a\mathcal{L}_t b$ holds, then $M(a) < M(b)$.*

Considering the real $M(a)$, for the object $a \in \mathcal{O}$, as an infinite binary sequence, we denote by $M(a)\restriction_n$ the dyadic rational corresponding to the prefix of size n of $M(a)$ and by a_n an object from $\mathcal{O}$ with that measure. Such an object a_n exists due to the convention of the toolbox of standards: once specified the *unit*, we have access to all its multiples and submultiples.

Definition 20. *The* complexity of a measurement map $M : \mathcal{O} \to \mathbb{R}$, *given the timed comparative relations $\mathcal{E}_t$ and $\mathcal{L}_t$ on the set $\mathcal{O}$ of objects, is the map $T : \mathbb{N} \to \mathbb{N}$ defined as follows*:

$$T(n) = \min\{t \in \mathbb{N}\{0\} : a_n \mathcal{L}_t a \text{ for some } a, a_n \in \mathcal{O} \text{ with } M(a_n) = M(a)\restriction_n\}.$$

For the *collider machine experiment*, the complexity of the measurement map is *exponential*. This complexity of measurement is, indeed, a lower bound on the time needed to get an answer from the machine, as can be seen in the proof of Proposition 6.

Now, we introduce an extra axiom for the physical apparatus.

Definition 21. *The apparatus satisfies the* separation property *for the measurement map $M : \mathcal{O} \to \mathbb{R}$ if for every objects a and b in $\mathcal{O}$, if $M(a) < M(b)$, then there exists a time bound t such that $a\mathcal{L}_t b$.*

To connect these ideas with Hempel's axiomatisation, we use the following definition:

Definition 22. *Given the timed comparative concept $\mathcal{E}_t$ and $\mathcal{L}_t$, for some time bound t, we define the following relations $\mathcal{E}_{lim}$ and $\mathcal{L}_{lim}$:*

(a) for every a and b in $\mathcal{O}$, $a\mathcal{E}_{lim}b$ if $a\mathcal{E}_t b$ for every time bound t, and
(b) for every a and b in $\mathcal{O}$, $a\mathcal{L}_{lim}b$ if there exists a time bound t such that $a\mathcal{L}_t b$.

Proposition 12. *If the two relations $\mathcal{E}_t$ and $\mathcal{L}_t$ define a timed comparative concept (Definition 17) and the physical apparatus witnessing the relations satisfies the separation property (Definition 21), then the two relations $\mathcal{E}_{lim}$ and $\mathcal{L}_{lim}$ define a comparative concept and M is a measurement map in the sense of Hempel (see Definitions 3 and 4).*

Proof. We have to prove that Hempel's axiomatization holds, which is straightforward.

1. $\mathcal{E}_{lim}$ is reflexive: Suppose that, for some object a in $\mathcal{O}$, $a\mathcal{E}_{lim}a$ does not hold. It means that, for some time bound t, $a\mathcal{E}_t a$ does not hold, which is a contradiction with the fact that $\mathcal{E}_t$ is reflexive.
2. $\mathcal{E}_{lim}$ is symmetric: Use Definition 16(b).
3. $\mathcal{E}_{lim}$ is transitive: Use Definition 16(c).
4. $\mathcal{L}_{lim}$ is transitive: Use Definition 17(b).
5. $\mathcal{L}_{lim}$ is $\mathcal{E}_{lim}$-irreflexive: Suppose that, for some objects a and b in $\mathcal{O}$, both $a\mathcal{E}_{lim}b$ and $a\mathcal{L}_{lim}b$ hold. Then, there is a time bound t such that $a\mathcal{L}_t b$. Since $\mathcal{L}_t$ is $\mathcal{E}_t$-irreflexive, we conclude that $a\mathcal{E}_t b$ does not hold, which is contradictory with the case that $a\mathcal{E}_{lim}b$ holds.
6. $\mathcal{L}_{lim}$ is $\mathcal{E}_{lim}$-connected: Suppose that, for some objects a and b in $\mathcal{O}$, $a\mathcal{E}_{lim}b$ does not hold. Then, there is a time bound t such that $a\mathcal{E}_t b$ does not hold. Consequently, since $\mathcal{L}_t$ is $\mathcal{E}_t$-connected, either $a\mathcal{L}_t b$ or $b\mathcal{L}_t a$, meaning that either $a\mathcal{L}_{lim}b$ or $b\mathcal{L}_{lim}a$.
7. Suppose that $M(a) \neq M(b)$. Then either $M(a) < M(b)$ or $M(a) > M(b)$. Consider the first case. By the separation property (Definition 21), there exists a time bound t such that $a\mathcal{L}_t b$ holds. Consequently, $a\mathcal{E}_t b$ is not the case and, therefore, $a\mathcal{E}_{lim}b$ is not the case.
8. If $a\mathcal{L}_{lim}b$, then there exists a time bound t such that $a\mathcal{L}_t b$ and, consequently, $M(a) < M(b)$.

And we are done! □

9.6.3 The Collider as an Example

Now we are in a position to prove that the CME is a measuring process and that the mass obtained by the collision experiment is a measurement map. We use $\mathcal{O}$ to denote the set of objects used in the collider experiment. For the collider experiment, we measure mass using Eq. 9.15, which is independent of the value of the initial velocity. The vital fact to remember is that the time t_{exp} taken to conclude the physical experiment for masses m_a and m_b is bounded by (for constants $A,\ B > 0$):

$$\frac{A}{|m_a - m_b|} \leq t_{exp} \leq \frac{B}{|m_a - m_b|}. \tag{9.18}$$

Proposition 13. *The map M, given values by Eq. 9.15, is a measurement map with exponential complexity. That is, the collider provides a model of the timed axioms of measurement.*

Proof. We start by providing the semantics of the predicates $\mathcal{E}_t$ and $\mathcal{L}_t$. We say that two objects a and b have *experimentally* the same mass – event $\perp$ – if when a collides with b, there is no answer from the oracle in time t. We say that the object a has less mass than b if when a collides with b, the object a bounces back in time t.

Note that the separation axiom provided in Definition 21 is valid for the collider machine experiment: for every objects a and b in $\mathcal{O}$, if $M(a) < M(b)$, that is if $m_a < m_b$, then the time needed to detect the bouncing of object a is

$$t \leq \frac{B}{|m_a - m_b|},$$

that is, $a\mathcal{L}_t b$.

The $\mathcal{E}$-irreflexivity and $\mathcal{E}$-connectivity follow directly from the fact that the experimental outcomes (for a given setup) are exactly one of $\{-1, \perp, +1\}$. The properties 16(d) and 17(d) on increasing time are true, as a result of ± 1 at time t guarantees the same result for any time $t' > t$.

Let us prove that the predicate $\mathcal{E}_t$ is a timed equivalence relation.

It is *reflexive*: if two copies of a are made to collide, then there is no answer from the oracle at any time – event $\perp$. Consequently there will be no answer in time t.

It is *timed symmetric*: if a collides with b with no answer from the oracle in time t, then

$$\frac{B}{|m_a - m_b|} > t.$$

Then, if b collides with a, then

$$\frac{A}{|m_b - m_a|} > \frac{A}{B}t.$$

Thus, $a\mathcal{E}_t b \Rightarrow b\mathcal{E}_{A/B\,t}a$.

It is *timed transitive*: Suppose that a collides with b with no answer in time t, and that b collides with c with no answer in time t. Then

$$\frac{B}{|m_a - m_b|} > t \quad \text{and} \quad \frac{B}{|m_b - m_c|} > t.$$

Since

$$|m_a - m_c| = |m_a - m_b + m_b - m_c| \leq |m_a - m_b| + |m_b - m_c|,$$

we have

$$|m_a - m_c| < \frac{2B}{t}.$$

Now if a collides with c, there will be no answer in time $A/(2\,B)t$.

The proof that the predicate $\mathcal{L}_t$ is a transitive relation follows the same guidelines as the proof given immediately above. If a collides with b and bounces back in time t and b collides with c and bounces back in time t, then

$$\frac{A}{|m_a - m_b|} \leq t \quad \text{and} \quad \frac{A}{|m_b - m_c|} \leq t.$$

Since, in this case,

$$|m_a - m_c| = |m_a - m_b + m_b - m_c| = |m_a - m_b| + |m_b - m_c|,$$

the upper bound on the experimental time required to distinguish a and c is

$$\frac{B}{|m_a - m_c|} = \frac{B}{|m_a - m_b| + |m_b - m_c|} \leq \frac{B}{2\,A}t.$$

The complexity of the map is determined by the analysis done in the proof of Proposition 6. □

The theory of the *collider machine experiment* CME as a measurement device can be developped and fully axiomatized. Of course Hempel's timed system of axioms is not complete for the CME: many further complex properties of the CME can be axiomatised. Mainly, those properties that dissect the entanglement of the relations $\mathcal{E}_t$ and $\mathcal{L}_t$ for arbitrary values of t.

Let us give an example. In Hempel's system, it can be proved that, for every objects *a, b*, and *c* in $\mathcal{O}$, if $a\mathcal{L}b$ and $b\mathcal{E}c$, then $a\mathcal{L}c$. In the timed system, it does not hold that, for every objects *a, b*, and *c* in $\mathcal{O}$, if $a\mathcal{L}_t b$ and $b\mathcal{E}_t c$, then $a\mathcal{L}_t c$. But for the collider this theorem can be replaced by a timed one in the following form:

Proposition 14. *For every objects a, b, and c in $\mathcal{O}$, for every time bound t, there is a $K \geq 2$ so that the following holds: If $a\mathcal{L}_t b$ and $b\mathcal{E}_{Kt}c$, then $a\mathcal{L}_{Kt}c$.*

Proof. If $a\mathcal{L}_t b$ and $b\mathcal{E}_{t'}c$, then

$$t > \frac{A}{m_b - m_a} \text{ and } t' < \frac{B}{|m_b - m_c|}.$$

If $t' = 2B/A\,t$ then we have

$$|m_b - m_c| < (m_b - m_a)/2,$$

and then

$$m_c - m_a \geq m_b - m_a - |m_b - m_c| > (m_b - m_a)/2.$$

Then an upper bound on the time taken to distinguish a and c is

$$\frac{B}{m_c - m_a} < \frac{2\,B}{m_b - m_a} < \frac{2\,B}{A}t. \qquad \square$$

Many propositions of this kind can be proved for the CME, namely introducing quantifiers. They show how masses can be compared in the less abstract timed system, where measurements take time, without further measurements.

We can also see how the CME fails to measure with arbitrary accuracy when used with a polynomial time limit:

Proposition 15. *Let $p(n)$ be a polynomial. For any a, a_n in $\mathcal{O}$ ($n \in \mathbb{N}$), such that $M(a_n) = M(a)\restriction_n$, there are only finitely many n so that $a_n\mathcal{L}_{p(n)}a$.*

9.6.4 Complexity

We propose that a measurement procedure has a "computational complexity" that can be derived from the intrinsic duration of the phenomenon considered.

If a is the object being measured and, for all $i \in \mathbb{N}$, a_i is the object from the toolbox of standards corresponding to the dyadic rational z_i, then we can restate Proposition 10 in the following terms:

Proposition 16. *Let $\mathcal{A}(\mathcal{T}, \zeta)$ be an experimental apparatus for physical theory $\mathcal{T}$ and physical concept value ζ. If the Turing machine with the physical oracle $\mathcal{O}(\mathcal{T}, \zeta)$ and a schedule can give instructions to set an infinite sequence of objects $\{a_i\}_{i\in\mathbb{N}}$ to be compared with object a in some attribute, by the bisection method, without timing out in any query, then*

$$M(a) = \lim_{i\to\infty} M(a_i). \tag{9.19}$$

Proposition 17. *If the Turing machine (experimenter) is equipped with the bisection algorithm, then the analogue-digital collider machine can serve as measurement apparatus for the measure of mass with complexity exponencial in the size of the query.*

Proof. The time of the experiment is exponential in the size of $z_i \sqcap \zeta$, where z_i is the i-th query and ζ the unknown mass. Using the bisection algorithm the size of the largest common prefix is $|z_i|$ up to 1 unit. Consequently, the time computed in this way is the same complexity class $(k'2^{kn})$. □

This last proposition shows that the bisection method is one of those methods that allows the experimenter, equipped with the toolbox of standards, to measure the unknown mass with a time schedule that does not depend on the unknown mass, although the experiment may time out assigning the two objects in the measurement context *the same mass* in the sense of relation $\mathcal{E}$.

We think these last propositions give a solid ground to understand our physical experiences of measurement and the role of the Turing machine as experimenter.

Now we introduce what we think is the most relevant concept:

Definition 23. *We say that a measurement in physical theory $\mathcal{T}$ has structural complexity T if the associated measurement map M has a computable complexity T in the sense of Definition 20.*

Then we can define complexity classes of measurements, such as:

Definition 24. *$\mathcal{T} - EXP$ is the class of measurements in physical theory $\mathcal{T}$ that have associated measurement maps with exponential time complexity, i.e., complexity $2^{O(n)}$.*

We can specify an open problem in measurement theory:

Conjecture 1. No reasonable physical measurement, based upon a reasonable physical theory $\mathcal{T}$, has an associated measurement map with polynomial time complexity.

The SME in Beggs and Tucker (2007) can be considered to be "unreasonable" since its behaviour is not *fully* governed by physical laws. This is because no physical law determines what happens in the "close vicinity" of the vertex of the wedge (cf. Froda 1959).

9.7 The Non-measurable Character of a Physical Concept

We start with a definition more general than Definition 14.

Definition 25. *A number ζ is said to be* measurable *over a physical theory $\mathcal{T}$ if there exists a Turing machine M with experimental apparatus $\mathcal{A}(\mathcal{T}, \zeta)$, specified by the physical theory $\mathcal{T}$, and physical oracle $\mathcal{O}(\mathcal{T}, \zeta)$ which, running over unbounded time, computes a sequence of rational approximations to (the binary expansion of) ζ.*

(Compare the quotations Geroch–Hartle 3 and 5.) We are now going to reconsider the collider experiment in Section 9.3. Let ζ denote the unknown value to be measured and $\{z_i\}_{i\in\mathbb{N}}$ be the sequence of words queried by the Turing machine.

From the sequence $\{z_i \sqcap \zeta\}_{i\in\mathbb{N}}$, introduced in Section 9.6, we can extract the sequence of sizes $\{|z_i \sqcap \zeta|\}_{i\in\mathbb{N}}$, which determines the lower bound of the time needed to perform the i-th consultation of the experiment, $i \in \mathbb{N}$.

We suppose there is a notion of *physical time* that belongs to the physical theory $\mathcal{T}$ underlying the measurement. Suppose the natural physical $\mathcal{T}$-time of the experiment has a lower bound exponential in the size of the largest common prefix of the unknown word and the query word. Then the sequence of lower bounds in the times needed for the consultations is $\{2^{|z_i \sqcap \zeta|}\}_{i\in\mathbb{N}}$. Therefore, even if the program for the Turing machine "cheats" for some $i \in \mathbb{N}$, by timing out some queries, an infinite subsequence of queries has to have time constraints. The proper way to formulate this property is via the Ω notation:

Proposition 18. *Let $\mathcal{O}(\mathcal{T}, \zeta)$ be an oracle to a Turing machine for a physical theory $\mathcal{T}$ and physical quantity ζ. Let physical $\mathcal{T}$-time be τ. Let the oracle consultation schedule be T. If the number ζ is measurable then $T \in \Omega(\tau)$.*

Now, we make a conjecture, which we will call the BCT Conjecture, stating:

Conjecture 2. For all reasonable physical theories $\mathcal{T}$, for all reasonable physical measurements of ζ based upon $\mathcal{T}$, the natural physical $\mathcal{T}$-time τ is at least exponential in the size of $z \sqcap \zeta$, where z is a query of the experimenter.

Our Conjecture 2 claiming *exponential in the size of the query* can be explored for the bisection algorithm. By *exponential* we generally mean a law of time of the form

$$\tau(n) = 2^{kn}, \tag{9.20}$$

for some value of k different from 0.

As an example, consider the speed of light of 299 792 458 ms^{-1}. Any attempt to prove that it is 299 792 458. 0^{ω} ms^{-1} will fail, according to our conjecture, but an attempt to prove that it is 299 792 458. $0^i d$ ms^{-1}, for some large i may succeed for some digit $d \neq 0$.

Conjecture 2 is suggested by our studies of *gedankenexperimente* in a variety of physical fields, measuring *length, mass, resistance, latitude, mass of a elementary particle*, and *Brewster's angle* in optics. All these experiments are fully described in Beggs et al. (2009c). The conclusion of each analysis is the same: the time needed to establish the nth bit of a value is at least exponential in n. Of course, if the statement of the conjecture is turned into a widely accepted thesis, or even a law about the process of measurement, then there will be deep consequences, both philosophical and physical.

The following propositions answer questions seen earlier in Section 9.4:

Proposition 19. *There are measurable numbers that are not computable.*

These are best seen through particular experiments such as Beggs and Tucker (2007).

Proposition 20. *There are computable numbers which are not measurable.*

Proof. Take any dyadic quantity ξ of size n and consider it measurable. Then, the Turing machine can produce a sequence $\{z_i\}_{i\in\mathbb{N}}$ of queries such that $\lim_{i\to+\infty} z_i = \xi$. As a consequence of the concept of limit provided by Definition 13, we know that there is an order $p \in \mathbb{N}$ such that, for $i > p$, $z_i = \xi$. For such queries z_i, $i > p$, the time of the experiment is infinite. □

This last Proposition 20 conspicuously challenges arguments in the quotation Geroch and Hartle 6 (recall Section 9.4). A reason is this: for Geroch and Hartle, a computable number is a priori, i.e., knowing that a number is computable we can prove it is computable. But, in our case, we do not know if a quantity being measured is computable or not.

We conclude that the Geroch and Hartle's Quotation 6 (see Geroch and Hartle 1986) is a difficult one. Our interpretation is that Geroch and Hartle are making distinguishing those numbers which can a priori be known to be computable and, consequently, measurable, and those numbers under the influence of an experimental apparatus. Indeed, what Geroch and Hartle state in Quotations 5 and 6, taken together, is that *all computable numbers predicted by physical theories are measurable*. This view is acceptable when only negative results are in context. But for the Philosophy of Physics, if it is a refutation what we are looking for, then even this exercise of Geroch and Hartle is not suitable.

The diference of knowing and not knowing in advance if a given quantity is computable or not is entangled in the following two propositions from Beggs et al. (submitted). The first tells us that, if we know a quantity in advance, then we can design a schedule (using that quantity as a *conventional oracle* (!)) that allows the experimenter to measure the number:

Proposition 21. *There are programs N_k (with integer $k \geq 1$), with specified waiting times (say T_k), so that the following is true: For any non-dyadic $\mu \in [0, 1]$ and any $n \geq 0$, there is a k so that program N_k will find the first n binary places of μ.*

But if that quantity is not known in advance than, for most numbers, there is a last bit that can be read. (cf. Proposition 19, stated in advance for the purpose of clarity.)

Proposition 22. *There are uncountably many $\zeta \in [0, 1]$ so that, for any program P with a specified computable schedule, having access to the oracle $\mathcal{O}(\mathcal{T}, \zeta)$, there is an n so that P cannot determine the first n binary places of ζ.*

We note that the *impression* that the non-algorithmic character of measurement is induced by the thresholds of sensitivity of the equipment is false. In the collider machine experiment the two flags are put at a finite non-zero distance from each other: notice that the non-measurability arises no matter how small is the distance

between the two flags. Besides that fact, there are uncomputable reals that are indeed measurable irrespective to the finite distance between flags of the collider.

Thus, a number is computable if there is a Turing machine that generates a sequence of rational approximations to the number.

A number is measurable if there is a Turing machine connected to the experiment that also produces rational approximations to that number – for the bisection method, the sequence of queries is that sequence of rational approximations.

The relation between the measurable and the non-measurable is as subtle as the relation between the computable and the non-computable. From what is non-measurable we can produce measurable numbers by suitable encoding. The same with the non-computable. Geroch and Hartle stresses this fact by giving the interesting example of a computable number made of non-computable numbers (see Geroch and Hartle 1986):

$$M = \sum_{n=1}^{\infty} \frac{3^{-n}}{s(n)}, \tag{9.21}$$

where $s(n)$ is the number of steps taken by the Turing machine encoded in n to halt. This function s is itself non-computable. However, the number M is computable. In order to approximate the number M to within error, say $\varepsilon = 0.01$, it suffices to deal only with the first ten terms in the sum, and, even for these, only either to determine $s(n)$ or else ensure that it exceeds 1,000. So, given $\varepsilon = 0.01$, our machine merely runs the first ten Turing machines for 1,000 steps each one, letting $s(n)$ be infinite for any machine that has not by then halted.

9.8 Conclusions

This paper is about measurement seen from a computational point of view. In our models of Turing machines with physical oracles, introduced in our papers (Beggs et al. 2008a, b, c, 2009a), we have been observing that our experiments make measurements (e.g., in Beggs et al. (2008a, 2009b)).

In Campbell (1928), Carnap (1966) and Hempel (1952), we find an established theory of measurement, axiomatized by Hempel (1952) extended by Carnap (1966). Campbell (1928), discusses the problem of measurement in experiments involving objects with almost identical attribute values.

According to a our framework all depends upon the physical theory chosen. For Newtonian mechanics we have shown that for some experimental quantities are *always* measurable (see Beggs et al. 2008c; Beggs and Tucker 2007) whilst for others there are quantities that are *not* always measurable. Our technical results can be used to show that the task of measuring quantities in physics can be classified by well known complexity classes. Principle 6, and the postulates, lead to a deeper understanding of experimenters and experiments which impose a *theoretical* and absolute limit on the measurability of a physical quantity.

In this paper we solved two problems: we were able to strongly root the ideas and results developed in Beggs et al. (2008a) in the Philosophy of Physics; and we were able to provide a decidable theory by adding time complexity measures into the Hempel's system of axioms.

Edwin Beggs, José Félix Costa and John Tucker would like to thank EPSRC for their support under grant EP/C525361/1.

References

Beggs E, Tucker JV (2006) Embedding infinitely parallel computation in Newtonian kinematics. Appl Math Comp 178(1):25–43

Beggs E, Tucker JV (2007) Experimental computation of real numbers by Newtonian machines. Proc R Soc Ser A (Math, Phy Eng Sci) 463(2082):1541–1561

Beggs E, Tucker JV (2008) Programming experimental procedures for Newtonian kinematic machines. In: Beckmann A, Dimitracopoulos C, Löwe B (eds) Computability in Europe, vol 5028 of Lecture notes in computer science. Springer, pp 52–66

Beggs E, Tucker JV (2009) Computations via Newtonian and relativistic kinematic systems. Appl Math Comp 215(2009):1311–1322

Beggs E, Costa JF, Loff B, Tucker JV (2008a) Computational complexity with experiments as oracles. Proc R Soc Ser A (Math, Phy Eng Sci) 464(2098):2777–2801

Beggs E, Costa JF, Loff B, Tucker JV (2008b) On the complexity of measurement in classical physics. In: Agrawal M, Du D, Duan Z, Li A (eds) Theory and applications of models of computation (TAMC 2008), vol 4978 of Lecture notes in computer science. Springer, pp 20–30

Beggs E, Costa JF, Tucker JV (2008c) Quanta in classical mechanics: uncertainty in space, time, energy. 2008. Accepted for presentation in Studia Logica International Conference on Logic and the foundations of physics: space, time and quanta (Trends in Logic VI), Belgium, Brussels, 11–12 December 2008

Beggs E, Costa JF, Loff B, Tucker JV (2009a) Computational complexity with experiments as oracles II. Upper bounds. Proc R Soc Ser A (Math, Phy Eng Sci) 465(2105): 1453–1465

Beggs E, Costa JF, Tucker JV (2009b) Physical experiments as oracles. Bull Eur Assoc Theor Comp Sci 97:137–151. An invited paper for the "Natural Computing Column"

Beggs E, Costa JF, Tucker JV (2009c) Physical oracles. Technical Report

Beggs E, Costa JF, Tucker JV (2010) Limits to measurement in experiments governed by algorithms. Technical Report, Swansea University, submitted for publication

Campbell NR (1928) An account of the principles of measurement and calculation. Academic, London and New York

Carnap R (1966). Philosophical foundations of physics. Basic Books, New York

Froda A (1959) La finitude en mécanique classique, ses axiomes et leurs implications. In: Henkin L, Suppes P, Tarski A (eds) The axiomatic method, with special reference to geometry and physics, studies in logic and the foundations of mathematics. North-Holland Publishing Company. Amsterdam

Geroch R, Hartle JB (1986) Computability and physical theories. Found Phy 16(6):533–550

Hempel CG (1952) Fundamentals of concept formation in empirical science, vol 2 of International encyclopedia of unified science. University of Chicago Press, Toronto Suppes P (1951) A set of independent axioms for extensive quantities. Portugaliæ Mathematica 10(2): 163–172

Suppes P (1951) A set of independent axioms for extensive quantities. Portugaliae Mathematica, 10(2):163–172

Chapter 10
Impredicativity of Continuum in Phenomenology and in Non-Cantorian Theories

Stathis Livadas

10.1 Introduction

Edmund Husserl held the early idea that pure mathematics belongs to the exact sciences dealing with idealities whereas phenomenology is a descriptive eidetic science of pure mental processes as viewed in the phenomenological attitude. They are fundamentally different in that they both use different cognitive tools and turn their attention to essentially different objects. This is Husserl's prevalent attitude to which he makes references especially in *Ideen I (Husserl 1982)*, where he supports that they can combine though they cannot take the place of one another.

It is my aim, though, in this article to demonstrate how the phenomenological analysis of time consciousness can not only provide a model of the intuitive and ultimately mathematical continuum, something that had already attracted the theoretical interest of prestigious mathematical names as that of H. Weyl and L.E.J. Brouwer in early twentieth century, but it can also motivate a new approach of the ontological nature of intuitive continuum and its ad hoc axiomatization in the language of non-Cantorian theories. On a phenomenological level, we are mostly based on the analysis of the phenomenological constitution of time as it is developed in Husserl's *Phänomenologie des inneren Zeitbewußtseins* (Husserl 1996) and the work of Patočka (1992) as well as on the more general Husserlian idea of genetical-kinetical constitution. The kinetical-genetical character of constitution in phenomenological analysis is based on the view that the world of objects is not given but as a unity of multiplicities as they are constituted in progression in the flux of multiplicities within consciousness.

Following this line Husserl confronted in *Phänomenologie des inneren Zeitbewußtseins* (Husserl 1996) the issue of a transcendental, non-temporal subjectivity objectified in the self-constituting unity of the flux of consciousness which in

S. Livadas (✉)
Division of Pedagogy, History and Philosophy of Mathematics, Department of Mathematics, University of Patras, Patras 26500, Greece
e-mail: livadas@math.upatras.gr

A. Carsetti (ed.), *Causality, Meaningful Complexity and Embodied Cognition*, Theory and Decision Library A 46, DOI 10.1007/978-90-481-3529-5_10,

a somehow circular turn is thought of later as constituted genetically in a kind of transcendental "genesis" constantly generating temporality. Becoming convinced that the transcendental ego is given in temporal profiles – "time is the universal form of all egological genesis" he professed in the *Fourth Cartesian Meditation* – he was inducing an impredicativity in the phenomenology of time, a radical transcendence.[1]

Nevertheless, the phenomenological constitution of time provides a model for the intuitive continuum and its impredicativity[2] a motive to reflect on its representation as an extension or a beyond the "horizon" factor in certain non-Cantorian mathematical theories that provide an alternative, phenomenologically oriented version of standard mathematics by discarding conventional infinity and following the ever shifting horizon of our incorporating Life-World (*Lebenswelt*) as is the case with Alternative Set Theory (AST) of the Prague School (Vopěnka 1979).

I claim here that the adoption of ad hoc extension principles or "external" predicates in non-Cantorian theories with respect to vagueness or fuzziness (that is, uncountable infinity) reflects on a formal-axiomatical level the impredicativity of the transcendental ego of consciousness in its Husserlian sense meant as the constituting factor of the continuous unity of the flux of internal time. This is also the case with respect to the intuitionistic approach to continuum by a choice sequence modeling, where a strong extension principle is adopted for the elements of the universal spread C (Van Atten et al. 2002, p 223). It should be noted again that intuitionistically oriented H. Weyl had already developed in *Das Kontinuum* (1918), a notion of the intuitive continuum based largely on the phenomenological description of the consciousness of internal time (Van Atten et al. 2002, p 203).

Finally, as we take a closer look of these alternative approaches to continuum we point to its inherent indescribability by means of a first-order formal language.[3] We hold that this indescribability manifests itself in the phenomenology of consciousness as the irreducibility of the continuous unity of the constituting flux of consciousness in-itself to the discrete mode of appearances of phenomena constituted as immanent unities within it.

In the following section we examine how the problematic of continuity arises as a circularity in the phenomenological description of the constitution of time consciousness.

---

[1] Husserl did not clarify to the end the meaning of the absolute ego in general of his *Cartesian Meditations* (Husserl 1931) and has drawn criticism on the part of philosophers like Theodor Adorno who claimed that Husserl did not succeed of getting rid of a grounded Cartesian ego (Adorno 1982, Ch 4, pp 227–228).

[2] The term impredicativity is used here in the sense of impredicative mathematical theories in which there is no stratification of the mathematical universe and, intuitively speaking, one cannot comprehend (or describe) the elements or the parts but in terms of the whole, or a big part of it. It should be noted in addition that in platonic sense impredicativity of an object is the impossibility to assign to it any predicates at all, defined thereby identically as the non-being (*Plato's Parmenides, 163D,* The Loeb Class.Library).

[3] A first-order formal language $\mathcal{L}$ is one that, roughly speaking, allows quantification only over countably many elements of this language and does not allow quantification over higher order objects, e.g. sets or functions.

10.2 Continuity in the Constituting Flux of Consciousness

In phenomenological analysis it is known that the conviction to an objective reality in an absolute sense is suspended by **Epochë** and substituted with a constituted reality approach. The constituted objects are immanent to the constituting flux of consciousness in which they are reflected in a certain mode, that is, in the *vor-zugleich* (anterior-simultaneous) mode that entails a continuum of phases trailing behind an original sensation each of which is a retentional consciousness of the preceding "present" (Husserl 1996, §38, p 104).

The temporal consciousness of appearances is the continuous unity of a whole, an all encompassing unity of the simultaneity and anteriority of the original sensations of actuality transforming continuously every group of original sensations, in the simultaneity, in a way that trails into an immediate posteriority which is a continuity and each of whose points is in the form of a homogeneous flow.[4]

Let us see how one can interpret in the phenomenological-kinetical fashion the continuous mode of the anterior-simultaneous flow of the original sensations with the "queue" of their retentions in the flux of consciousness. Husserl responded to this problematic by appealing to what he called double intentionality of the flux of consciousness, the immediate retention of an immanent object in the flux of consciousness (e.g. the sonore effect of a sound) on one hand and the intentional constitution of the "descending" sequence of retentions of this primary sensation in the flux as a continuous unity, always in the anterior-simultaneous mode of flow, which means that each new continuity of phases that present themselves instantaneously in simultaneity is a retention with respect to what is group continuity in simultaneity in the anterior phase. "Thus, the flux is traversed by a longitudinal intentionality which, in the course of flux, overlaps in itself continuously", (Husserl 1996, §39, pp 106–107, transl. by the author), see also Bernet (1983).

In this retentional–protentional mode of the constitution of the flux of consciousness – where by the term protention we understand a-thematic expectation similar but asymmetrical to retention – lacks however a definition in ontological or kinetical terms of the term continuity in the characterization of the mode of constitution of the flux in-itself. Husserl uses this term as the mode in which retentions are constituted in a "descending" sequence form, each term of the sequence being a continuous retention of the continuity of preceding phases. This is also the case in the second part of *Phänomenologie des inneren Zeitbewußtseins*. There, referring to the retentional structure of the flux, Husserl talks about the essential nature of every linear continuum that makes possible, departing from a point of intensity to think of every other point as produced continuously from that one, continuous production being

[4] "The totality of the group of original sensations is bound to this law: It transforms into a constant continuum (in ein stetiges Kontinuum) of modes of conscience, of modes of being-in the flow and in the same constance, an incessantly new group of original sensations taking originally its point of depart, to pass constantly (stetig) in its turn in the being-in the flow. What is a group in the sense of a group of original sensations remains as such in the modality of the being-in the flow." (Husserl 1996, §38, p 102, transl. by the author).

production in continuous iteration. In this way the constituted continuum of time is a flux of continuous production of the modifications of modifications (Husserl 1996, suppl. I, p 130).

The term continuity is treated here as a modality, without any further specification, in the description of the double intentionality character of the retentiontional flux and it makes possible to fix the regard to the flux in-itself constituted as an objective unity of consciousness. This would sooner or later lead to difficulties as ultimately one reaches the transcendence in the constitution of the flux in-itself. We may note, in passing, that one is left *in vacuum* as to the ontological nature of continuum in classical philosophy too.[5]

In this phenomenological interpretation the transcendence "underlies" the self-constitution of the continuous unity of the flux in contrast to the constitution of the discrete multiplicities of appearances (*Erscheinungen*) as immanent objects of the intentionality of the flux. Referring to this transcendence, Husserl asserted that it is impossible to extend the phases of this "flux" in a continuous succession, to transform it mentally in a way that each phase "extends" identically on itself, a certain phase of it belonging to a present that constitutes or to a past that also constitutes (not constituted) to the degree that it is an absolute subjectivity beyond any predicate and whose retentional continuity in the constituting flux is not but its objectification, its ontification by its "mirror" reflexion (Husserl 1996, §35, p 98).

We should also take into account that Husserl had already explicitly stated in *Philosophie der Arithmetik* (Husserl 1970a, II, pp 24, 25) the impossibility of the description of a collection of objects in phenomenological representation as a temporal instantaneity (*zeitliches Zugleich*) which evidently points to the structure of inner experience.

The self-appearance of the flux as a phenomenon in itself is not but an objectification of what is the ultimate subjectivity in the flux of consciousness put later in the most general terms as the absolute egoin *Cartesian Meditations*. This phenomenological transcendence, the source of all temporality as Husserl came to believe, will

[5] This problematic arises from the difficulty to describe ontologically under the same terms the continuum as a whole and its constituent unities. In Plato's *Parmenides* the instantaneous change in the state of a physical body is attributed to the effect of a somehow intermediate state between rest and motion, the $'\varepsilon\xi\alpha\acute{\iota}\varphi\nu\eta\varsigma$, not expressible in spatiotemporal terms (*Parmenides*, 156D, E, The Loeb Class. Library), whereas in Aristotle's *On Coming to Be and Passing Away* material points or lines are defined as limits, $\acute{o}\rho\iota\alpha$, of material bodies which in their turn cannot be composed by points or attachments but by indivisible bodies or magnitudes (*Coming to Be and Passing Away*, 320b 15 and 316b 15, The Loeb Class. Library).

In R. Descartes, physical space in its infinitely divisible parts (up to extentional points) is defined as a primary substance, *Res extensa*, filled up with matter, as spatial extension is a substantial characteristic of matter (*Discours de la Méthode,* pp 168–169, Garnier Flammarion, Paris).

These extensional individualities are defined in Leibnizian monadology as incarnations of unique monadic localities representing in particular the body which they "affect" and whose entelechy realize. Space is thus, what results from those uniquely defined monads taken together (G. Leibniz, *Fifth Letter to Clark*, §47).

be conceptually linked to the axiomatization of continuum in certain non-Cantorian mathematical theories as phenomenologically constituted time can be regarded as the basis of the intuition of all continuity phenomena.

10.3 Impredicativity of Phenomenological and Mathematical Continuum

10.3.1 Phenomenological Recurrence to Absolute Subjectivity

Husserl came gradually to examine thoroughly the idea of the absolute ego in general and in the Fifth Logical Investigation, around 1913, thought of the phenomenologically reduced ego as a "residuum" resisting all reductions "identified with the unity of the set of structures which cause the various acts of consciousness to glue together into a single self-related stream" (Husserl 1970c, LI, V, §4, p 541). It was after the 1920s that Husserl faced squarely the problem of the articulation of the nature and role of the transcendental ego in general as deeply related to the source of time consciousness (Moran 2000, Ch 5, p 173) and his deepest account of it will emerge at the end of the 1920s in Cartesian Meditations.[6]

This is not to be meant that the problematic of a transcendental or absolute subjectivity and its relation to time consciousness had not preoccupied him earlier as it is evident from his lectures of 1904–1905 at Göttingen and his work up to 1910 and further, published as *Phänomenologie des inneren Zeitbewußtseins* under the nominal editorship of Martin Heidegger. In *Ding und Raum*, edited out of his 1907 lectures, Husserl talked also of the unity – in fact continuous, unbroken unity – that is the primary characteristic associated with the phenomenological perception (*Wahrnehmung*) in the constitution of all spatiotemporal phases in consciousness out of pre-phenomenal experience. The particular abstract phases in this unity cannot be taken as such on their own but only as abstractions out of the continuous unity of *Wahrnehmung*. In this part of the description of the spatiotemporality of things Husserl acknowledged that he had not yet reached deep enough in the constitution of temporality and was conscious of the difficulties posed by the problem. It is very important to underline at this point his aporeia as to how the moments of *Wahrnehmung* make that appear in temporal constitution, so substantially different a temporal point and a time interval and make also apprehensible that wondrous difference between "now" and "just passed" (Husserl 1973, Ch 4, §19, pp 60–65).

[6] "The Universe of living that composes the "real" content of the transcendental ego is not co-possible but as the universal form of the flux, a unity in which all particular elements are integrated as flowing by themselves. . . . We can see in them (the forms of the states of living) the formal laws of the universal genesis, according to which, thanks to a certain noetico-noematical structure, are constituted and united continuously the modes of flux: past, present, future." (Husserl 1931, Fourth Meditation, pp 63–64, transl. by the author).

If the double intentionality of the retention is a mode of constitution both of an immanent object as such and the internal flux in-itself, it is evident that it concerns an objectification of the flux in its extension. This means if it is the ultimate reduction to be effected, we have to apprehend what is most subjective in the subject refusing to admit whatever is constituted and thus presented in a temporal extension.

What is left after this ultimate reduction is a transcendence that Husserl called *nunc stans* or stationary actuality which is itself an oxymoron as actuality is something which by necessity passes (Patočka 1992, VII, p 165). This *nunc stans* which is a name for the transcendental ego of consciousness (or the ego in-act) cannot be brought into reflection but only through its "mirror" autoreflexion, so there must be something interposed between its subjectivity acting now and its mirror objectification in the reflection.

Patŏka implies that a kind of retention is interposed between pure ego in-act and reflected ego without which pure or transcendental ego of consciousness would be inaccessible. But retention, as a presupposition of every possible reflection on the transcendental ego is by itself a circularity since it has been previously put as a modal characteristic of the objectified flux of consciousness without any ontological or other designation. Patoka infers, in any case, that the function of the transcendental ego independently of its "autoalienation", from which it is inseparable, is inaccessible to us without objectification. It follows that since each objectification points to something already objectified one can ensure the assumption of an absolute subjectivity making possible the unity of the flux as a whole and which transcends temporality in the phenomenological sense. Since, in turn, temporality is the necessary condition of every individuality and every existence in the world as well as of every first degree transcendental reflection one cannot attribute to this transcendental ego the predicates of existence and individuality (Patočka 1992, VII, pp. 163–168).

We are thus led to the impredicativity of the transcendental ego of consciousness in supplementary reduction that is, in fact, already implicitly present in the description of the double intentionality mode of the constitution of the flux.[7]

This recurrence to impredicativity is implicit in Husserl's subsequent reduction, in *Formale und Transzendentale Logik*[8] (Husserl 1984), of the laws of analytical logic to subjective laws of evidence. In the structure of analytical judgements which must ultimately refer to the "things in-themselves", one is led to a group of judgements referring directly to individuals for the possibility and essential structure of which nothing can be said in analytical terms even that they necessarily possess a temporal form, a duration and a qualitative plenitude of duration (Husserl 1984, §82, p 276).

[7] In this approach we follow the lessons of 1893–1917. We leave aside, in this article, Husserl's subsequent views on the matter putting in doubt the distinction between immanent and absolute time, see *Die Bernauer Manuskripte* (1997/1998), Hu XXXIII.

[8] The original German text under the title *Formale und Transzendentale Logik. Versuch einer Kritik der logischen Vernunft* was published in 1929 in the *Jahrbuch für Philosophie und phänomenologische Forschung* edited by E. Husserl, Vol. X, pp 1–298.

Further, as individuals in-themselves are given in the "lowest level" by intentional experience in the sense of direct reference to individuals- as such, it can be inferred that "individuals" are the contents of original judgements based on the most primary evidence which is that of experience. In view of this, Husserl turns to the phenomenological – transcendental principle of universal genesis of consciousness to provide a theoretical foundation to the passing from predicative evidences donated as individualities-in-themselves to the impredicativity of experience as such, genetically constituted in every being's unity of the flux of consciousness (ibid., §86, 89, respectively, pp. 282–286, 293–296).

We should note in relation to this phenomenologically induced impredicativity with regard to the flux of time consciousness, G. Longo's view in *Naturalizing Phenomenology* in which he states that the intuitive circularity in phenomenological temporality is reflected in the apparent paradoxes of mathematical construction where the "impredicativity of analysis permits a possible formalization of this intuitive circularity" (Longo 1999, Ch 14, pp 406–407).[9]

It is also of significance to note that the intuitive continuum as conceived by L.E.J. Brouwer and H. Weyl is largely connected with the Husserlian description of the consciousness of internal time and both Brouwer and Weyl distinguished between "internal" intuitive time and "external" or scientific, measurable time. Moreover Brouwer's idea of the primordial intuition of mathematics largely concerns internal intuitive time (Van Atten et al., 2002, pp 203–204).

10.3.2 The Continuum in Alternative and Internal Set Theories

It seems purposeful, at first, to refer to the Husserlian idea of scale invariance, as an evident generic similarity which can lead to minima visibilia as point-like ultimate minimalities bearing the same eidetic relationships "discovered" in the macroscopic universe (see Husserl 1973, §48, p 166). This idea seems to have an important effect on the "shift of the horizon" principle in AST.[10]

In *Krisis der europäischen Wissenschaften und die transzendentale Phänomenol ogie* (Husserl 1970b), Husserl made more broadly known his idea of Life-World (*Lebenswelt*) as the sense-intuited preidealized world which is the grounding soil

[9] Impredicative notions in mathematics are generally those in which the *definiens* uses the *definiendum* e.g. an open set (open interval) of the real line is not defined as the union of singletons (one-element sets) but in terms of other basic open sets (open intervals).

[10] In the sense that this evident generic similarity justifies the transposition of the eidetic relationships "discovered" in the universe of common intuition to that beyond this "horizon". Although P. Vopěnka implicitly assumes this phenomenological principle in his Prolongation Axiom he seems to deny it in a later expository article on the philosophical foundations of Alternative Set theory where he allows for the possibility of a complete collapse of our intuitions beyond a genuinely qualitative shift of the horizon to apeiron (Vopěnka 1991). But this eventuality contradicts with the Husserlian idea of our Life-world as *gründenden Boden* (grounding soil) of an ever shifting horizon.

for the "objective-true" world of the sciences of exactness. Out of this sense-intuited world is substructed the classical mathematical world of idealized limit-shapes and the plena to which they belong. This intuitively given world can be intuited as an endlessly and ever shifting horizon with reference to which all particular causalities can be anticipated and are not themselves given. It was this particular idea of the shifting horizon of *Lebenswelt* that motivated P. Vopěnka's definition of the countability of a class (Vopěnka 1979, Ch I, p 39):

> If a large set x is observed then the class of all elements of x that lie before the horizon need not be infinite but may converge toward the horizon. The phenomenon of infinity associated with the observation of such a class is called countability.

The fundamental ideas of Alternative Set Theory as exposed in (Vopěnka 1991) are those of natural infinity in contrast to idealized classical infinity and the sharpening of the horizon to infinity involving by necessity the presence of an observer. In this phenomenological perspective natural infinity presents itself to us as a converging series of finite "appearances" to an ever shifting horizon, the closer to which they are the less definite and sharp they seem. So in this sequence, natural infinity in its most basic form presents itself as countable natural infinity. Classical countable infinity is derived from this sequence by constantly sharpening our view, that is by moving the horizon further and further so that it stabilizes as "unchangeable, definite and sharp" (Vopěnka (1991, p 118). The most radical and qualitative shift of the horizon beyond the bounds of which natural infinity no longer sharpens but becomes vague and uncountable is axiomatized by the adoption of Prolongation Axiom[11] whose deeper content is reflected in classical Cantorian mathematics in the application of the Power Set Axiom which refutes, in effect, the original Cantorian conception of a set as anything that can be counted (Lavine 1994, Ch IV, p 95).

This extension axiom together with an existence axiom (axiom of existence of proper semisets) which are "external" to a first-order axiomatical system with a built-in predicate for the natural numbers in Weyl's sense, axiomatize the indefiniteness and blurriness of infinity beyond the horizon of countability. In this sense they reflect, in axiomatization, the impredicativity of the intuitive continuum that is irreducible to a countable infinity advancing to an ever shifting horizon.[12] The conceptual motivation is again explicitly stated by P. Vopěnka as that of extending the sense-intuited AST universe of countable classes to the vagueness of infinities beyond countability, shifting in effect the horizon of Husserlian *Lebenswelt*.

[11] If F and G stand for functions, which in the AST extended universe are sets or classes, the Prolongation Axiom states that: *For each countable function F, there is a set function f such that* $F \subseteq f$. It is important here to have in mind that countability of a function is, in fact, countability of a class of ordered pairs of elements and that a set function can be an uncountable set of ordered pairs of elements.

[12] In a formal context, AST works with sets and classes as objects. Sets are definite (might be very large) but sharply defined and finite from the classical point of view in the sense that in its universe of sets, AST accepts the axioms of Zermelo–Fraenkel system with the exception of the axiom of infinity. Classes represent indefinite clusters of objects such as the class N of natural numbers in the classical sense. In this context the notion of a semiset represents, roughly, blurriness and non-surveyability in the observation inside a very large set (see Vopěnka 1979, Ch I).

Countability of classes is an ever shifting finiteness whereas uncountability is the vagueness or indiscernibility beyond the horizon of intuitive hereditary finiteness. But in formalizing continuous unity beyond the horizon of discreteness, AST has to adopt extension principles that are, in fact, "transcendental" principles in a first-order universe of a countable domain. It is the Prolongation Principle that lies as axiomatical basis in the definition of continuum properties in AST by means of indiscernibility equivalences[13] which though they are formal-mathematical notions incorporate in their definition the "transcendence" of this principle.

In my view these axioms of AST reflect the impredicativity in the phenomenology of time consciousness in the sense that these ad hoc axioms "bridge" the gap between the intuited discreteness of the countable path to the horizon and the continuous unity intuited as a vagueness beyond it (compare, in parallel, the discreteness of the multiplicities of appearances as immanences in time-consciousness held together in the continuous unity of the flux).

This impredicativity is reflected in the axiomatization of continuum in Internal Set theory (IST) by the use of the external predicate *standard* in its syntax. Internal Set theory is generally considered, if properly interpretated, as a non-Cantorian version of Robinson's nonstandard analysis (Drossos 1990). We may remark here that the non-Cantorian designation means generally that the Axiom of Choice (AC)[14] of the Cantorian system does not hold. Moreover, since AC implies the Excluded Middle Principle, the negation of the latter implies the negation of AC (Diaconescu 1975). But on the grounds that negation of actual or Cantorian infinity is a conceptual presupposition of the negation of AC, it follows that any theory that denies actual infinity in its axiomatization can be characterized as non-Cantorian.

Internal set theory adjoins to ordinary mathematics (the Cantorian in structure Zermelo-Fraenkel system plus the Axiom of Choice, or ZFC) a new undefined unary predicate called *standard* (Nelson 1977, pp 1165–1166). In this respect it is linked with the intensional development of nonstandard analysis in which infinitesimals and infinitely large numbers do not exist in an objective way as in the extensional case[15] (Robinson's nonstandard approach, ultrapower constructions,

[13] The indiscernibility equivalences $\doteq$ are binary relations that are Π-classes (having the reflexive, symmetric and transitive property among others) and equipped in addition with the property of compactness; that is, for each infinite set U there is at least a pair (x, y) with $x, y \in U$ such that, $x \neq y$ and $x \doteq y$. For further details and topological definitions of AST (monads, figures, closures, connected sets, etc) based on the indiscernibility equivalences (see Vopěnka, 1979, Ch III, §1, 2, 3).

[14] In a less strict mathematical formulation the Axiom of Choice states that: *Given a non-empty class of non-empty sets a set can be formed containing precisely one element taken from each set in the given class*. Although the Axiom of Choice might strike someone as being intuitively obvious it may be less so if one has to deal with sets or classes of uncountably infinite cardinalities. An intuitive version of AC is produced by AST theorist A. Sochor concerning countable classes in AST sense (Lano 1993, p 152).

[15] Concerning the extensional development of nonstandard analysis, mainly A. Robinson's version, one is led to the introduction of nonstandard elements endorsing, in effect, the Axiom of Choice or Zorn's lemma in the use of free ultrafilters both in the construction of the nonstandard structures themselves and in the proof of significant theorems (e.g. Loś theorem, Mostowki collapsing

superstructures) but their existence has a subjective meaning related to the limitations of an "observer" in his witnessed universe for whom the predicate *standard* plays the semantic role of "fixed" or "sharp" in informal mathematical discourse.

E. Nelson's IST essentially adopts the axioms of ZFC together with three new axioms taking care of the "action" of the undefined predicate *standard*. These axioms, the Idealization axiom (I), the Standardization axiom (S) and the Transfer axiom (T) can be thought of as semantical content equivalents of the extension axioms of AST in the sense that they express in the syntax of IST a shift of the horizon of "fixedness" (Nelson 1986, Ch 1, pp 2–12). The intuition, e.g., behind the Idealization axiom is that we can only fix a finite number of objects at a time and the intuition behind the Transfer axiom is that if something is true for all standard (fixed) but arbitrary x's then it is true for all x's.

By direct application of the Idealization axiom one can prove that there exists at least a nonstandard element in every infinite set. In particular there exists at least a nonstandard natural number, a fact that by itself implies that vagueness or indiscernibility is not necessarily linked to the real number structure but it is rather the result of the adoption of the external predicate standard, in the particular case, along with ad hoc axiomatical equipment; it is by this axiomatical means that the horizon of IST standardness is "shifted". As it is the case with AST-sense indiscernibility all subsequent definitions involving vagueness along infinitity (or infinitesimality) and relevant topological notions are expressed in terms of external formulas involving the predicate *standard*.[16]

Topological and continuity properties and more generally vagueness are ultimately reduced to the action of the external to Cantorian system predicate *standard* which in spite of its rather syntactical role in the context of IST, as supported by E. Nelson, acquires by the addition of the three ad hoc axioms a significant underlying semantic role in axiomatizing vagueness.[17]

I claim again that this "embedded" shift of the horizon of "fixedness" to the vagueness of continuum inside IST, by the adoption of the external predicate *standard*, reflects on a formal-axiomatical level the impredicativity of the intuitive continuum associated with the phenomenological description of temporal consciousness. This impredicativity was reflected in a more closely phenomenological fashion in AST by the adoption of the Prolongation and Existence of Proper Semisets Axioms.

---

function). Zorn's lemma, in fact, is logically equivalent to the Axiom of Choice which in this context, in its stronger form of Global Choice, "induces" indirectly a notion of classical (actual) infinity. For more details, one is referred to Robinson (1966), Stroyan (1976) and especially to Connes et al. (2000) for a more intuitive presentation of the (uncountable) Axiom of Choice.

[16] For instance, regarding any object that can be described uniquely within internal mathematics as standard such as the set of real numbers $\Re$, a real element x is *infinitesimal* in case for all standard $\epsilon > 0$ we have $|x| \leq \epsilon$. Next, $x \cong y$ (x is *infinitely close* to y) in case x–y is infinitesimal and further, if $E \subseteq \Re$ and E standard, E is *compact* in case for all x in E there is a standard x_0 in E with $x \cong x_0$. Regarding the definition of classical mathematical continuity, if f and x are standard then f is continuous at x in case $y \cong x$ implies $f(y) \cong f(x)$, see (Nelson 1986, Ch 1, pp 2, 13).

[17] For a thorough development of the ideas of general topology based on the predicate and axioms of standardness in IST, we refer to (Diener and Diener 1995, Ch 6, p 109).

The impredicativity engendered by the Husserlian absolute ego of consciousness is essentially produced out of the impossibility to express under the same predicates of existence and individuality the subjectivity of the flux in itself as a self-constituting continuous unity and its constituent immanences of phenomena within it together with their retentions–protentions. It seems that this phenomenologically induced impredicativity of continuum underlies its formalization inside non-Cantorian theories by making use of ad hoc extension principles.

10.3.3 The Intuitionistic Approach to Continuum

We have already noted that Weyl based largely intuitive continuum on the Husserlian description of the consciousness of internal time whereas Brouwer's idea of intuitive continuum can be also readily comprehended in connection with Husserl's analysis of the phenomenology of internal time. In this context, intuitionistic approach to natural numbers is based on abstraction from a temporal process in which they are intuited as durationless sensations in discrete succession. Evidently the concept of duration does not apply to natural numbers but the same cannot be held to be true of real numbers which in intuitionism are viewed as "incomplete" or "unfinished" objects. It is very important to state that in intuitionistic view "an act of abstraction that would give us a real number as a durationless point is not something of which we would be capable" (Van Atten et al. 2002, p 207).

Just as we do not experience durationless points in time we do not experience extensionless points in space, so intuitive continuum, as Brouwer and Weyl thought, cannot be understood as a set of durationless or extensionless points. As was the case with the alternative to Cantorian context theories previously described, sooner or later one would be confronted with the impredicative nature of intuitive continuum, so both Brouwer and Weyl held that in order to capture the fluidity of intuitive continuum one should replace the element/set relation with a part/whole relation in which parts are of the same lowest genus as the undivided whole.

On this account, Brouwer "split" the continuum or interval (a non-denumerable set) into parts (subintervals) homogenous to the whole under the relation of inclusion (part/whole) in which the order relation between disjoint intervals is the natural order of the continuum abstracted from the progression of time. By this construction a point P or a real number P is an indefinitely proceedable sequence of nested λ-intervals and the difference with the classical approach lies in the fact that the point P is not something like the limiting point to which these nested intervals converge in which case it would be defined as the accumulation point of midpoints of these intervals.[18] The point P is the sequence itself and the λ-intervals are parts of the point P (Van Atten et al. 2002, p 212).

[18] Let λ-intervals be intervals of the form $(\frac{\alpha}{2^{\nu-1}}, \frac{\alpha+1}{2^{\nu-1}})$. Then L.E.J Brouwer (1992, p 69) defined real numbers as follows:

We ... consider an indefinitely proceedable sequence of nested λ -intervals $\lambda_{\nu_1}, \lambda_{\nu_2}, \lambda_{\nu_3}, \ldots$ which have the property that every $\lambda_{\nu_{i+1}}$ lies strictly inside its predecessor λ_{ν_i} $(i = 0, 1, 2, \ldots)$. Then, by the definition of λ-intervals the length of each interval $\lambda_{\nu_{i+1}}$ at most equals half the length

It could be argued, though, that even in intuitionistic approach the part/whole relation could be reduced to an element/set one if one assigns to each of the nested intervals a corresponding natural or rational number. Then the indefinitely proceedable sequence of those intervals could, in AST terms, stand for the countable class FN of natural numbers which represents a path to some vastly remote horizon inaccessible to us, beyond that of hereditarily finite countability.

This particular example is just a case of choice sequences as an intuitive foundation for real analysis. The underlying idea is that these sequences need not be lawlike, in the sense that given an initial segment of them there need not be a law prescribing in advance any future terms of the sequence except of course for the natural assumption of the assignment of a unique value in each step. Brouwer's rationale was that by an indefinite procession of a sequence independently of whether it is a lawlike or a lawless one preserves the intuition of continuum in the sense that it cannot be reduced to a durationless point as it is always in progress, whereas Weyl had certain reservations about the status of lawless sequences as genuine and individual mathematical objects something that was also Husserl's point of view.

In any case both had to adopt, as it is the case with previously examined non-Cantorian theories, certain ad hoc extension principles in the form of the "Weak Continuity for Numbers", the stronger "Continuity Principle for Universal Spreads" (L. Brouwer) and "The Principle of Open Data" (H. Weyl) in order to formalize the intuition of continuum.[19] These are extension principles that, given an initial

---

of λ_{v_i}, and therefore the lengths of the intervals converge to 0. (...) We call such an indefinitely proceedable sequence of nested λ-intervals a point P or a real number P.

It should be noted that the point P is thought to be the sequence $\lambda_{v_1}, \lambda_{v_2}, \lambda_{v_3}, \ldots$ itself and not something as the limiting point (the unique accumulation point of the midpoints of these intervals) to which according to the classical conception these nested λ-intervals converge. Each of these λ_{v_i} is considered then as part of the point P (Van Atten et al. 2002, p 212).

[19] By virtue of the definition of a spread as a "law on the basis of which, if again and again an arbitrary complex of digits (a natural number) of the sequence ζ (the natural number sequence) is chosen, each of these choices either generates a definite symbol, or nothing, or brings about the inhibition of the process together with the definitive annihilation of its result; ... Every sequence of symbols generated from the spread in this manner (which therefore is generally not representable in finished form) is called an element of the spread. We also speak of the common mode of formation of the elements of a spread M as, for short, the spread M", L. Brouwer formulated the Continuity Principle for the universal spread C as follows:

A law that assigns to each element g of C an element h of A (the natural numbers), must have determined the element h completely after a certain initial segment α of the sequence of numbers of g has become known. But then to every element of C that has α as an initial segment, the same element h of A will be assigned (see Van Atten et al. 2002, pp 222–224).

The principle of Open Data in its simplest form can be stated symbolically as follows: $A\alpha \rightarrow \exists n\ (\alpha \in n$ and $\forall \beta \in n, A\beta)$ where A is a syntactical variable for any mathematical formula and α, β stand for lawless sequences. This principle essentially identifies under predication a lawless sequence α with all those lawless sequences of its neighborhood starting with the same initial segment n, see Troelstra (1977, §2.6, p 14). In a stronger than this, continuity principle, if one denotes by Cont_{LS} the class of lawlike operations on lawless sequences assigning natural numbers to lawless sequences such that:

for $\Gamma \in \text{Cont}_{LS}, \forall \alpha \exists x \forall \beta \in < \alpha_0, \alpha_1, \ldots, \alpha_{(x-1)} > (\Gamma\alpha = \Gamma\beta)$,

then:

$\forall \alpha \exists x\ A(\alpha, x) \rightarrow \exists \Gamma \in \text{Cont}_{LS} \forall \alpha\ A(\alpha, \Gamma\alpha)$ see (Troelstra 1977, §2.6, pp 14–19).

segment of choice sequences, treat them as complete, individual objects by means of continuity (lawlike) operations, essentially in an extension to a vague horizon approach. In this way intuitive continuum is "grasped" axiomatically by ad hoc extension principles shifting the natural bounds of the finite and discrete which, in the case of choice sequences, is represented by their initial segments.

10.4 Conclusion: A Reflection on the Impredicative Character of Continuum

As we followed Husserl's path to the transcedence of the absolute ego of conscience in general, we noted that the idea of intuitive continuum is deeply related to the source of temporal consciousness leading ultimately to a transcendence beyond temporality that is a necessary presupposition for the objectified unity of one's time consciousness. The resulting impredicativity is inevitably induced as it is impossible to describe the constitution of the continuum of the flux of consciousness in-itself in terms of its immanent unities in the protention–retention schema without falling into the circularity of continuity in the retention of immanences. It is the essentially impredicative nature of continuum in the genetical constitution approach of Husserlian analysis that is by necessity reflected in the axiomatization of continuum in nonstandard and non-Cantorian approaches (AST, IST, ultrafinitist theories) whenever they have to shift the hereditarily finite bounds of the local "environment" of a potential "observer" towards vague infinity.

We may note here that Cantorian mathematics incorporates only formally in axiomatization (but evidently does not "grasp") continuum by adopting the Continuum Hypothesis as an axiom independent to ZFC, since it cannot be proved or disproved by means of the other axioms of Zermelo-Fraenkel system. In connection with Continuum Hypothesis (CH) Gödel was already at pains to show that it is meaningful and determinate enough to expect a forthcoming unambiguous answer in a natural extension of Set theory in his 1947 paper *What is Cantor's continuum problem?* (Tieszen 1998, p 194). Truth or falsity of CH is still a matter of an ongoing theoretical controversy among set theorists (see, for example, Woodin 2001).

Adopting the Husserlian view of *Lebenswelt* intuited as an endlessly open and ever shifting horizon in AST or inducing vagueness at infinity as nonstandardness in IST, non-Cantorian theories and intuitionistic ones are "trapped" in the impredicativity of the continuum when they shift the boundaries beyond naturally intuited countability in our witnessed universe. They have to adopt extra ad hoc extension axioms "external" to the inner structure of a first-order axiomatical system something that essentially reflects the impossibility of an ontology of the continuum. As Gödel was refuting Carnap's syntactical program using his second incompleteness theorem, he was proving in effect that *"the mathematical essences we intuit could not be linguistic conventions. There are constraints on them that we do not freely invent or create. One might also say that this content or meaning will be "abstract" relative to the rules of syntax. Mathematical intuition will therefore not*

be eliminable. In Husserl's language, categorial intuition will not be eliminable. Thus, instead of clarifying the meaning of abstract and non-finitary mathematical concepts by explaining them in terms of syntactical rules, abstract and non-finitary concepts are used to formulate the syntactical rules." (Tieszen 1998, p 193).

The intuition of continuum in phenomenological sense can be considered as associated with such a categorial intuition based on the possibility of reflection on the continuous unity of the temporal flux of consciousness.

The impredicativity of continuum manifests itself also in the impossibility to describe topological structures which incorporate the intuition of the continuum[20] by first-order expressional tools. Back in 1918, Weyl stated that it is an "act of violence" to assume the perfect coincidence of the analytical construction of the continuum with that of phenomenal space and time "*... that is, the continuity given to us immediately by intuition (in the flow of time and motion) has yet to be grasped mathematically*" (Weyl 1977). It seems very doubtful that it could be ever grasped mathematically in any sense that would reflect the existence of an ontology of the continuum.

References

Adorno T (1982) *Against epistemology*: *A metacritique* (trans: Willis Domingo). Blackwell, Oxford

Bernet R (1983) La présence du passé dans l'analyse husserlienne de la conscience du temps. *Revue de métaphysique et de morale* 2:178–198

Brouwer LEJ (1992) In: van Dalen D (ed) *Intuitionismus*. Bibliographisches Institut, Wissenschaftsvelag, Mannheim

Connes A, Lichnerowicz A, Schützenberger MP (2000) *Triangle of thoughts* (trans: Gage J). Editions Oedile Jacob, Paris

Diaconescu R (1975) Axiom of Choice and complementation. Proc AMS 51:175–178

Diener Fr, Diener M (1995) *Nonstandard analysis in practice*. Springer, Berlin

Drossos CA (1990) Foundations of fuzzy sets: a nonstandard approach. *Fuzzy Sets Syst* North-Holland 37:287–307

Husserl E (1931) *Méditations Cartésiennes* (trans: Emm. Levinas G, Peiffer). Librairie Armand Colin, Paris

Husserl E (1970a) In: Lothar Eley (ed) *Philosophie der Arithmetik*. Husserliana 12. M. Nijhof, The Hague, Netherlands

Husserl E (1970b) *The crisis of European sciences and transcendental phenomenology* (trans: Carr D). Northwestern University Press, Evanston

Husserl E (1970c) *Logical Investigations*, 2 vols (tran: Findlay JN). Humanities, New York

[20] In Husserl's phenomenology there is an analogy between mathematical and perceptual intuition in the sense that questions concerning mathematical intuition should be simply more specific cases of questions concerning intentional reference and directedness to objects. A counterargument is offered by those who insist on the difference between mathematical and perceptual intuition on the grounds that while in perception objects of intuition are determinate and individually identifiable, that seems to be what is missing in the case of mathematical objects e.g. in the mathematical intuition of the symbol 'A' standing for the empty set, see Tieszen (1984, pp 399–400).

Husserl E (1973) In: Claesges U (ed) *Ding und Raum: Vorlesungen*. Husserliana 16. M. Nijhoff, The Hague
Husserl E (1982) *Ideas pertaining to a pure phenomenology and to a phenomenological philosophy* (trans: Kerten F). Martinus Nijhoff publishers, London
Husserl E (1984) *Logique formelle et logique transcendantale* (trans: Bachelard S) Editions PUF, Paris
Husserl E (1996) *Leçons pour une Phénoménologie de la conscience intime du temps* (trans: Dussort H). Editions PUF, Paris
Lano K (1993) The intuitionistic alternative set theory. *Ann Pure Appl Logic* 59:141–156
Lavine Sh (1994) *Understanding the infinite*. Harvard University Press, Cambridge MA
Longo G (1999) The mathematical continuum: from intuition to logic, ch 14, *Naturalizing phenomenology*. Stanford University Press, Stanford, CA, pp 401–425
Moran D (2000) *Introduction to phenomenology*. Routledge, New York
Nelson E (1977) Internal set theory: a new approach to nonstandard analysis. *Bull Am Math Soc* 83(6): 1165–1198
Nelson E (1986) *Predicative Arithmetic. Mathematical notes*. Princeton University Press, Princeton
Patočka J (1992) *Introduction à la Phénoménologie de Husserl* (trans: Abrams Er) Ed. Millon, Grenoble
Robinson A (1966) *Non-standard analysis*. North-Holland, Amsterdam
Stroyan KD, Luxembourg WAJ (1976) *Introduction to the theory of infinitesimals*. Academic Press, New York
Tieszen R (1984) Mathematical intuition and Husserl's phenomenology. *Noûs* 18(3):395–421
Tieszen R (1998) Gödel's path from the incompleteness theorems (1931) to Phenomenology (1961). *Bull Symb Logic* 4(2):181–203
Troelstra A (1977) *Choice sequences. A chapter of intuitionistic mathematics*. Clarendon Press, Oxford
Van Atten M, Van Dalen D, Tieszen R (2002) Brouwer and Weyl: the phenomenology and mathematics of the intuitive continuum. *Philosophia Mathematica* 10(3):203–226
Vopěnka P (1979) *Mathematics in the alternative set theory*. Teubner-Texte zur Mathematik, Teubner Verlag, Leipzig
Vopěnka P (1991) The philosophical foundations of alternative set theory. *Int J Gen Syst* 20:115–126
Weyl H (1977) *Das Kontinuum* (Italian Edition, care of B. Veit) Bibliopolis, Napoli
Woodin HW (2001) The continuum hypothesis I, the continuum hypothesis II. *N Am Math Soc* resp. 48(6):567–576 and 48(7):681–690

Part IV
Epistemic Complexity and Causality

Chapter 11
Reasons Against Naturalizing Epistemic Reasons: Normativity, Objectivity, Non-computability

Julian Nida-Rümelin

11.1 Naturalism

In order to present reasons against the possibility of naturalizing reasons it is necessary to present an account of naturalization. Naturalism with respect to a domain D is the view that all entities or properties out of D can be naturalized. There are many different kinds of characterizing naturalism. The broadest account takes *naturalism* as being the view that nature is a coherent whole, and that human beings and all their properties are a part of nature. This account is not quite clear-cut and I would like to avoid answering the question "are you for or against naturalism?" understood in this sense. Aristotle and present day proponents of Aristotelian metaphysics and ontology adhere to this kind of naturalism. More radical Aristotelian views do not only subsume human aspirations, human beliefs and intentions, human reasons under this broadly understood *natural order*, but tend to interpret the natural order itself as teleologically or intentionally structured. To name only one of the more prominent Aristotelians of this sort: Hans Jonas puts his *Imperative of Responsibility* into such a broader teleological account of nature.[1] Deep teleology, the view that all entities in nature are embedded into a broader teleological frame and that explanation of natural phenomena must refer to it, stands undoubtedly in sharp contrast to modern natural science, the practice and the theory of natural sciences. There are many and competing accounts of what explanation in the natural sciences is but there exists an almost unanimous consensus that reference to *telē* cannot be a legitimate part of explanation in the natural sciences. Put differently: Teleological explanation is different from causal explanation and the natural sciences aim at causal explanations only. For example, take game theoretic models in evolutionary theory. Game theory has developed from the analysis of human agents. Utility and probability functions that one can attribute to these human agents constitute

J. Nida-Rümelin (✉)
Seminar für Philosophie Ludwig Maximilians-Universität München, Germany
e-mail: sekretariat.nida-ruemelin@lrz.uni-muenchen.de

[1] Cf. Hans Jonas: *The Imperative of Responsibility. In Search of an Ethic for the Technological Age*. Chicago 1985.

A. Carsetti (ed.), *Causality, Meaningful Complexity and Embodied Cognition*, Theory and Decision Library A 46, DOI 10.1007/978-90-481-3529-5_11,

its conceptual frame.[2] But the evolutionary story is exclusively causal. The talk of "selfish genes" (Dawkin) is merely metaphorical. The causal explanation contains no reference to intentions, aspirations, reasons, *telē*. Explanation in the natural sciences is causal – deterministic or probabilistic –, it deduces *explananda* (natural events) from causes (antecedent natural events) together with natural laws. The *explananda* and the antecedent natural events do not contain intentional states and a fortiori do not contain reasons. The *eidos* or the *telos* of a tree does not explain how it grows whereas the average ankle of the sun rays can be part of the explanation.

In every discipline of natural science there are standards of good explanatory theories. Although there is a wide spectrum of methodological and conceptual differences in this spectrum and notwithstanding the ongoing debate in the philosophy of science about the criteria of good explanatory theories, there is a robust consensus that teleological elements are to be excluded from explanatory theories in the national sciences. This should not be understood as a meta-theoretical position among others but a descriptive element of established scientific practice. Therefore we can use this trait of scientific practice in order to define naturalism. *Naturalism* is the view that the methods of natural science suffice to describe and explain not only those events that are generally accepted as natural events in the sense of being adequate objects for scientific[3] explanation, but also of events that are usually not objects of natural science. In this reading "naturalism" is the meta-theoretical view that all events can in principle be explained by natural science. It is obvious that this meta-theoretical view makes sense only if it is based on a more general naturalistic world view regarding the ontological constitution of the entities and the range of the laws of nature.

In the following we are not concerned with naturalism in general but with the question whether epistemic reasons can be naturalized. If naturalism*[4] were to be true, epistemic reasons could be naturalized. If epistemic reasons could not be naturalized, naturalism* would not be true. In this paper we shall introduce three reasons against the possibility of naturalizing epistemic reasons: The argument from normativity, the argument from objectivity, the argument from non-computability. Before we come to these three arguments in Sections 11.3–11.5, we need to clarify the concept of epistemic reasons in the next section.

11.2 Epistemic Reasons

Epistemic reasons justify (rational) beliefs. Practical reasons justify (rational) decisions. The standard view, mostly called Humean, is that genuine reasons are epistemic ones. I agree with it, but for quite different reasons. The Humean view is based on the idea that practical reasons result from or are even implied by epistemic

[2] Cf. JNR: "Rational Choice: Extensions and Revisions". In: *Ratio* VII (1994), S. 122–144.

[3] Here and in other cases I use "scientific" referring to the natural sciences.

[4] "naturalism*"is the kind of naturalism we described in the last paragraph.

reasons given desires, the desires of the acting person. The person has a good reason to do A if she has a desire that is best fulfilled by her given beliefs. Epistemic reasons qualify beliefs as rational, whereas desires are given. Therefore, for the standard view there are no genuine practical reasons. There are only reasons to believe. There are only epistemic reasons – and these are based on empirical evidence. Reasons to act are derivative.

The *non-standard view* I am arguing for, rejects this dichotomy between theoretical and practical reasons and it rejects the idea of desires as given, desires which cannot be criticised and modified. In giving up the idea of given desires we reject foundationalism regarding practical reasons. The non-standard view is coherentist. The practice of giving and taking reasons is not split into two separate parts with different rules of inference. A reason to act results in a belief that this act would be a good one. Beliefs regarding goodness or rightness or justice refer to (normative) propositions. The fact that some of these normative propositions have practical implications does not change the form of reasoning. Reasons speak for or against a propositional attitude. Some of these propositional attitudes have practical implications in the sense that a rational person having this propositional attitude acts accordingly.

This description of my non-standard view is compatible with a close linkage between theory and practice, between propositional attitudes and actions. Propositional attitudes reveal themselves in acting. Preferences reveal themselves in choices. Wishes reveal themselves in motivations for action etc. A person may say that she believes that p, but if she acts as if p were not the case, we may doubt whether the person indeed has this belief. Reasons are epistemic. Reasons justify propositional attitudes. Propositional attitudes represent practices or, to put it more generally, whole *forms of live*.

Actions, desires and beliefs cannot be ascribed independently from each other, even if there are cases in which the ordinary linkage between these three types of ascriptions dissolves. For example it is a strong stoicist view that desires taken for themselves do not lead to action, that there is an intermediate role for *synkatathesis* or *krisis*, that the deciding person may have wishes or desires, which she after deliberation does not want to be fulfilled. She decides not to fulfil a certain desire, even if this desire does not vanish being confronted with the result of deliberation.

Our everyday language allows for delicate discriminations. We tend to speak of desires if we want to let it open whether in the end we shall decide to act upon them. Whereas we tend to use the term "wish" if it is a propositional attitude that resulted from deliberation already. If our normative judgements differ from our wishes we are confronted with some kind of *incoherence*. A coherent form of life represents a coherent system of empirical and normative beliefs. Wishes are propositional attitudes that are not independent from beliefs. If our normative beliefs and our wishes differ, we try to reconcile them. Against Harry Frankfurt I assume that reconciliation goes *via* reasoning, not via second order desires.[5] Desires are not independent

[5] H. Frankfurt, "Alternate Possibilities and Moral Responsibility", *Journal of Philosophy* 66 (1969). "Freedom of the Will and the Concept of a Person", *Journal of Philosophy* 68 (1971).

from reasons, not even second order desires are independent from reasons. The person acting should be identified with her reasons, not with her (second order) desires. Reasons justify belief. Some of our beliefs are to a more or lesser degree practically relevant. There are no isolated beliefs; instead the whole of our beliefs resembles an organism with many interdependent parts. All beliefs have some practical implications.

Therefore we can conclude in saying that *all reasons can be transformed into epistemic ones*. A reason for acting is ipso facto a reason for the normative belief that this acting is the *right* one. The question whether there are practical implications or not does not discriminate between empirical and normative reasons. Some normative reasons (e.g. in favour of a certain theory of justice) might not have (immediate) practical consequences. But at least in an indirect and implicit way all reasons have some practical implications taken as a whole. The system of reasons corresponds with (justified) practice. If we argue against the possibility of naturalizing epistemic reasons, we argue against the possibility of naturalizing reasons *in general*, including reasons for acting. The reason why we focus on arguments against the possibility of naturalizing reasons for belief is that this case is more obvious. It is easier to show that it is impossible to naturalize epistemic reasons. In the context of this conference the *argument from non-computability* of epistemic reasons may attract more attention. But before we come to this we have to consider two other interdependent arguments: The *argument from normativity* of epistemic reasons and the *argument from objectivity* of epistemic reasons.

11.3 The Argument from Normativity

Epistemic reasons are reasons to believe p. p can be an empirical proposition, a proposition that is based on empirical experience, and then the reasons in favour of p are empirical reasons. If p e.g. is the statistical finding that m out of n individuals are taller than six feet then the belief in p is based on roughly m positive results out of a representative sample of n individuals. We have to believe a lot of other things in order to be justified to belief in p: that the sample is representative; that the collected data are reliable, that they were counted correctly; that nobody involved cheated etc.

If the belief of yours is not justified you *ought* not to have that belief. Justification is obviously a *normative* concept. Moore's open question argument[6] can be applied to every naturalistic explication of justified belief. Every property of your reasoning may be a good indication for you being justified by good reasons, but none of these properties does exclude the question whether it is exactly this property that makes your belief a justified one. Life world and scientific discourses about justified

The "Frankfurt type examples" are discussed in part 5 of *The Oxford Handbook of Free Will*. Ed R. Kane. Oxford 2002. I criticize the concept in *Über menschliche Freiheit*. Stuttgart 2005, chap III.

[6] *Principia Ethica*, Oxford 1903, Kap. II.

and unjustified beliefs, rational and irrational beliefs, well-founded and unfounded beliefs are normative. To deprive these discourses of their normativity would make them senseless, would cut out the issue of discussion.[7] Giving and taking epistemic reasons is about justified belief. Whether a belief is justified is a normative question.

Almost all ethicists agree that naturalism in ethics cannot be upheld. The fact that moral properties supervene on empirical properties does not transform moral properties into empirical ones. Likewise the property *being justified* of beliefs is not a natural property. Epistemic reasons in favour of beliefs decide whether the belief is justified. Epistemic reasons cannot be naturalized. It is not possible to describe and explain instances of *justified beliefs* with the conceptual and explanatory means of natural science because *justified* is not an empirical, but a normative property. The normativity of epistemic reasons speaks against the possibility of naturalizing epistemic reasons. As far as we do not dismiss the established forms of life world and scientific discourses as globally erring, the normativity of epistemic reasons excludes their naturalization. Formulated in a different terminology: from a pragmatic point of view epistemic reasons cannot be naturalized. Only a radically sceptical point of view allows taking a naturalistic stance on epistemic reasons. But even then we would not understand what this naturalistic stance is, since we would miss the concept of a justified point of view, we could not say anymore that the naturalistic stance is justified.

11.4 The Argument from Objectivity

If somebody holds that r is a good reason to believe that p, he thinks that this is *objectively* so. Even if he does not know the concept "objective", he would reject any interpretation of this proposition as being about subjective states of his. If he e.g. says "the fact that uttering S hurts him is a good reason not to utter S" he does not report his subjective attitude, but he expresses the normative belief that it would be wrong to utter S given the fact that the addressee would be hurt. Some argue at this point that the empirical fact that the addressee would be hurt can be a reason not to make this utterance only on the basis of an ethical principle that one has to accept in the first place. I am convinced in the meantime that this assumption is wrong, that we are not in need of a principle that transforms certain empirical facts into reasons for normative beliefs. Principles systematize reasons; they do not form a fundamental basis from which reasons can be deduced.[8] Most ethicists would agree up to this point. They would accept that ethical reasons formulate objective duties that are independent from wishes and desires of the acting person. Some might doubt whether these objective reasons really exist and may retreat to a kind of second-order subjectivism approving that in every day life they give and take reasons as if

[7] Hilary Putnam makes a similar point in "Why Reason Can't Be Naturalized". In: *Synthese* 52, 1982.

[8] Cf. JNR: Philosophie und Lebensform (Frankfurt a.M.: suhrkamp 2009) in print.

they were objective.[9] But, aren't there many reasons for normative beliefs that are undoubtedly subjective? My normative belief that I should go to the travel agency now because later on the price for my holiday flight will rise may be taken as an example for a reason that seems to be subjective. It is my wish to take this flight and the rest follows from the wish alone (given that my expectation concerning the ticket prices is sound). But this is obviously wrong. There is no immediate relation between your wish to spend the holidays in Egypt paying as less as possible for your flight and the normative belief that you ought to go to your travel agency now. There may be reasons to believe that the fulfilment of this wish would cause a lot of troubles for you (terrorism, your friend at home, ...). Some may find the argument convincing that a few days on vacancy do not justify that amount of energy waste. We should treat wishes of ourselves like other empirical properties that can be relevant for giving reasons to believe, but not as reasons themselves. We argue for and against normative and empirical beliefs in order to find out whether these beliefs are well founded or not. The mode of this exchange of reasons is not one of expressing subjective attitudes, but one of finding out what *objectively* speaks in favour of a belief. The form of our everyday practice of giving and taking epistemic reasons makes sense only if we take epistemic reasons to be objective. In fact it seems to me that the common expression "subjective reason" is like "subjective fact" a contradiction in itself. There might be subjective beliefs regarding facts, but there are no subjective facts. There might be facts about subjective states (certain types of mental states), but this doesn't make these facts subjective. The analogy holds for epistemic reasons.

If epistemic reasons are objective they cannot be identified with mental states, a fortiori they cannot be identified with the neuro-physiological correlates of mental states. The critique of psychologism in logics most forcefully formulated by Gottlob Frege and Edmund Husserl, cannot be made plausible by referring to logical inferences and logical languages. It is based on this (you might say "metaphysical") account of objectivity of epistemic reasons for which I argued above. The propositional content of epistemic states is objective. Supposing p to be the case includes the possibility that the speaker is wrong about it. The speaker cannot say "I believe that p, but p is probably not the case". The inferential form of reasons speak in favour of an objectivist account. In other words: reasons are not internal, they are no part of mental states or processes, although *accepting* reasons results in mental states. Reasons are external, but since reasons are normative, they cannot be empirical. Therefore, the objectivity of epistemic reasons speaks against the possibility of naturalizing reasons. Reasons are reasons *for* – in this sense they are inferential. Giving reason for x has the form of assuming a proposition p, that the addressee supposedly agrees with and stating that p speaks in favour of x. Reasons have a *propositional content* and are used *inferentially*. Both the propositional

[9] The most prominent philosopher who took this stance was John Mackie in *Ethics. Inventing Right and Wrong,* Oxford University Press, 1977.

content and the inferential use have to be interpreted objectively. *p* is the case or *p* is not the case independently of epistemic states (if the epistemic states are not themselves part of the propositional content). The inferential use is valid or invalid, independently from epistemic states.

If naturalization is confined to the entities that in principle have a physicalist description and explanation as we argued in section I, subjectivist accounts of epistemic reasons could not count as naturalistic ones. However, naturalism as it is related to Quine's work gives a behaviourist interpretation of mental states and in this way allows for a naturalization of epistemic reasons qua identifying them with causal effects of sensory stimuli. This form of naturalizing epistemic reasons is blocked by this argument from objectivity.

Another form of naturalizing the subjective (the mental) would be neuro-physiological: Mental states are then identified with neuro-physiological ones and reasoning becomes a neuro-physiological process leading from one neuro-physiologically epistemic state to another. This form of naturalizing epistemic reasons is blocked also by the argument from objectivity.

Analogously, all forms of naturalizing epistemic reasons via naturalizing the mental are blocked by the argument from objectivity.

11.5 The Argument from Non-computability

Epistemic reasons speak in favour of beliefs. To make things simple we can assume that epistemic reasoning has the form of a sequel of propositions whereas the validity of epistemic reasoning shows itself in a sequel compatible with inferential rules. Formal logic should be understood as the enterprise to systemize parts of the inferential rules of everyday (and scientific) reasoning. Life world reasoning is usually much more complicated. The interplay of giving and taking reasons is essential for it. The responses show whether the propositional content or the inferential assumptions are accepted by the addressee. If the addressee opposes some propositional content or some of the inferential moves (doubting the validity of it) the reasoner (the person who tries to show that a belief of his is valid) has to respond using new propositional or inferential assumptions. Epistemic reasoning of this kind is usually not algorithmic. It is usually not algorithmic because it contains as essential part reasoning according to rules of formal logic. The theorems of first order predicate logic and a fortiori richer logical languages cannot be proven algorithmically. The sequel of lines necessary to produce the proof of a theorem of first order predicate logic cannot be produced by a turing machine. Proving a theorem of first order predicate logic is obviously one form of epistemic reasoning. More complex forms of epistemic reasoning contain moves that mirror proofs of theorems of first order predicate logic. Even much richer epistemic reasoning cannot dismiss the inferential rules that are systemized by first order predicate logic. Therefore epistemic reasoning cannot be identical with some causal-deterministic neurophysiological process, because causal deterministic processes in principle can be produced by

turing machines. This is obviously true for the classical deductive-nomological model of causal explanation, but it can be extended to more complex models of causal explanation including probabilistic ones.

The validity of the argument from non-computability depends heavily on theories of causal relations. Whereas natural scientists stick to the classical model of causality as algorithmic, philosopher of science developed accounts of causality during the last decades that made causal relations part of epistemic reasoning. These neo-pragmatist versions of causality including Bayesian causality are not at stake here. If they were adopted in the natural sciences, "naturalism" as defined in Section 11.1 would cease to exist. But as far as causality is understood as a relation between natural, empirically accessible events, whereas this relation is lawful and this natural law allows producing the sequel of caused events by a turing machine, non-computability is a strong argument against the possibility of naturalizing epistemic reasons.

Chapter 12
Some Remarks on Causality and Invariance*

Raffaella Campaner and Maria Carla Galavotti

The last few decades have seen a proliferation of theories on causality. Currently, one of the main trends is pluralism, with a few kinds of pluralism highlighting possible intersections and contact-points between different and more or less distant positions. This paper focuses on a notion that has been finding great fortune, being so-to-speak "transversal" to various contemporary approaches to causality and causal explanation, namely invariance. The notion will be used as a lens to view a portion of the latest debate on causation developed with respect to various fields.

Taking as a starting point the debate arising from Woodward's view, the first part of the paper will show how some authors have recently appealed to invariance as crucial to the assessment of causal nexus within specific scientific disciplines. Attention will be drawn to difficulties arising from such attempts and to the different uses they make of the notion of invariance. The second part of the paper will consider further views which provide suggestions as to how to conceive of invariance within a pluralistic perspective and devote due attention to the role of context.

Part 1 Invariance Under Intervention

12.1 Woodward's Interventionist Theory

Roughly in the last decade, Woodward has been developing a manipulative theory of causality which conjugates manipulation and counterfactuals, is meant to be a theory of causal explanation, and also admits of mechanisms. Invariance is the fundamental notion on which the whole theory is grounded. In the most articulated explication of his manipulability conception of causal explanation, presented in *Making Things Happen* (2003), Woodward defines invariance as "the key feature a relationship must

R. Campaner (✉) and M.C. Galavotti
Department of Philosophy, University of Bologna, Italy
email: Raffaella.Campaner@unibo.it; MariaCarla.Galavotti@unibo.it

*Part 1 is by Raffaella Campaner, Part 2 is by Maria Carla Galavotti.

A. Carsetti (ed.), *Causality, Meaningful Complexity and Embodied Cognition*, Theory and Decision Library A 46, DOI 10.1007/978-90-481-3529-5_12,

possess if it is to count as causal or explanatory", and believes a generalization is to be counted as "invariant [...] across certain changes if it holds up to some appropriate levels of approximation across those changes" (Woodward 2003a, p 239). The invariance of the relationship between X and Y under at least some interventions on X is a necessary condition for the relationship between X and Y to be regarded as causal. "Invariance under intervention" is here taken to stand for "invariance under some testing interventions on variables figuring in the generalization", and it is not an all-or-nothing matter: "most generalizations that are invariant under some interventions and changes in background conditions are not invariant under others" (ibidem, p 257). A generalization is regarded as more invariant than another if it is invariant under a "larger or more important set of changes and intervention" (ibidem, p 257). Invariance is hence maintained to come in degrees and to be relative to a class of changes: if the class of changes under which relationship R_1 is invariant is a subset of the class of changes under which R_2 is invariant, then R_2 is claimed to be more invariant than R_1. What changes are taken into account can vary: certain sorts of changes can be regarded as particularly important for the assessment of invariance for a discipline and a subject matter under consideration, but not for another. The more invariant the generalization included in the causal explanation is, the better the explanation itself. In other terms, degrees of invariance affect degrees of explanatoriness as well. Moreover, as in any manipulative approach, a strong interest in controlling is expressed: "wiggling" on a given X that is a relatively invariant cause of Y gives one some control over whether Y obtains.

Causal generalizations are generalizations that are invariant under *some* (actual or ideal) interventions, and can be expressed in counterfactual terms: they are such that they *would have* continued to hold *if* various sorts of changes *had been made* to occur. The notion of invariance is thus presented as a modal notion, having to do "with whether a relationship would remain stable if, perhaps contrary to actual fact, certain changes or interventions were made to occur" (Woodward 2004, p 235). Given that invariance is meant as "invariance under intervention", Woodward defines his account as an "interventionist account", and the counterfactuals appealed to in this analysis of causality are hence labelled "interventionist counterfactuals" or "active counterfactuals". With respect to the type-token issue, Woodward believes that to state "X is causally relevant to Y" is to state that changing the value of X instantiated in particular, spatiotemporally located, individuals will change the value of Y instantiated in other particular individuals. "The truth of a claim such as (S) 'smoking causes lung cancer' depends on relationships that do or would obtain (under appropriate manipulations) at the level of particular individuals", but it is also assumed that "the claim (S) would be true even if no one were to smoke, as long as it is the case [...] that manipulating whether some particular human being [...] smokes will change whether they develop [...] lung cancer" (Woodward 2003a, p 40).

According to Woodward, in order to qualify as an intervention on X with respect to Y, a manipulation must change the value of X in such a way that if the value of Y changes, it does so *only because* of the change the manipulation produced in X, and by no other means. The parts of the system to which X and Y belong must operate independently enough to allow an exogenous cause to change the values of

X without producing changes in other parts of the system which can influence the value of Y independently of the manipulation of X. Modularity is hence required to hold: "a system of equations will be modular if it is possible to disrupt or replace (the relationships represented by) any one of the equations in the system by means of an intervention on (the magnitude corresponding to) the dependent variable in that equation, without disrupting any of the other equations. [...] It is natural to suppose that if a system of equations correctly and fully represents the causal structure of some system, then those equations should be modular" (Woodward 2003a, p 48). The modularity requirement provoked a wide debate: it is controversial whether such a requirement is as "natural" as Woodward claims.[1]

Without playing a central role, causal mechanisms are also admitted, and described as "organised or structured sets of parts or components" (Woodward 2002a, p S375), governed by invariant generalizations. Such generalizations are the key to explanation, which is conceived of as growing out of highly practical interests. Information acknowledged as relevant to explaining an outcome causally involves the identification of factors such that, *if* manipulations of these factors *were* possible, these manipulations *would* prove a way to alter the phenomenon in question. An explanation is thus an answer to a *what-if-things-had-been-different* question. According to both Woodward and Christopher Hitchcock, this theory of explanation has a number of virtues over other views,[2] and is able to make sense of the intuition that some explanations are deeper than others. Such a merit – it is argued – rests on the fact that "one generalization can prove a deeper generalization [...] if it is invariant under a wider range of interventions", that is, if it is more general, where generality is to be understood "with respect to hypothetical changes in the system at hand" (Hitchcock and Woodward 2003, p 198).

Woodward's appeal to invariance echoes economic and econometric literature and his position is widely discussed. Among critics, we can recall Stathis Psillos, Paul Humphreys and Jim Bogen, who have all raised objections against Woodward's use of counterfactuals. While believing that counterfactuals and mechanisms can be combined in a theory of causality, Psillos criticises Woodward for characterizing counterfactuals for causal analysis in terms of experiments. More precisely, he accuses him of keeping evidence-conditions and truth-conditions apart: evidence-conditions of Woodward's counterfactuals are fully specified in terms of experiments, whereas truth-conditions are not. The problem arises from such statements as the following: "doing the experiment corresponding to the antecedents of [counterfactual claims] doesn't *make* [them] have the truth-values they do. Instead the experiments look like ways of *finding out* what the truth values of [the counterfactual claims] were all along. On this view of the matter, [counterfactual claims] have non-trivial values [...] even if we don't do the experiments of realizing their antecedents. Of course, we may not *know* which of [two counterfactual claims] is true and which false if we don't do these experiments and don't have evidence form some

[1] See, e.g., Cartwright (2001, 2007).

[2] See Woodward and Hitchcock (2003).

other source, but this does not mean that [they] both have the same truth values" (Woodward 2004, p 46). Psillos concludes that, while giving us a relatively detailed account of the evidence-conditions of counterfactuals, Woodward does not provide anything remotely like that for their truth-conditions, thus implicitly maintaining that "there is something more to causation – *qua* an intrinsic relation – than just invariance under intervention" (Psillos 2004, p 302).

Humphreys' objections arise from a distinction between explanation and understanding, believing that Woodward's emphasis on non-actual situations is more appropriate to understanding than to explanation. According to Humphreys, counterfactuals can be useful to increase our understanding, but not to elaborate explanations. For example, although we all agree that it is a law-like relationship that opposite charges attract, "consider what would happen if some hypothetical intervention occurred so that opposite charges did not attract one another is to enter a realm *of no relevance to an explanation* of why these two charges attracted. Here, then, is perhaps where one part of the boundaries between explanation and understanding lies. Although it can enhance our scientific understanding to explore models that violate the laws of our universe, such models cannot be used in explanations" (Humphreys 2006, pp 42–43).[3]

Objections to basically the same effect are raised by Bogen, who too maintains that in an adequate causal explanation we reveal what, *as a matter of fact*, links X and Y, by virtue of the fact that the one brings about the other. Bogen is strongly against the idea that whether one thing causes another "depends in part upon what would have been [...] the case if something that did not happen had happened" (Bogen 2004, p 3). For Bogen, it is something "factual", which is "the opposite of counterfactual", that does all the explanatory work.

In Woodward's view, any theory of causal explanation whatsoever needs to account for causal and explanatory *relevance*, and the relevance is what counterfactuals provide. More serious criticisms Woodward is called to face regard his theory's invariance assumptions, not adequately justified, and his failing to distinguish between genuinely experimental situations and observational ones. While clearly acknowledging that in many circumstances actual manipulations are impossible, right at the beginning of *Making Things Happen* Woodward says: "my idea is that one ought to be able to associate with any successful explanation a hypothetical or counterfactual experiment: [...] the explanation must enable us to see what sort of difference it would have made for the explanandum if the factors cited in the explanans had been different in various possible ways. [...] Our interest in causal explanation represents a sort of generalization or extension of our interest in manipulation and control from cases in which manipulation is possible to cases in which it is not" (Woodward 2003a, p 11). The problem seems to be sidestepped too quickly. Elsewhere he states: "invariance is a *relative* notion: a relationship can be invariant under one set of interventions or changes in background conditions (or as I will say under some domain or regime) and not others. A relationship can thus be invariant within some domain without being exceptionless and universal in the way that,

[3] See in the same volume also Sober (2006) and Woodward (2006).

according to many philosophers, laws of nature are" (Woodward 1997, p S33, italics in the text). That a relationship proves invariant under just *some* interventions seems too weak a requirement for causal explanations, aimed at tokens. Furthermore, "domains" would need specifying. Such crucial issues, and possible alternative perspectives in which the context is acknowledged a central role, will be addressed in Part 2 below.[4]

12.2 Invariance and Stability in Biology

Woodward has engaged in an interesting debate with Sandra Mitchell, who has been working on the presence and role of laws in biology and on the form explanation takes in such a discipline. The debate between the two has been centred on differences – and possible analogies – between the idea of stability supported by Mitchell and that of invariance advanced by Woodward.

According to Woodward, as in other disciplines (belonging both to the physical and the social sciences), in biology too explanations must appeal to generalizations that satisfy his requirement of invariance.[5] Such generalizations do not need to satisfy the criteria usually required for lawfulness, such as being exceptionless or supported by a wider theory; invariant generalizations suffice to figure in biological explanations. Sandra Mitchell puts forward the idea of "stability", which, though far from identical to that of invariance, shares some interesting aspects with it. Mitchell stresses, first of all, that biological systems are highly complex: they are evolved, multi-component and multi-level and "their features are, in large part, historically contingent. Their behaviour is the result of the interaction of many component parts that populate various levels of organization from gene to cell to organ to organism to social group" (Mitchell 2002, p 329).[6] High complexity is claimed to have implications on obtaining scientific knowledge and scientific representations of biological systems, which fail to be characterised – according to Mitchell – by universal, exceptionless, true and necessary generalizations. Should we then conclude that biology is completely lawless? Mitchell answers in the negative: for example, patterns of behaviour in a social insect colony or patterns of genetic frequencies over time in a population subject to selection are caused, predictable, explainable, and used to manipulate biological systems reliably, and all this can be made sense of by a *revised* notion of law.

Unlike what is usually maintained, laws cannot be detached from their evidential context and claimed to apply to all regions of space-time. According to Mitchell, the difference between laws and accidental generalizations "is better

[4] See also Galavotti (forthcoming), Sections 3.b and 5.

[5] See, e.g., Woodward (2000, 2001).

[6] Contingencies "are as important to good sciences as are the regularities that can be abstracted from distributions of their contextualised applications" (ibidem, p 330).

rendered as degrees of stability of conditions upon which the relations described depend" (Mitchell 2000, p 257), and this affects their acceptance and application to further situations. We must "rethink the idea of a scientific law pragmatically or functionally, that is, in terms of what scientific laws *let us do* rather than in terms of some ideal of a law by which to judge the inadequacies of the more common (and very useful) truths" (Mitchell 2002, p 333). The modified sense of law Mitchell suggests is meant to give adequate importance to contingencies generalizations depend upon. "The problem of laws in the special sciences is not just a feature of our epistemological failings; it is a function of the nature of complexity displayed by the objects studied. [...] We can [...] ask not whether biological claims can be transformed into strict laws, but rather when and how do biological claims perform the functions that laws are thought to serve? That is, how can less than strict laws explain, predict and assist in intervention?" (ibidem, pp 343–344).

Woodward's and Mitchell's views share some similarities. They both: (a) reject universality and exceptionlessness as necessary for generalizations to be deemed laws and to explain; (b) have arguments against the *ceteris paribus* clause, which unduly forces laws into the standard view;[7] (c) believe that generalizations have properties that come in degrees.

According to Woodward, the relevant continuum is that of invariance, and having *some* domain of invariance is sufficient for explanation. Mitchell, on her hand, does not identify a single continuum, but *a number of* different continua "that generalizations can be located within – in particular ontological ones of stability and strength, and representational ones of abstraction and cognitive accessibility" (ibidem, p 345). As an example, she mentions Fourier's law of thermal expansion and Mendel's law of segregation. The difference between them is claimed not to lie in the one functioning as a standard law and the other not, or in the one being necessary and the other contingent, but in the stability of the conditions upon which the relations are contingent: "Once the distribution of matter in the primordial atom was fixed, presumably shortly after the Big Bang, the function described by Fourier's laws would hold. [...] Instead, the conditions that both gave rise to the evolved structure of sexual reproduction and meiotic process of gamete reproduction are less stable" (ibidem, p 345). Mitchell draws methodological consequences from variations of degree in stability and strength, stressing that we very often cannot detach the relations we discover from their evidential context. Evidence must be carried from the contexts of discovery and confirmation along to new situations and "as the conditions required become less stable, more information is required for application. Thus the difference between the laws of physics, the laws of biology, and the so-called accidental generalizations is better rendered as degrees of stability of conditions upon which the relations described depend" (ibidem, p 345).

In sum, stability is defined by Mitchell as "a measure of the range of conditions that are required for the relationship described by the law to hold", which she takes to include "the domain of Woodward's invariance. However, stability can be

[7] See Mitchell (2002) and Woodward (2002b).

a feature of relationships that are not invariant under ideal interventions" (ibidem, p 346). On the other hand, for Woodward mere stability under some or even many changes is not sufficient for explanatoriness. Mitchell believes that the disagreement between them can be fundamentally traced back to the fact that in Woodward's view a causal generalization reports a counterfactual dependence, it describes a relationship which remains true under certain episodes of "other things being different" and which does not need to be true under all such episodes. In other terms, "Woodward lets domain restricted generalizations count as explanatory in just those domains where the relationship described in the generalization holds. Stability does just the same work, however it is weaker and includes what might turn out to be correlations due to a non-direct causal relationship" (ibidem, p 347). While in Woodward's account the notion of invariance under interventions is meant to do the crucial work of distinguishing between genuinely causal from merely accidental generalizations, a work that "is done by the notion of a *law of nature* in other philosophical accounts" (Woodward 2003a, p 16), stability is given a less ambitious task, with no specific link to causality whatsoever.

Whereas Woodward does not attempt to specify what the context exactly amounts to, Mitchell tries to do so: "the context sensitivity of complex dynamical systems, like those studied by biology, entails a shift in our expectations. We should not be looking for single, simple causes. We should not be looking exclusively for universal causal relationships. And we must record and use not only the causal dependencies detected in a particular system or population to understand other systems and populations, but also the features that define the contexts present in the system under study". To understand what happens in a given domain, the different ways in which it can be restricted must be considered, such as "temporal and spatial restrictions that are the result of the evolutionary process; contextual restrictions in which certain parameter values or background conditions change the functions that describe the causal dependence; and contextual restrictions in which the operation of other causal mechanisms can interact in ways in which the effects of a cause are amplified, damped, made redundant or evoked". How to account for "all that variety of contingency" (Mitchell 2002, p 348) adequately and precisely seems to remain an open problem.

12.3 Social Norms and Invariance

An appeal to the notion of invariance as a fundamental clue to explanation was also recently made by David Henderson, who pursues the idea with respect to the social sciences. The framework Henderson chooses is the erotetic approach to explanation, according to which explanations are answers to why-questions. As is well known, in such approach both the questions and the answers (i.e. the constituents of explanations) are claimed to strongly depend on the context in which they are formulated. Why-questions having to do with actions of an individual or a group of individuals are examples of why-questions frequently asked in the social sciences: why did the agent do such and such? why did those folks do such and such? While

adopting it, Henderson raises a common objection against the erotetic approach: the (contextually appropriate) relevance relation which should ground the explanation is usually not sufficiently spelled out, the approach thus running the risk of accommodating *any* state or event as explanatory with respect to *any other* state or event. Descriptions of social norms – Henderson argues – can solve the problem: they qualify as explanatory and can serve as answers to such why-questions.

Norms are defined by Henderson "dispositions to coordinated patterns of action and evaluation within some group of people. To characterize a people as having such-and-such a norm is to characterize a pattern of action exhibited in the 'fitting circumstances', and to say that members of the group have dispositions to conform to such a pattern and to evaluate action (or actors) with respect to its conformity"; "descriptions of norms are, in effect, generalizations regarding a group's coordinated dispositions to action and evaluative stance-taking" (Henderson 2005, pp 327, 330). Here too contingencies are brought to the foreground: it is emphasized how social sciences are concerned with systems that result from historical situations, and how the regularities governing the behaviour of individuals or groups cannot but be the result of historical and social contingencies.

Regularities are strongly dependent on background conditions, and certain changes in those conditions can destroy or markedly modify the systems. Which features of social norms make them, then, explanatory? Explicitly drawing upon Woodward's theory, Henderson holds that the degree of invariance exhibited by norms does. High contingency of norms notwithstanding, "some significant ranges of changes in background conditions would not lead to breakdowns – and thus, within limits, the systems and regularities have a degree of stability" (ibidem, p 325). As an example, Henderson supposes that it is a norm among adolescent males in a given community not to show deference to males in a position of authority. Several adolescent males are stopped in the hall of a school by a teacher and asked a question regarding their plans for that afternoon. Each replies in a highly disrespectful way. Had a different authority figure inquired (another teacher, the headmaster, an administrator, a security guard, and so on), the adolescents would have given a similar response. "Descriptions of norms provide what is, in effect, a generalization regarding the kind of historically contingent system – a group or society. Such a generalization has a significant and interesting degree of invariance" (ibidem, p 324). What if someone were to object that norms themselves are precisely what social sciences want to explain, and cannot hence play an explanatory role? Henderson believes the story should be just the same: in addressing the issue of explanation of norms, social sciences have usually to appeal to further invariant generalizations.

In Henderson's account too invariance comes to be closely related to explanation and, through it, to the context: "substantive analyses of erotetic explanations are, it seems, *not general analyses of explanation simpliciter*, but *analyses of explanations-in-a-particular-kind-of-context*"(ibidem, p 324, italics in the text). The link with causation, although not insisted upon, is considered, recognising that the relevance relation sought to formulate an answer to a why-question is often one of causation. Henderson's view gives rise, though, to some doubts: How should one define the

"degree of invariance" that descriptions of social norms are said to exhibit? What changes are admitted in social norms and, thus, in their descriptions? Henderson simply claims that descriptions of norms are to be understood as generalizations with "*appropriate* degrees and kinds of invariance", which "hold across *some range* of possible interventions" (ibidem, pp 327, 334, italics in the text). These are all open questions.

12.4 Invariance and Intervention in Reasoning and Learning

Both Mitchell's and Henderson's accounts deal with invariance with respect to causal generalizations. Let us now turn to two authors interested in *models* in psychology, Steven Sloman and David Lagnado, who focus on the notion of invariance with respect to human reasoning, learning and use of language. In doing so, they inextricably link cognition with invariance, and invariance with intervention. Cognition is claimed to depend on what does not change; perception to involve discovering the hints that consistently signal things of interest to us; prediction to require identifying variables whose behaviour is constant over time (in order to derive their future behaviour from their present one); explanation to involve "assimilating an observation or phenomenon to a process or representation that applies generally, that emanates from or instantiates relations that are regular. Perhaps most important, control requires knowing the systematic relations between actions and their outcomes, so the right action can be chosen at the right time. In all these cases, the secret is to identify and use invariance, the constant, regular, systematic relations that hold between the objects, events, and symbols that concern cognition. [...] The hypothesis we pursue here is that the invariant that guides human reasoning and learning about events is causal structure. [...] Causal structure is part of the fundamental cognitive machinery" (Sloman and Lagnado 2004, pp 3–5).[8]

Although they stress that causal models are not the only kind of useful tools for representing the world, Sloman and Lagnado maintain that people use them more than other sorts of models when learning and reasoning about events. This claim – they state – is an empirical one: they refer to a series of experiments aimed at demonstrating that causal explanations quickly become independent of the data from which they are derived, assuming a general value. Some experiments are also mentioned to reveal belief perseveration in the face of discredited evidence, since participants are shown to continue to assert the relations they had causally explained regardless of

[8] Sloman and Lagnado do not refer directly to Woodward's view. For hints on the application of his interventionist theory to psychological cases, see Woodward (2007). Appealing to experiments, Woodward maintains that the notion of causation as advanced by 'interventionism' "is involved in or connected with our ability to separate out means and ends in causal reasoning" and "even young children are able to reason causally about the consequences of combinations of interventions" (p 23).

updated information. In other terms, causal beliefs are shown to shape our thinking even when they come to be divorced from observation.

How do people manage to acquire knowledge of causal structure, given that the correlational information provided by observations is seldom sufficient? This is one of the crucial questions Sloman and Lagnado want to answer, and that is where intervention comes into play. Even more closely than to observation of correlations or temporal order, causation is linked to action, and the learning of invariant structures is linked to intervention. Acting on variables in the world affords us a critical cue to causal status, and the ability to exert control over certain variables in our environment allows us to identify possible confounding causes. Furthermore, by wiggling a putative cause and observing subsequent wiggling of an effect, we can rule out the possibility that these are both effects of an unknown common cause. Although intervention can also increase the difficulty of learning, Sloman and Lagnado show how in general experiments demonstrate an advantage of intervention over observation. Stressing the fundamental difference between experimental and observational regimes, they describe some experimental settings in which, as a matter of fact, interveners performed much better than observers, proving that intervention facilitates learning. Why is this the case? It is argued especially that intervention permits us to discriminate between structures that are impossible to distinguish by observation alone. Whereas observation of a correlation between two variables X and Y is not sufficient to establish whether X causes Y or vice-versa, interventions on these variables can tell. "Even with complex networks people were much better at inferring causal structure when they were able to intervene. It appears that the dynamic nature of the display, and of people's interaction with the system, facilitates quick detection of the relevant dependencies. [...] Moreover, once they had established these dependencies, people were able to select additional interventions that allowed them to infer unique structures" (ibidem, p 48).[9]

Sloman and Lagnado conclude their reflections on reasoning and learning with the following considerations: "Prediction and control come from knowing what to *manipulate* to achieve an effect and how to perform the manipulation. When manipulations are dangerous or consume substantial resources, the ability to perform the manipulation mentally can be invaluable. *Causal models* are intended to represent the knowledge necessary to perform this function. [...] If we're right that the major source of *invariance* in human experience are the *causal principles* that generate the *mechanisms* that govern what we observe, then *causal structure* seems the place to look to find out what people are sensitive to. The existence of a coherent and powerful theoretical framework to do causal analysis gives cognitive scientists a foothold on representing that structure" (ibidem, pp 49–50, italics added). The notion of "causal structure" used here extensively rests on Judea Pearl's view[10] discussed in Section 12.5 below. Sloman and Lagnado's declared target is to stress a number of elements: the value of causal modeling in a variety of domains; the impor-

[9] See also Lagnado and Sloman (2004).

[10] See also Sloman (2005).

tance of agency and action in human thought; the importance of understanding the effects of action; and the distinction between action and observation. Interventions are held to be essential for correct detection of causal nexus, which are regarded as the key to the invariant structures aimed at by human perception and knowledge. Such notions as those of causal structure, manipulation and invariance appear hence to be inextricably linked in the context of psychology, and the precise ways in they are intertwined, and appeal to each other, require further reflections.

From what has been presented above, it emerges that attempts to analyse causality within different scientific disciplines shape what the notions of invariance and interventions are taken to mean and what they imply to capture causation. A few issues are not addressed in a univocal fashion and call for deeper inspection. Among them, we can recall the relationship between type and token causation, the role of theoretical assumptions underlying models, the distinction between experimental and observational domains, and the characterization of the context. More on these crucial issues is to be found in Part 2.

Part 2 Invariance and Context

This second part of the paper first addresses the approach to causality taken by statisticians, who devote great attention to the subject. It is argued that the main lesson to be learned from statisticians is their recommendation to make explicit the invariance assumptions that are made when establishing causal relationships and building causal models. There follows an overview of Patrick Suppes' viewpoint, which is deeply pluralistic and urges the importance of context in connection with a useful account of probabilistic causality. Finally, some remarks are made on the notion of context, whose importance is increasingly acknowledged by the literature on causality.

12.5 Causality, Invariance and Statistics

Statisticians have traditionally been concerned with causality, and have developed a number of causality theories. Statistical literature establishes a strong association between causality, manipulability and invariance, and devotes great attention to the distinction between type and token causation. The statistician Irving John Good was one of the pioneers of probabilistic causality with two articles published in 1961–1962 under the title "A Causal Calculus". His approach starts precisely from the distinction between two kinds of probabilistic causality: the *tendency* of an event of a certain kind to cause another event, and the *degree* to which one particular event caused another event. While the tendency to cause concerns types of events (*type*, or *general*, causation), the degree of causation concerns single events, considered after they have occurred (*token*, or *singular*, causation). Good regards them as dif-

ferent types of causal analysis, to be defined and measured by different conceptual tools.[11] Remarkably, Good does not establish a strong link between causation and explanation and defines a notion of "explicativity" taken as "the extent to which one proposition or event explains why another one should be believed."[12] According to his perspective, causality and explanation are related, but not identical. This stance seems to be shared by most authors working within statistics.

Among them Judea Pearl, who developed an influential theory of causality. Building on the techniques of representing statistical associations by means of graphs, Pearl suggests that causal relationships be represented by means of "directed acyclic graphs" (DAG), also called Bayesian networks "to emphasize three aspects: (1) the subjective nature of the input information; (2) the reliance on Bayes' conditioning as the basis of updating information; (3) the distinction between causal and evidential models of reasoning, a distinction that underscores Thomas Bayes' paper of 1763" (Pearl 2000, p 14). The causal interpretation attached to such networks results from a combination of the functionalist notion of mechanism and manipulability. Put briefly, causal Bayesian networks are taken to represent ordered structures of variables exhibiting certain stability conditions, which can lead to manipulations. Such a "mechanism-based conception of interventions" (ibidem, p 24) is the cornerstone of a theory of causality that regards the latter as a useful instrument for prediction and intervention. A crucial feature of this approach amounts to the clear-cut distinction between "seeing" and "doing" underlying Pearl's treatment of causation, where the quantities determined through observation are systematically distinguished from those obtained through experiment. This distinction plays a crucial role in connection with prediction of the results of controlled experiments from observed probabilities, which is the main task of causality.

Pearl also considers the explanatory use of causal models "to provide an 'explanation' or 'understanding' of how data are generated" (ibidem, p 25), or to convey information on "how things work" (ibidem, p 26). A crucial role is assigned to the stability of causal structures, which should be durable over time and *invariant* across a variety of situations. Models characterized by such features of robustness allow for predictions and manipulations that are bound to hold for wide ranges of circumstances. So conceived, "the explanatory account of causation is merely a variant of the manipulative account, albeit one where interventions are dormant" (ibidem). Remarkably, Pearl's recent work on "actual causes" and explanation, in collaboration with Joseph Halpern, reaches the conclusion that when taken in the explanatory sense causality is context dependent. This simply follows from the fact that the whole edifice of causation is made to rest on modeling, which in turn requires various assumptions so strictly linked with the context as to justify the claim that "the choice of model is a subjective one" and "depends to some extent on what the model

[11] Good's theory of causality was first developed in Good (1961–1962). A more accessible version was published together with other articles on causation in Good (1983). See also Good (1988).

[12] Good (1983), p 219.

is being used for" (Halpern and Pearl 2005, p 878). Furthermore, explanation is defined "relative to the agent's epistemic state" (ibidem, p 897).

Although well received, especially among computer scientists and psychologists, Pearl's theory of causation provoked much debate. Its most controversial aspect lies in the Markov condition required by DAGs, whose assumption is considered by many authors problematic in a number of situations.

Another important theory developed within the statistical literature is the so-called *potential response* (PR) approach of Donald Rubin, Paul Holland and others.[13] This is a model for inferring the *effects of causes* (type causation) that makes use of counterfactual reasoning couched in statistical terms. As described by its most resolute opponent, namely Philip Dawid, the "special feature of the PR approach is that it represents a response ('effect') variable Y by two or more random variables, one for each of the possible values of the 'cause variable' X" (Dawid 2007, p 510). The idea is to compare the (assumed) values of potential responses with the effects observed in experimental units, under appropriate assumptions.

Potential responses are defined counterfactually, namely the model assumes that even though a certain response is observed (say a certain subject u has given a certain response to drug treatment) "there is still a fact of the matter about what the subject's u response would have been, had she been given [a different] treatment" (Psillos 2004, p 303). Such a counterfactual assumption is really the gist of the PR method, which derives from it its capability of treating also *causes of effects*. Stathis Psillos, who has called the attention of philosophers of science to this method, quite popular among statisticians, deems it "a big step forward" (ibidem, p 307), and argues in favor of a theory of causality that combines counterfactuals and mechanisms.

By contrast, the counterfactual assumption underlying the PR approach is judged metaphysical by Philip Dawid, who opposes to it a *decision-theoretic approach* (DT), which is entirely in terms of conditional probabilities and expectations based on information, known or knowable. The DT approach avoids counterfactuals, to use only "models and quantities that are empirically testable and discoverable" (Dawid 2000, p 408). According to Dawid, type causality, or the analysis of effects of causes, can be done entirely with the DT machinery. However, the same cannot be said of token causality, or the analysis of causes of effects, which is more strictly related to counterfactuals. In that connection the DT approach leaves some problems open.

Obviously, all statistical methods for establishing causal links require invariance assumptions of various kinds. Dawid mentions the following as assumptions that are often made: "unit homogeneity" (for instance, experimental units of people who are treated for headache with a certain drug); "temporal stability" (constancy of response to treatment over time); "causal transience" (the effect of causes and measurement processes in control groups is transient and does not affect the response to treatment measured later); and "constant effect" (the effect of treatment on each and every experimental unit is the same). Obviously, the extent to which

[13] See for instance Holland (2001) and the bibliography included.

invariance assumptions like these hold within a certain context is a crucial and often controversial matter on which the soundness of the conclusions that are drawn entirely depend. It is therefore recommended that whenever adopted such assumptions should be explicitly stated.

Dawid draws attention to the need for a systematic separation between observational and interventional regimes. The task of causal analysis, he claims, is to use past data to make choices about future interventions, and "... this requires that we understand very clearly the real-world meaning of terms such 'observational regime' or 'interventional regime', since there are many possible varieties of such regimes" (Dawid 2007, p 529). One important assumption that is commonly made is that invariance holds across observational and experimental situations. This is a fundamental assumption underlying the PR approach. As we saw, the same assumption is to be found within Woodward's theory of causality. In Woodward's words: "using experiments to learn about causes requires that some relationships remain stable or invariant across the manipulated and unmanipulated systems" (Woodward 2003b, p 102). The uncritical adoption of invariance assumptions of this kind is sharply criticized by Dawid, who regards the fact that invariance across different regimes (observational and interventionist) is not essential to the application of the method as a merit of his own DT approach. As he puts it: "DT never *obliges* us to make any connexions between distributions across different regimes, unless we consider it appropriate" (Dawid 2007, p 526). Without going into technical details, it is worth pointing out that strong invariance assumptions like the one under consideration are questionable.

To sum up, Dawid's analysis of probabilistic causality highlights the fact that modelling of "statistical causality" requires various invariance assumptions, which should be fully specified case by case, because they vary according to different modelling techniques. However, the distinction between observational and interventionist regimes and the justification of invariance assumptions can only be accomplished with reference to the context in which one operates. As stated by Dawid: "... appropriate specification of context, relevant to the specific purposes at hand, is vital to render causal questions and answers meaningful" (Dawid 2000, p 422).

A similar conclusion is reached by the literature on causality produced by econometricians, who have been traditionally concerned with the topic. A well known theory of causality is that of the econometrician Clive Granger,[14] who holds that the suitability of a definition of causality can only be asserted on pragmatical grounds: "the effect of a causal definition on the decisions taken by a decision maker in a realistic setting is the only way that its usefulness can be discussed" (Granger 2007, p 295). Like other econometricians, Granger regards causality as strictly linked with manipulation because it provides solid grounds for decisions regarding economic policy. Nevertheless, he insists that causality should not be *identified* with control over economic variables. Against other authors, who establish a strong connection between causality and control, he argues that "many writers seem to want to find causality so that the relationship obtained can be used to control or manipulate the

[14] See especially Granger (1980).

effect. However, for this to succeed one needs the assumption that the relation does not change when control is attempted" (ibidem, p 294). In other words, Granger also warns us that the assumption of invariance across different regimes cannot be taken to hold in all situations, and needs justification. Granger ends up with the recommendation that the assumptions underlying causal attributions, as well as prediction and decisions on policy making, be made explicit and justified.

12.6 Invariance and Causality Within Suppes' Pluralistic Epistemology

As suggested by the title of his monumental book *Representation and Invariance of Scientific Structures*, the notion of invariance is assigned a crucial role by Patrick Suppes, upholder of a pluralistically oriented epistemology and one of the pioneers of the probabilistic approach to causality. Inspired by the pragmatist conception of science as a perpetual problem-solving activity, Suppes believes that scientific theories are constructs which "like our own lives and endeavours... are local and are designed to meet a given set of problems" (Suppes 1978, pp 14–15). In this spirit, starting from the early 1960s Suppes developed a pluralistic view of theories based on models, according to which theories are representable by means of a hierarchy of models characterized by different degrees of abstraction, which range from empirical models, or "models of data" describing experimental evidence, to abstract mathematical models characterizing the theory. The models linking a theory to empirical phenomena can be shown to preserve a certain structure under certain operations. In the author's words: "the structure of a set of phenomena under certain empirical operations is the same as the structure of some set of numbers under arithmetical operations and relations" (Suppes 1967, p 59). Invariance, taken as the capacity to preserve structure, is therefore a pivotal feature of this view.

It is important to note that Suppes regards "empirical structures" as an object of investigation no less important than "logical structures", which had for a long time been privileged by philosophers of science. This move, which dates back to the Sixties, extended epistemological analysis to experimentation, including measurement, techniques for the analysis of data, statistical methodology for testing hypotheses, and the like. These components of investigation had been relegated by logical empiricists to the "context of discovery" and excluded from the realm of epistemology, which they saw as concerned with the sole "context of justification".[15]

Conceived as an essential ingredient of the representation of scientific structures, invariance is not only relative to some group of transformations, being "more fundamentally and essentially relative to some theory" (Suppes 2002, p 105). In other words, "invariant properties of models of a theory are the focus of investigations of invariance" (ibidem). This brings out the relevance of context in this perspec-

[15] See Galavotti (1994, 2006) for a discussion of Suppes' philosophy of science.

tive. Within Suppes' pluralistic viewpoint context plays a pivotal role. Epistemology has a local character: epistemological notions should be analysed within a specific context. It is stressed that the complexity of phenomena and the variety of practical situations in which phenomena are investigated are such that important notions of science, and philosophy too, cannot be forced into some definition given once and for all. In other words, there is no unique way of representing scientific structures; on the contrary, a multiplicity of representations can be produced, resulting in a multi-faceted view of scientific knowledge.

Equally pluralistic is the account of causality given by Suppes, who was one of the first to conceive it in probabilistic guise. In his classical monograph *A Probabilistic Theory of Causality*, published in 1970, he does not attempt to work out a univocal definition of probabilistic causality, but gives a flexible, context-dependent account which is centred on models rather than laws. Suppes does not impose particular requirements on causal chains, and claims that causality can be defined both in terms of random variables and events, without specifying in a univocal fashion what counts as an "event". Remarkably, no "ultimate genuine causes" are contemplated. By contrast, the notion of cause, genuine or spurious, is strictly linked to the specification of the set of concepts on which the set of events that can serve as causes in a given context is to be defined. In other words, both the notion of event and that of cause are linked to the specification of the set of concepts characterizing a given context.

As to the specification of context, Suppes calls attention to the distinction between those contexts that are characterized by extensive experimentation (with randomization) and contexts in which experimentation cannot be performed and data are obtained only through observation. He brings out this distinction in the course of a discussion of the problem of passing from type to token probabilistic causation. In spite of the fact that this passage is unwarranted, predictions of individual events based on type kind relations are made every day, for instance by meteorologists. According to Suppes, such inferential procedures depend ultimately on "judgment as to how the knowledge one has is used and assessed" (Suppes 1984, 1993, p 130). If such knowledge includes widely accepted generalizations (or statements of correlations) backed by extensive scientific experimentation, the shift from type to tokens will be more warranted, but there will always be additional elements to be evaluated case by case. If the information available does not stand on a strong experimental basis, additional care will be required. Other elements, such as the qualitative complexity and diversity of cases under investigation, also matter.

12.7 Towards a Characterization of Context

Not surprisingly, a full-fledged account of context is not available in the literature. There are myriad elements that enter into the characterization of context, and it seems hopeless trying to fix a definition of "context" once and for all. Nevertheless, a number of relevant elements can be enumerated.

An important element is the level of detail required (or desirable) for the description of the phenomena under investigation. The dependency on context imposed by the level of detail of explanatory accounts has been stressed by Bas van Fraassen, upholder of a *pragmatics of explanation* that regards explanation as a "three-term relation, between theory, fact, and context" (van Fraassen 1980, p 156). An explanation is an answer to a why-question posed by somebody to somebody else. What kind of answer is sought and offered depends on the kind of information available to the actors involved, the purpose behind the request of explanation and the use to be made of it. Although van Fraassen's pragmatics concerns explanation and not causation, it could easily be extended to the latter.

As a corollary of van Fraassen's pragmatics, it should be admitted that the conceptual setting describing the phenomena under study also matters to the characterization of context. If reference is made to some accepted theory, the set of laws governing such phenomena should be specified. Obviously not all causal attributions – indeed, only a few – are grounded on widely accepted theories. In most cases one relies upon generalizations that do not meet with general consensus, or are accepted subject to constraints of various kinds. It should be added that the notion of "law" is far from being amenable to a satisfactory characterization, as testified by the ongoing debate on the topic.[16] Given the difficulties besetting it, an increasing number of authors turn to the idea that the notion of "law" is context-dependent. Following Suppes, we can identify one such contextual element with the degree to which the correlations (laws) adopted to support causal attributions are supported by (or amenable to) experimental testing.

A further element that enters in the characterization of context is the nature of the available information. Data can be obtained in a variety of different ways, including direct and/or indirect evidence, observation and/or experimentation, simulation, analogy, and so on. Suffice to think of the complex procedures leading to the collection and exploitation of statistical data.

Furthermore, it should be admitted that the disciplinary context surrounding empirical research determines various peculiarities. There is no doubt that within different disciplines different methods, terminology and practices are preferred, partly due to consolidated habits and traditions. In addition, different disciplines retain different aims of enquiry, especially in connection with explanation and/or prediction, and also with mechanisms and/or manipulation. In some contexts, especially in the natural sciences, what is sought is knowledge of mechanisms. In other contexts what is being sought is prediction, which may or may not lead to manipulation. Furthermore, there are contexts in which the mechanical and the manipulative notions of causation are both relevant. A case in point is offered by medicine, as we have argued elsewhere.[17]

The preceding pages call attention to the assumptions underlying a number of theories of causality, which involve different methods of causal attribution.

[16] See, for instance, the special issue on ceteris paribus laws of Erkenntnis 57 (2002), n. 3.

[17] See Campaner and Galavotti (2007). See also Galavotti (2008).

In general, one can say that the assumptions underlying the conceptual tools adopted for representation and inference – including in the first place model building techniques – constitute the grounds on which description, prediction and (causal and non-causal) explanation ultimately rest. The soundness of all such conceptual operations depends on the assumptions that are made, and these can only be justified with reference to the context.

References

Bogen J (2001) What we talk about when we talk about causality. http://philsci-archive.pitt.edu.
Bogen J (2004) Analyzing causality: the opposite of counterfactual is factual. Int Stud Philos Sci 18:3–26
Campaner R, Galavotti MC (2007) Plurality in causality. In: Machamer P, Wolters G (eds) Thinking about causes. From Greek philosophy to modern physics. University of Pittsburgh Press, Pittsburgh, pp 178–199
Cartwright N (2001) Modularity: it can – and generally does – fail. In: Galavotti MC, Suppes P, Costantini D (eds) Stochastic causality. CSLI, Stanford, pp 65–84
Cartwright N (2007) Hunting causes and using them. Cambridge University Press, Cambridge
Dawid P (2000) Causal inference without counterfactuals. J Am Stat Assoc 95:407–424
Dawid P (2007) Counterfactuals, hypotheticals and potential responses: a philosophical examination of statistical causality. In: Russo F, Williamson J (eds) Causality and probability in the sciences. College Publications, London, pp 503–532
Galavotti MC (1994) Some observations on Patrick Suppes' philosophy of science. In: Humphreys P (ed) Patrick Suppes: a mathematical philosopher, vol 3. Kluwer, Boston, pp 245–264.
Galavotti MC (2006) For an epistemology 'from within'. An introduction to Suppes' work. Epistemologia XXIX:215–224
Galavotti MC (2008) Causal pluralism and context. In: Galavotti MC, Scazzieri R, Suppes P (eds) Reasoning, rationality and probability. CSLI, Stanford, pp 233–252
Galavotti MC (forthcoming) Probabilistic causality, observation and experimentation. In: Gonzalez W (ed) Methodological perspectives on observation and experimentation in science. Netbiblo, A Coruña
Good IJ (1961–1962) A causal calculus I and II. Br J Philos Sci 11:305–318; 12:43–51; "Errata" and "Corrigenda" (1963) 13:88
Good IJ (1983) Good thinking. University of Minnesota Press, Minneapolis
Good IJ (1988) Causal tendency: a review. In: Skyrms B, Harper W (eds) Causation, chance and credence. Kluwer, Dordrecht, Boston, London, pp 73–78
Granger CWJ (1980) Testing for causality: a personal viewpoint. J Econ Dyn Control 2:329–352
Granger CWJ (2007) Causality in economics. In: Machamer P, Wolters G (eds) Thinking about causes. From Greek philosophy to modern physics. University of Pittsburgh Press, Pittsburgh, pp 284–296
Halpern J, Pearl J (2005) Causes and explanations: a structural-model approach. Part I: causes. Part II: explanations. Br J Philos Sci 56:843–887, 889–911
Henderson D (2005) Norms, invariance and explanatory relevance. Philos Soc Sci 35:324–338
Hitchcock C, Woodward J (2003) Explanatory generalizations, Part II: plumbing explanatory depth. Noûs 37:181–199
Holland P (2001) The causal interpretation of regression coefficients. In: Galavotti MC, Suppes P, Costantini D (eds) Stochastic causality. CSLI, Stanford, pp 173–187
Humphreys P (2006) Invariance, explanation, and understanding. Metascience 15:39–44
Lagnado D, Sloman S (2004) The advantage of timely intervention. J Exp Psychol Learn Mem Cogn 30:856–876

Mitchell S (2000) Dimensions of scientific laws. Philos Sci 67:242–265
Mitchell S (2002) Ceteris paribus. An inadequate representation for biological contingency. Erkenntnis 57:329–350
Pearl J (2000) Causality. Models, reasoning, and inference. Cambridge University Press, Cambridge
Psillos S (2004) A glimpse of the secret connexion: harmonizing mechanisms with counterfactuals. Perspect Sci 12:288–319
Sloman S (2005) Causal models. Oxford University Press, Oxford
Sloman S, Lagnado D (2004) Causal invariance in reasoning and learning. In: Ross B (ed) Psychology of learning and motivation, vol 44. Elsevier Science, San Diego, pp 287–325
Sober E (2006) Invariance, explanation, and understanding. Metascience 15:45–53
Suppes P (1967) What is a scientific theory? In: Morgenbesser S (ed) Philosophy of science today. Basic Books, New York, pp 55–67
Suppes P (1970) A probabilistic theory of causality. North-Holland, Amsterdam
Suppes P (1978) The plurality of science. In: Asquith PD, Hacking I (eds) PSA 1978, vol II. Philosophy of Science Association, East Lansing, pp 3–16. Also in Suppes (1993), pp 41–54
Suppes P (1984) Conflicting intuitions about causality. In: French P, Yuehling T, Wettstein H (eds) Causation and causal theories. Midwestern studies in philosophy, vol IX, pp 151–168; reprinted in Suppes (1993), pp 121–140
Suppes P (1993) Models and methods in the philosophy of science: selected essays. Kluwer, Dordrecht, Boston
Suppes P (2002) Representation and invariance of scientific structures. CSLI, Stanford
Van Fraassen B (1980) The scientific image. Clarendon Press, Oxford
Woodward J (1997) Explanation, invariance and intervention. Philos Sci 64 (Supplement. Proceedings PSA 2006):S26–S41
Woodward J (2000) Explanation and invariance in the special sciences. Br J Philos Sci 51:197–254
Woodward J (2001) Law and explanation in biology: invariance is the kind of stability that matters. Philos Sci 68:1–20
Woodward J (2002a) What is a mechanism? A counterfactual account. Philos Sci 69 (Supplement. Proceedings PSA 2000):S366–S377
Woodward J (2002b) There is no such thing as a ceteris paribus law. Erkenntnis 57:303–328
Woodward J (2003a) Making things happen. Oxford University Press, Oxford
Woodward J (2003b) Experimentation, causal inference, and instrumental realism. In: Radder H (ed) The philosophy of scientific experimentation. University of Pittsburgh Press, Pittsburgh, pp 87–118
Woodward J (2004) Counterfactuals and causal explanation. Int Stud Philos Sci 18:41–72
Woodward J (2006) Invariance, explanation, and understanding. Author's Response. Metascience 15:53–66
Woodward J (2007) Interventionist theories of causation in psychological perspective. In: Gopnik A, Schulz L (eds) Causal learning: psychology, philosophy and computation. Oxford University Press, New York, pp 19–36
Woodward J, Hitchcock C (2003) Explanatory generalizations, Part I: A counterfactual account. Noûs 37:1–24

Chapter 13
Epistemic Complexity from an Objective Bayesian Perspective

Jon Williamson

13.1 Introduction

This paper will focus on a particular kind of epistemic complexity, namely complexity of evidence. In particular we will look at the question of how complex evidence should impact on the strengths of an agent's beliefs.

It is a platitude to say that the strengths of our beliefs should depend on our available evidence, but it is notoriously hard to say exactly *how* evidence constrains appropriate degrees of belief. Bayesian epistemology begins to tackle this question, but typically considers only the simplest kinds of evidence, e.g., the case in which the evidence consists of a set of atomic propositions, or the case in which the evidence consists of a large database of good quality data. Reality, of course, is rarely if ever so simple. Evidence can be structured in a number of ways – causally, hierarchically, logically, for instance – and tends to be multifarious, a mixture of different kinds of structure from a mixture of different sources.

In this paper I will show how *objective Bayesianism* – one particular version of Bayesian epistemology – can help shed light on the precise relation between complex evidence and belief. Causal evidence will be considered in Section 13.4, hierarchically structured evidence in Section 13.5, logical structure in Section 13.6, and varied structure in Section 13.7. First, a crash-course on objective Bayesianism.

13.2 Objective Bayesian Epistemology

Some preliminaries An agent's *language* $\mathcal{L}$ is the means by which she expresses the propositions that concern her. Her *epistemic background* or *evidence* $\mathcal{E}$ is taken to consist of everything she *takes for granted* in her current operating context. This includes background knowledge, observations, theoretical assumptions and so on. (We will not assume that this evidence is in any way articulable, let alone articulable in $\mathcal{L}$.)

J. Williamson (✉)
Department of Philosophy and Centre for Reasoning, University of Kent, U.K.
e-mail: j.williamson@kent.ac.uk

A. Carsetti (ed.), *Causality, Meaningful Complexity and Embodied Cognition*,
Theory and Decision Library A 46, DOI 10.1007/978-90-481-3529-5_13,

According to objective Bayesian epistemology, the agent's beliefs should satisfy certain norms, the first of which says:

Probability The strengths of the agent's beliefs should be representable by probabilities.

Suppose, for example, that the agent's language $\mathcal{L}$ can express n different elementary (i.e., non logically complex) propositions $A_1, \ldots, A_n$. An *atomic state* ω on $\mathcal{L}$ is a sentence of the form $A_1^{j_1} \wedge \cdots \wedge A_n^{j_n}$ where $j_1, \ldots, j_n \in \{0, 1\}$, A_i^0 is $\neg A_i$ and A_i^1 is just A_i. Let Ω be the set of atomic states. Then the Probability norm says that the strengths of the agent's beliefs in the 2^n atomic states should be representable by non-zero real numbers that sum to 1; the degree to which she should believe an arbitrary proposition θ should be representable by the sum of her degrees of belief in those atomic states that logically imply θ. Thus the strengths of the agent's beliefs should be representable by a *probability function* over $\mathcal{L}$: a function P such that (i) $P(\omega) \geq 0$ for each ω, (ii) $\sum_{\omega \in \Omega} P(\omega) = 1$, and (iii) $P(\theta) = \sum_{\omega \in \Omega, \omega \models \theta} P(\omega)$.[1]

A second norm says that beliefs should fit with evidence:

Calibration The agent's degrees of belief should satisfy constraints imposed by evidence.

Evidence $\mathcal{E}$ can constrain degrees of belief in a variety of ways. If $\mathcal{E}$ implies that proposition θ is true, then the agent should fully believe θ. More generally, if $\mathcal{E}$ implies that the empirical probability function P^* on $\mathcal{L}$ lies in a non-empty set $\mathbb{P}^*$ of probability functions, then the probability function $P_{\mathcal{E}}$ that represents the degrees of belief that the agent should adopt on the basis of $\mathcal{E}$ lies in the convex hull $[\mathbb{P}^*]$ of $\mathbb{P}^*$. Other kinds of constraint imposed by $\mathcal{E}$ will be discussed in subsequent sections of this paper. Let $\mathbb{E}$ denote the set of probability functions that are compatible with the agent's evidence (e.g., $\mathbb{E} = [\mathbb{P}^*]$). Then the calibration norm says that $P_{\mathcal{E}} \in \mathbb{E}$.[2]

The third norm says that beliefs should only be as bold as evidence warrants:

Equivocation Degrees of belief should otherwise be as equivocal as possible.

Here 'be as equivocal as possible' is just 'be as close as possible to maximally equivocal'.[3] The probability function that is maximally equivocal – the *equivocator* $P_=$ on $\mathcal{L}$ – is the function that gives each atomic state the same probability, $P_=(\omega) = 1/2^n$ for all $\omega \in \Omega$. The distance from one probability function to another is measured by *cross entropy*, $d(P, Q) = \sum_\omega P(\omega) \log P(\omega)/Q(\omega)$.

[1] This norm is typically justified by an appeal to a Dutch book argument or Cox's theorem – see, e.g., Paris (1994, Chapter 3).

[2] This norm is typically justified on the grounds that degrees of belief are used to make predictions, and calibrated degrees of belief lead to optimal predictions in the long run (Howson and Urbach, 1989, §13.e). Strictly speaking $P_{\mathcal{E}}$ depends on $\mathcal{L}$ as well as $\mathcal{E}$; we will write $P_{\mathcal{E}}^{\mathcal{L}}$ where we need to emphasise this dependence, but drop reference to $\mathcal{L}$ and $\mathcal{E}$ where the context permits. Williamson (2005, Chapter 12) discusses language change in the context of objective Bayesianism.

[3] This norm may be justified on the grounds that degrees of belief are used as a basis for action, extreme degrees of belief lead to riskier actions, and one should only take on risk to the extent that evidence demands – see Williamson (2007).

Distance to the equivocator, $d(P, P_=) = \sum_\omega P(\omega) \log(2^n P(\omega))$, is minimised just when *entropy* $H(P) = -\sum_\omega P(\omega) \log P(\omega)$ is maximised. Hence our three norms give us:

Maximum Entropy Principle An agent's degrees of belief should be representable by a probability function $P_\mathcal{E}$, from all those that satisfy constraints imposed by evidence $\mathcal{E}$, that has maximum entropy: $P_\mathcal{E} \in \{P \in \mathbb{E} : P = argmax H\}$.

Note that once we have these three norms we don't need any further principle to guide the updating of degrees of belief in the light of new evidence. As the evidence $\mathcal{E}$ changes to $\mathcal{E}'$, the agent's belief function will correspondingly change from $P_\mathcal{E}$, a maximally equivocal probability function from those compatible with $\mathcal{E}$, to $P_{\mathcal{E}'}$, a maximally equivocal probability function from those compatible with $\mathcal{E}'$. On a language $\mathcal{L}$ expressing n elementary propositions, $P_\mathcal{E}$ and $P_{\mathcal{E}'}$ are selected by successive applications of the maximum entropy principle (Williamson 2008).

13.3 Objective Bayesian Nets

Objective Bayesianism tells us how we should set our degrees of belief. Of course we can only be expected to follow the norms of objective Bayesianism to the extent that we *can* follow these norms. But following these norms is non-trivial: abiding by the maximum entropy principle is at first sight computationally demanding, since the number 2^n of atomic states grows exponentially with the number n of expressible elementary propositions. Fortunately, there are computational tools that mitigate this computational challenge. The machinery of *objective Bayesian nets* allows one to compute objective Bayesian probabilities more efficiently. In this section we shall introduce the concepts of Bayesian net and objective Bayesian net.

First some notation. For $i = 1, \ldots, n$ the propositional variable A_i takes one of two possible values, *true* or *false*; let a_i or a_i^1 signify the assignment $A_i = true$ and $\bar{a}_i$ or a_i^0 signify the assignment $A_i = false$. It is taken for granted that an agent's degree of belief that a proposition is true (respectively false) is just her degree of belief in the proposition itself (respectively in its negation): $P(a_i) = P(A_i)$ and $P(\bar{a}_i) = P(\neg A_i)$.

A *Bayesian net* offers an efficient way of representing and manipulating a probability function. A Bayesian net on $A_1, \ldots, A_n$ consists of a directed acyclic graph whose nodes are $A_1, \ldots, A_n$, as in Fig. 13.1 for instance, together with the probability distribution $P(A_i|Par_i)$ of each variable conditional on its parents in the graph. An assumption called the *Markov Condition* is made; this says that each

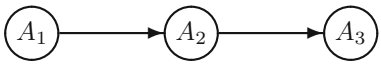


Fig. 13.1 A directed acyclic graph

variable is probabilistically independent of its non-descendants in the graph, conditional on its parents, written $A_i \perp\!\!\!\perp ND_i \mid Par_i$. Under this assumption, the Bayesian net determines a probability function over $\mathcal{L}$ via the identity $P(\omega) = P(A_1^{j_1} \wedge \cdots \wedge A_n^{j_n}) = \prod_{i=1}^{n} P(a_i^{j_i} | par^{\omega})$, where par^{ω} is the assignment to Par_i that is determined by ω, and where $j_1, \ldots, j_n \in \{0, 1\}$. Conversely, any probability function P over a finite language can be represented by a Bayesian net: simply (i) determine the independencies that are satisfied by P, (ii) represent as many of these as possible by a directed acyclic graph satisfying the Markov condition with respect to P, and (iii) add the conditional probability functions $P(A_i|Par_i)$. A wide variety of algorithms have been developed for calculating probabilities from a Bayesian net. If the graph in the Bayesian net is relatively sparse, the size of the net can increase sub-exponentially with n, meaning that it may be computationally feasible to represent and reason with a probability function even where n is very large.

An *objective Bayesian net* is just a Bayesian net that represents an objective Bayesian probability function $P_{\mathcal{E}}$, which in turn represents degrees of belief that are appropriate on the basis of evidence $\mathcal{E}$. An objective Bayesian net can be constructed by (i) determining conditional independencies that $P_{\mathcal{E}}$ must satisfy; (ii) representing these independencies by a directed acyclic graph, and (iii) maximising entropy to find the conditional probability distributions $P_{\mathcal{E}}(A_i|Par_i)$. Fortunately a maximum entropy function $P_{\mathcal{E}}$ will normally satisfy a large number of probabilistic independencies. Construct an undirected graph by linking two variables if they both occur in the same constraint imposed by $\mathcal{E}$: then $X \perp\!\!\!\perp Y \mid Z$ for $P_{\mathcal{E}}$ if in this undirected graph the variables in Z separate those in X from those in Y. Hence the graph in an objective Bayesian net will typically be sparse and it will typically be feasible to handle objective Bayesian probabilities.

For example, suppose $\mathcal{E}$ imposes the following constraints: $P(A_1|\neg A_2) \geq 0.7$, $P(A_2 \vee A_4) = P(A_3)$, $P(\neg A_5 \wedge \neg A_3) = 0$, $P(A_4) \in [0.4, 0.5]$. Then Fig. 13.2 represents the independencies of $P_{\mathcal{E}}$. This can be transformed into a directed acyclic graph Fig. 13.3 that represents the same independencies via the Markov Condition. All that remains is to determine the conditional probability distributions. See Williamson (2005, Chapter 5) for a full algorithm for constructing an objective Bayesian net.

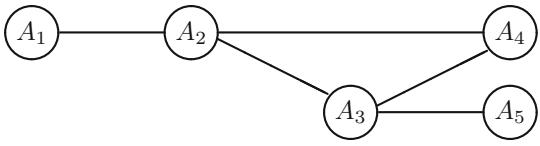


Fig. 13.2 An undirected graph representing the independencies of $P_{\mathcal{E}}$

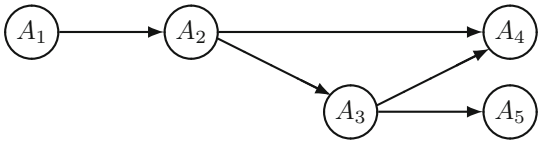


Fig. 13.3 A directed acyclic graph representing the independencies of $P_{\mathcal{E}}$ via the Markov Condition

13.4 Causal Structure

In Section 13.2 we saw that evidence of empirical probability constrains degrees of belief in a rather straightforward way: the set of probability functions compatible with evidence is just the convex hull of the set of functions in which (according to the evidence) the empirical probability function lies – written $\mathbb{E} = [\mathbb{P}^*]$. But evidence can contain information other than information about empirical probability, and the question arises as to what constraints $\mathcal{E}$ imposes on degrees of belief in such cases. In this section we shall look at the case in which evidence of causal relations is available to the agent.

Suppose for example that the agent grants the following evidence $\mathcal{E}$: smoking causes bronchitis and lung cancer, each of which cause chest pains; 30% of the population are smokers, 4% of the population but 10% of smokers get bronchitis, 2% of the population but 5% of smokers get lung cancer, 5% of the population but 99% of those with bronchitis or lung cancer have chest pains, Bob is a non-smoker with chest pains. Suppose further that $\mathcal{L}$ can express the elementary propositions S: *Bob is a smoker*, B: *Bob has bronchitis*, L: *Bob has lung cancer*, C: *Bob has chest pains*. The agent's causal evidence can be represented as in Fig. 13.4.

Causal evidence imposes constraints on degrees of belief in the following way. Causality is an *influence relation* in the sense that learning just of new non-influences provides no grounds for changing degrees of belief (Williamson 2005). More precisely, if the language $\mathcal{L}$ is extended to $\mathcal{L}'$, which expresses a new proposition, and it is known that the corresponding variable is not a cause of any of the former variables, and other information in $\mathcal{E}$ does not indicate otherwise, then the agent's degrees of belief over the former language should not change: $P_{\mathcal{E}}^{\mathcal{L}'}(\theta) = P_{\mathcal{E}_{\mathcal{L}}}^{\mathcal{L}}(\theta)$ for each sentence θ of $\mathcal{L}$, where $\mathcal{E}_{\mathcal{L}}$ is the evidence in $\mathcal{E}$ that concerns $\mathcal{L}$. Hence causal evidence imposes equality constraints on degrees of belief.

In our example $P_{\mathcal{E}}^{\mathcal{L}}(S) = P_{\mathcal{E}_{\{S\}}}^{\{S\}}(S)$ is but one such constraint. In fact these equality constraints ensure that the objective Bayesian net can be constructed by taking the causal graph Fig. 13.4 as the directed acyclic graph, and by iteratively maximising entropy to find the conditional probability distributions. See Williamson (2005, §5.8) for a detailed description of the procedure for constructing an objective Bayesian net in the presence of causal constraints. Ignoring for the moment the information that Bob is a non-smoker with chest pains, the objective Bayesian net has conditional probability distributions specified by:

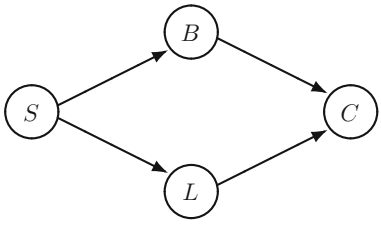


Fig. 13.4 Smoking causes Bronchitis and Lung Cancer, each of which cause Chest Pains

$$P(s) = \frac{3}{10};$$

$$P(b|s) = \frac{1}{10}, P(b|\bar{s}) = \frac{1}{70};$$

$$P(l|s) = \frac{1}{20}, P(l|\bar{s}) = \frac{1}{140};$$

$$P(c|bl) = \frac{99}{100}, P(c|\bar{b}l) = \frac{99}{100}, P(c|b\bar{l}) = \frac{99}{100}, P(c|\bar{b}\bar{l}) = \frac{40491}{659100}.$$

Now if we take the information specific to Bob into account by instantiating S to $\bar{s}$ and C to c in the network we get $P_{\mathcal{E}}(b) = P(b|c\bar{s}) \simeq 0.65$ and $P_{\mathcal{E}}(l) = P(l|c\bar{s}) \simeq 0.33$.

13.5 Hierarchical Structure

Causal structure provides one kind of evidential complexity, but there are others. In this section we shall look at evidence of hierarchical structure. Hierarchical structure occurs in descriptions of mechanisms. For instance, in describing mechanisms in the human body we often need to talk simultaneously about processes that occur at the level of the body as a whole (e.g., the circulation of the blood), those at the level of the cell (e.g., oxygenation of haemoglobin), and those at the level of the genome (e.g., mutation of a single nucleotide of the β-globin gene). Hierarchical structure also occurs in describing causal relationships, because causal relations can themselves act as causes and effects. For example, *smoking causing cancer* causes governments to restrict tobacco advertising, which prevents smoking and thereby prevents cancer (Fig. 13.5). This example shows that the same variable can occur at more than one level in the hierarchy.

Consider a simple example of hierarchically structured evidence. The National Farmer's Union needs to decide whether to lobby government for more subsidies. The evidence globally is that lobbying L is a cause of national agricultural policy A (Fig. 13.6). Here A is a hierarchical variable: one assignment a corresponds to the case in which farming F causes subsidy S (Fig. 13.7); a second assignment $\bar{a}$ is the case in which there is no link between farming and subsidy (Fig. 13.8). The evidence is that if farming causes subsidy, $P^*(s|f) = 1$, but 5% of the population receive

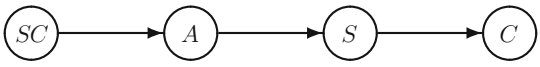


Fig. 13.5 SC: smoking causes cancer; A: tobacco advertising; S: smoking; C: cancer

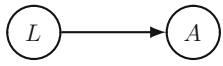


Fig. 13.6 Lobbying L causes Agricultural Policy A

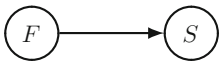


Fig. 13.7 a: Farming F causes Subsidy S

Fig. 13.8 $\bar{a}$: Farming F unrelated to Subsidy S

subsidies in any case, since fishing and other industries are subsidised. 10% of the population are farmers, and lobbying raises the probability of getting policy a by 20%.

In order to make sense of such an example we need to be clear about how evidence of hierarchical structure constrains degrees of belief. Call variable A *superior* to variable B if A occurs at a higher level in the hierarchy to B. Plausibly, *superiority* is an influence relation: learning of a new variable that is not superior to any of the current variables (and is not an influence in another respect – e.g., a causal influence) provides no grounds for changing degrees of belief concerning the current variables. So if the agent's language changes from $\mathcal{L}$ to $\mathcal{L}'$ and it is known that new propositions are not hierarchically superior to the old, then the agent's degrees of belief over the old language should not change: $P_{\mathcal{E}}^{\mathcal{L}'}(\theta) = P_{\mathcal{E}_{\mathcal{L}}}^{\mathcal{L}}(\theta)$ for each sentence θ of $\mathcal{L}$, where $\mathcal{E}_{\mathcal{L}}$ is the evidence in $\mathcal{E}$ that concerns $\mathcal{L}$. Hence hierarchical evidence imposes equality constraints on degrees of belief in the same way that causal evidence imposes such constraints.

Our example contains a mixture of causal and hierarchical evidence, but since both are evidence of influence relations, both can be treated alike. In this case the objective Bayesian net is a hierarchical or *recursive* Bayesian net (Williamson 2005). At the higher level is a network based on Fig. 13.6 – here the conditional probabilities are:

$$P(l) = \frac{1}{2};$$

$$P(a|l) = \frac{3}{5}, P(a|\bar{l}) = \frac{2}{5}.$$

At the lower level, the network for a, based on Fig. 13.7, has probabilities

$$P_a(f) = \frac{1}{10};$$

$$P_a(s|f) = 1, P_a(s|\bar{f}) = \frac{1}{20}.$$

The network for $\bar{a}$, based on Fig. 13.8, has probabilities

$$P_{\bar{a}}(f) = \frac{1}{10};$$

$$P_{\bar{a}}(s) = \frac{1}{20}.$$

Given this hierarchical net, the probability of a farmer receiving subsidy after lobbying, $P(s|lf)$, is 0.62, while with no lobbying $P(s|\bar{l}\,f) = 0.46$. These probabilities can be helpful for calculating the change lobbying will make to the expected subsidy, and thus helpful for the decision facing the National Farmer's Union.

13.6 Logical Structure

In the case of logical structure, we shall consider two kinds of evidential complexity. The first kind involves evidence of logical implications. The second kind involves evidence concerning the probabilities of logically complex propositions.

Logical Influence

The first kind – evidence of logical implications – proceeds analogously to the cases of causal structure (Section 13.4) and hierarchical structure (Section 13.5). To take a rather elementary example, suppose the agent's evidence includes the knowledge that $\theta \rightarrow \varphi$ and that θ logically implies φ. As well as this logical structure, the agent knows that if Socrates was a man then he was mortal ($A \rightarrow B$), and that it is as at least twice as likely as not that Socrates existed and was a man (A).

Now logical connection is an influence relation: a new proposition that does not, together with some current propositions, logically imply a current proposition, provides no grounds for changing degrees of belief over the current propositions. So if the agent's language changes from $\mathcal{L}$ to $\mathcal{L}'$ and it is known that new propositions are not influences (logical or otherwise) of the old, then the agent's degrees of belief over the old language should not change: $P_{\mathcal{E}}^{\mathcal{L}'}(\theta) = P_{\mathcal{E}_{\mathcal{L}}}^{\mathcal{L}}(\theta)$ for each sentence θ of $\mathcal{L}$, where $\mathcal{E}_{\mathcal{L}}$ is the evidence in $\mathcal{E}$ that concerns $\mathcal{L}$. Hence logical evidence imposes equality constraints on degrees of belief in the same way that causal or hierarchical evidence imposes such constraints.

In our example, the objective Bayesian net has the graph of Fig. 13.9. (Here C can be considered to be a hierarchical variable where assignment c corresponds to a net whose graph has nodes A and B and an arrow from A to B.) The probabilities are

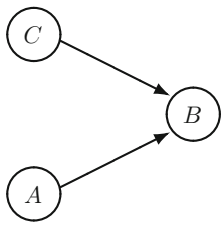


Fig. 13.9 C: if Socrates was a man then he was mortal; A: Socrates was a man; B: Socrates was mortal

$$P(c) = 1;$$

$$P(a) = \frac{2}{3};$$

$$P(b|ca) = 1, P(b|\bar{c}a) = \frac{1}{2}, P(b|c\bar{a}) = \frac{1}{2}, P(b|\bar{c}\bar{a}) = \frac{1}{2}.$$

In particular, the agent should believe that Socrates was mortal to degree $P(b) = 5/6$. See Williamson (2005, Chapter 11) for a full discussion of logical influence.

Predicate Languages

The second kind of complexity arises where the agent's evidence concerns logically complex propositions. The framework of Section 13.2 already handles arbitrary propositions in the propositional calculus. For instance, if the evidence says just that the physical probability of proposition θ is at least 0.8, $P^*(\theta) \geq 0.8$, then the agent's degrees of belief should be representable by probability function $P_{\mathcal{E}}$ which is closest to the equivocator, from all those in $\mathbb{E} = [\mathbb{P}^*] = \mathbb{P}^* = \{P : P(\theta) \geq 0.8\}$. But the question arises as to how handle evidence and beliefs concerning propositions with predicates, relations, constants, variables, quantifiers, etc. – i.e., propositions expressed in a predicate language.

If $\mathcal{L}$ is a predicate language, then the objective Bayesian method can be developed by appealing to the same three norms introduced in Section 13.2. Let $A_1, A_2, \ldots$ enumerate the *atomic propositions* of $\mathcal{L}$, i.e., the statements of the form Ut where U is a predicate or relation symbol and $t = (t_1, \ldots, t_k)$ is a tuple of constants of corresponding arity. An atomic n-state ω_n is an atomic state involving the first n of these atomic propositions: ω_n has the form $A_1^{j_1} \wedge \cdots \wedge A_n^{j_n}$ where $j_1, \ldots, j_n \in \{0, 1\}$. Let Ω_n be the set of atomic n-states.

Probability The strengths of an agent's beliefs should be representable by probabilities.

Here a probability function is a function P such that (i) $P(\omega_n) \geq 0$ for all ω_n, (ii) for each n, $\sum_{\omega_n \in \Omega_n} P(\omega_n) = 1$, (iii) for quantifier-free θ, $P(\theta) = \sum_{\omega_n \in \Omega_n, \omega_n \models \theta} P(\omega_n)$ where n is chosen large enough such that $A_1, \ldots, A_n$ includes all the atomic propositions in θ. Note that quantified sentences can be assigned probabilities as follows: $P(\exists x \theta(x)) = \lim_{m \to \infty} P(\bigvee_{i=1}^{m} \theta(t_i))$ and $P(\forall x \theta(x)) = \lim_{m \to \infty} P(\bigwedge_{i=1}^{m} \theta(t_i))$, where the $t_1, t_2, \ldots$ are the constant symbols, and where it is assumed that each element of the domain is named by precisely one constant symbol.

Calibration The agent's degrees of belief should satisfy constraints imposed by evidence.

Here, as before, the set $\mathbb{E}$ of probability functions compatible with evidence $\mathcal{E}$ is determined as follows. First take the convex closure of the set $\mathbb{P}^*$ of probability

functions in which the empirical probability function is presumed to lie. Then remove those functions which do not satisfy the equality constraints imposed by structural evidence – e.g., evidence of causal, hierarchical or logical influence considered above. Hence $\mathbb{E} = [\mathbb{P}^*] \cap \mathbb{S}$ where $\mathbb{S}$ is the set of probability functions that satisfy the structural constraints.

Equivocation Degrees of belief should otherwise be as equivocal as possible.

Again, 'be as equivocal as possible' is just 'be as close as possible to maximally equivocal'. The equivocator is defined by $P_=(\omega_n) = 1/2^n$ for all ω_n. Let $d_n(P, Q) = \sum_{\omega_n \in \Omega_n} P(\omega_n) \log P(\omega_n)/Q(\omega_n)$. Then take P to be closer to the equivocator than Q if there is some N such that for all $n \geq N$, $d_n(P, P_=) < d_n(Q, P_=)$. Thus the recipe is just as for the propositional case outlined in Section 13.2: the agent's degrees of belief should be representable by a probability function $P_\mathcal{E}$ from $\mathbb{E}$ that is closest to the equivocator.

Note that P is closer to the equivocator than Q if there is some N such that for all $n \geq N$, $H_n(P) > H_n(Q)$, where H_n is the n-entropy defined by $H_n(P) = -\sum_{\omega_n \in \Omega_n} P(\omega_n) \log P(\omega_n)$. If we deem P to have *greater entropy than* Q if this condition holds (i.e., $\exists N, \forall n \geq N, H_n(P) > H_n(Q)$), then we have a version of the maximum entropy principle for predicate languages:

Maximum Entropy Principle An agent's degrees of belief should be representable by a probability function $P_\mathcal{E}$, from all those that satisfy constraints imposed by evidence $\mathcal{E}$, that has maximum entropy in the sense outlined above.

Consider a simple example. Suppose that the agent's evidence says that *all men are mortal* has empirical probability at least $3/4$, *All those who are virtuous are men* has empirical probability at least $3/5$, and that Socrates is virtuous has probability $4/5$. The graph of the resulting objective Bayesian net is depicted in Fig. 13.10. The corresponding probabilities are

$$P(v) = \frac{4}{5};$$

$$P(h|v) = \frac{3}{4}, P(h|\bar{v}) = \frac{1}{2};$$

$$P(m|h) = \frac{5}{6}, P(m|\bar{h}) = \frac{1}{2}.$$

It turns out then that the agent should believe that Socrates is mortal to degree $P(m) = 11/15$. See Haenni et al. (2010) for more on objective Bayesianism with predicate languages, and on how to construct objective Bayesian nets in such cases.

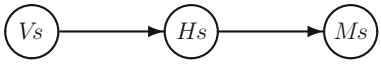


Fig. 13.10 V: virtuous; H: (hu)man; M: mortal; s: Socrates

13.7 Varied Evidence

Examples concerning the mortality of Socrates can seem remote from practical applications; in this section we shall look at a more realistic case study which exhibits a variety of kinds of evidence.

We will consider the application of objective Bayesian nets to breast cancer prognosis, described in detail in Nagl et al. (2008). The problem here is that a patient has breast cancer and an agent must make an appropriate treatment decision. Some treatments have harsh side-effects and it would not be justifiable to inflict these on low-risk patients. Broadly speaking, the higher the probability of recurrence of the cancer, the more aggressive the treatment that should be given. So it is important to determine the degree to which the agent should believe the patient's cancer will recur.

This is a genuine case of epistemic complexity in the sense that the evidence available is multifarious and exhibits various kinds of structure. Evidence includes the following. There are a variety of clinical datasets describing the clinical symptoms and disease progress of past patients. There are genomic datasets describing the presence or absence of molecular markers in past patients. There are scientific papers that provide evidence of causal relations, mechanisms, and statistical information that quantifies the strength of connection between the variables under study. Causal relationships and mechanisms can also be elicited from experts in the field, such as clinicians and researchers in cancer systems biology. And there are also a whole host of prior medical informatics systems which provide a variety of evidence: e.g., evidence of ontological relationships between variables in medical ontologies, evidence of logical relationships in medical argumentation systems.[4]

Traditional machine learning methodology would take one of two standard courses. One option is to choose the best piece of data – e.g., a clinical dataset – and to build a model – e.g., a Bayesian net – that represents the distribution of that data. The resulting model would then be used as a basis for decision. Clearly this approach ignores much of the available evidence, and will not yield useful results if the chosen data is not plentiful, accurate and relevant. A second option is to build a model from each piece of evidence and to combine the results – e.g., by each model taking a vote on the recommended decision and somehow aggregating these votes. There are several difficulties with this approach. One is that most machine learning methods only take a dataset as input; consequently the qualitative causal evidence and the evidence concerning hierarchical mechanisms is likely to be ignored. A second is that the resulting models may be based on mutually inconsistent assumptions, in which case it is not clear that they should be combined at all. A third difficulty is that the problem of aggregating the judgements of the various models is itself fraught (Williamson 2009). In contrast, the approach taken in Nagl et al. (2008)

[4] Ontological or semantic evidence may be understood in terms of influence relations, just as can causal, hierarchical and logical evidence – see Williamson (2005, §11.4).

is to construct a *single* model – an objective Bayesian net – that takes into account the full range of evidence. We considered four evidential sources, which will now be described.

The first source is the SEER study, a clinical dataset involving 3 million patients in the US from 1975–2003; of these 4731 were breast cancer patients. This dataset measures the following variables: Age, Tumour size (mm), Grade (1–3), HR Status (Oestrogen/Progesterone receptors), Lymph Node Tumours, Surgery, Radiotherapy, Survival (months), Status (alive/dead). A sample of the dataset appears below.

Age	T size	Grade	HR	LN	Surgery	Radiotherapy	Survival	Status
70–74	22	2	1	1	1	1	37	1
45–49	8	1	1	0	2	1	41	1
...	...	...	...	...	...	...	...	...

If standard machine learning methods for learning a Bayesian net that represents the empirical probability distribution of this dataset were invoked, they would generate a net with a graph similar to that of Fig. 13.11. In our case, however, we treat this empirical distribution as a constraint on appropriate degrees of belief. An agent's degree of belief in any sentence that involves only variables measured in this dataset should match the empirical probability of that sentence as determined by the dataset: $P_{\mathcal{E}}(\theta) = P^*(\theta)$ for all θ involving just variables in the dataset.

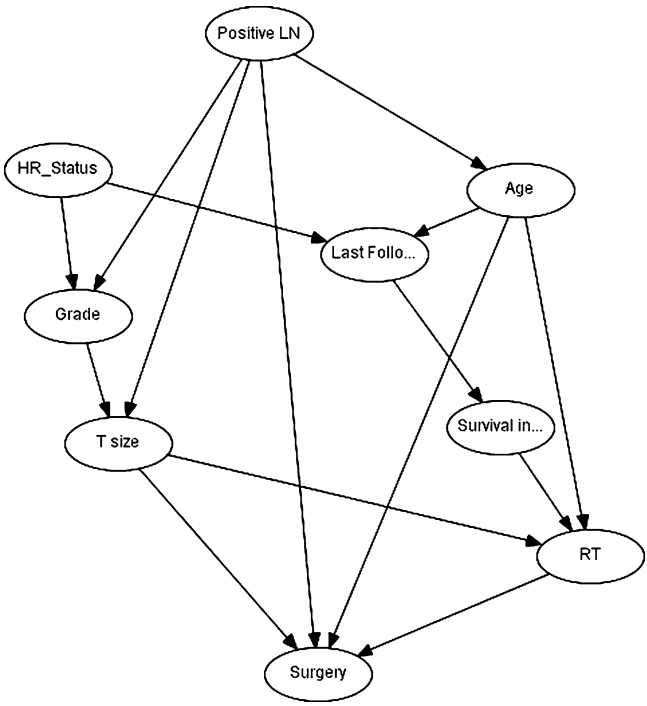


Fig. 13.11 Graph of a Bayesian net representing the empirical distribution of the clinical data

Table 13.1 Graph of a Bayesian net representing the empirical distribution of the clinical data

1p31	1p32	1p34	2q32	3q26	4q35	5q14	7p11	8q23	20p13	Xp11	Xq13
0	0	0	1	−1	0	0	1	0	0	0	−1
0	0	1	1	0	0	0	−1	−1	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...

Table 13.2 Graph in a Bayesian net representation of a genomic dataset

Lymph Nodes	1q22	1q25	1q32	1q42	7q36	8p21	8p23	8q13	8q21	8q24
0	1	1	1	1	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...

The second source consists of genomic data from a Progenetix dataset, with 502 cases. A sample appears in Table 13.1.

The empirical distribution of this dataset is represented by a Bayesian net with the graph of Fig. 13.12. Again, from an objective Bayesian point of view, this data imposes the constraint that $P_{\mathcal{E}}(\theta) = P^*(\theta)$ for all θ involving just variables in the dataset.

The third source was a further genomic dataset (119 cases with clinical annotation) from the Progenetix database (Table 13.2).

The fourth source was a paper published study (Fridlyand et al. 2006), which contains causal and quantitative information concerning the probabilistic dependence between the variables HR_status and 22q12 – this provided a further bridge between clinical and genomic variables represented in Fig. 13.13.

The resulting objective Bayesian net has the graph depicted in Fig. 13.14. This kind of representation is attractive in that it involves both clinical and molecular variables, permitting inferences from one kind of variable to the other. Thus one can use molecular as well as clinical evidence to determine an appropriate prognosis. See Nagl et al. (2008) for a fuller discussion of the construction and uses of this objective Bayesian net.

13.8 Conclusion

Complexity of evidence is one kind of epistemic complexity. In this paper we have seen how objective Bayesian epistemology can begin to tackle this kind of epistemic complexity. Objective Bayesianism offers a unifying framework for integrating and interpreting not just evidence of empirical probability, but also evidence of causal, hierarchical and logical structure. Objective Bayesian probability can be defined over predicate languages as well as propositional languages, and the machinery of objective Bayesian nets can be used to represent and reason with objective Bayesian degrees of belief.

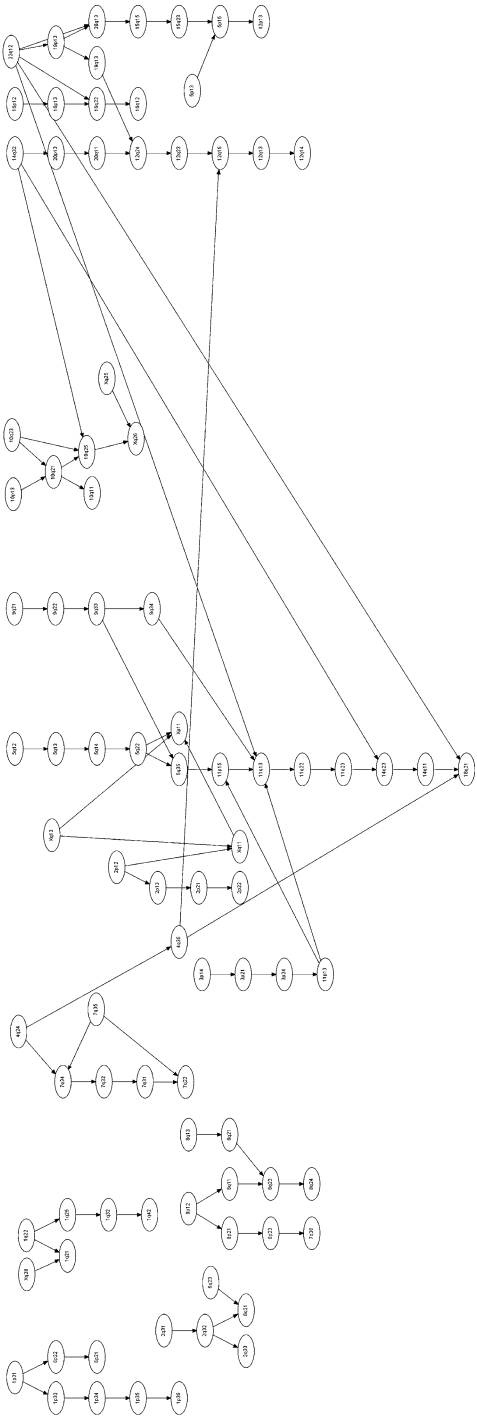


Fig. 13.12 Graph in a Bayesian net representation of a genomic dataset

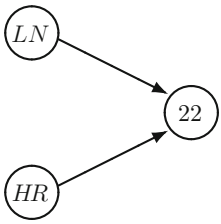


Fig. 13.13 Lymph nodes status, hormone receptor status and 22q12

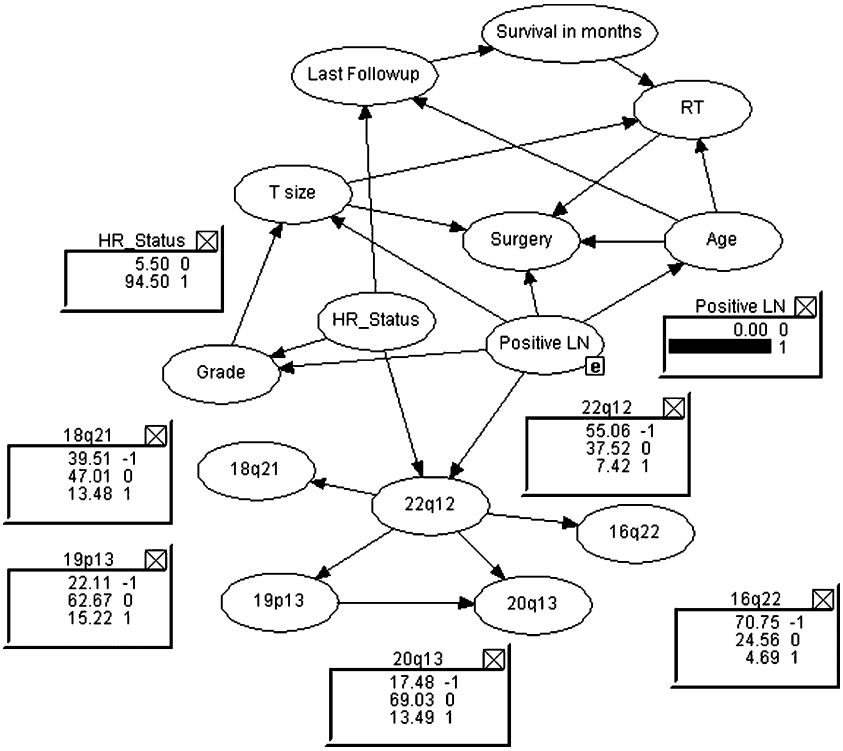


Fig. 13.14 Graph of the objective Bayesian net

Acknowledgment This research was carried out as a part of the project *progicnet: Probabilistic logic and probabilistic networks*, supported by the Leverhulme Trust.

References

Fridlyand J, Snijders A, Ylstra B, Li H, Olshen A, Segraves R, Dairkee S, Tokuyasu T, Ljung B, Jain A, McLennan J, Ziegler J, Chin K, Devries S, Feiler H, Gray J, Waldman F, Pinkel D, Albertson D (2006) Breast tumor copy number aberration phenotypes and genomic instability. BMC Cancer 6:96

Haenni R, Romeijn J-W, Wheeler G, Williamson J (2010) Probabilistic logic and probabilistic networks. Springer

Howson C, Urbach P (1989) Scientific reasoning: The Bayesian approach, 2nd edn. Open Court, Chicago, IL

Nagl S, Williams M, Williamson J (2008) Objective Bayesian nets for systems modelling and prognosis in breast cancer. In: Holmes D, Jain L, (eds) Innovations in Bayesian networks: Theory and applications. Springer, Berlin

Paris JB (1994) The uncertain reasoner's companion. Cambridge University Press, Cambridge

Williamson J (2005) Bayesian nets and causality: Philosophical and computational foundations. Oxford University Press, Oxford

Williamson J (2007) Motivating objective Bayesianism: From empirical constraints to objective probabilities. In: Harper WL, Wheeler GR (eds) Probability and inference: Essays in honour of Henry E. Kyburg Jr., College Publications, London, 151–179

Williamson J (2008) Objective Bayesianism, Bayesian conditionalisation and voluntarism Synthese DOI 10.1007/s11229-009-9515-7

Williamson J (2009) Aggregating judgements by merging evidence. J Logic Comput 19(3): 461–473

Part V
Embodied Cognition and Knowledge Construction

Chapter 14
The Role of Creativity and Randomizers in Human Cognition and Problem Solving*

Werner Leinfellner

14.1 Practical Reasoning, Default Rules, and Genetic Algorithms as New Inductive and Non-linear Evolutive Mental Processes

Genetic algorithms demonstrate that a higher organism in its environment or society can modify its behavior (humans their societal decisions) by a selective and adaptive learning process which is regimented by ad-hoc game-theoretical and statistical societal default rules. These rules may change even genetically fixed rules; their use can generate new ones which our brain evaluates (Holland's "credit assignments"; Holland 1995, p 53); the organism must store all of them in its memory system. In short, animals learn (mostly) unconsciously by using default rules (Holland 1995, p 45), humans consciously by using default rules stored in the higher linguistic and the cultural memory$_3$ system. Thus, evolutionary processing by learning, rule generation, and rules of innovations (Holland 1995, p 61) can totally describe the evolutionary and evolutive dynamic. It is characteristic for mental evolutive processing after randomizations to progress gradually by using default rules, step by step, beyond the established knowledge. The use of default rules by humans can lead, as we will show, to mental innovations and the creation of entirely new solutions of conflicts between different mentifacts, sociofacts, artifacts, and technifacts. The advances of scientific research in democratic societies are produced by inventions, teaching, transmitting, and by storing past and new solutions of societal conflicts, as well as by the ultimate successful realizations. They rest upon serial default rules stored by the gigantic, cultural, scientific, evolutive process in our cultural memory$_3$. The process is rule-bound; this is one salient property of TSO (Götschl) theories.

As already known, the rules of majority voting function like certain rules in population genetics. The democratically accepted voting rules are default rules. Voting results indicate, e.g., whether a custom as mentifact has been evaluated as

W. Leinfellner (✉)
University of Nebraska, Quaringasse 22/2/19, A-1100 Wien, Austria

A. Carsetti (ed.), *Causality, Meaningful Complexity and Embodied Cognition*,
Theory and Decision Library A 46, DOI 10.1007/978-90-481-3529-5_14,

favorable for the individuals and the society by more than half of the population. This increase in frequency can be regarded as a practical optimal solution. It can influence the course of societal evolution, at least for a while. How customs are accepted, used, and placed under democratic rules has been explained by Sen's (Nobel Prize winner of 1998 in economics) collective choice theory. It is easy to understand that individual and collective learning is an adaptive evolutionary process which uses the "spooling off" of batches of default rules, one after the other, and adds new ones according to Holland (1995, pp 11–90). Conscious learning in democratic societies as well as in cognitive processes leads to structurally the same kind of rules, evolutive Bayesian default rules, as used already by animals (Holland 1995, p 55). Since humans have learned to express default rules linguistically, the rules can be stored far more effectively, scientifically; they can be retrieved at will, used again and again, and changed, if necessary. Moreover, they are open to improvements by new empirical evidence, in accordance with Bayesian structures. In the evolution of society, therefore, conscious learning runs in the batch mode of default rules, in the same way as we follow, for example, our democratic rules. Democratic rules can be executed step by step, iterated, and improved by learning, just like Hebb's neuronal networks. When new societal conflicts arise and need to be solved and the solutions realized, we can add new default rules to the old ones; these rules work like genes steering evolutionarily the metabolism or the protein syntheses. The neuronal, cognitive, and evaluative processing and the memory storage of past successful conflict solutions work in a similar mode, too. Series of default rules (instructions) are applied sequentially; but when loops, iterations, and random events are included or certain complex evolutionary algorithms are used, the sequential order disappears. Finally, default rules have the tremendous advantage of being rules-generating rules (Holland 1995, p 50f). Any additional rule can lead to new creative solutions of societal conflicts. It is really no wonder that batches of default rules for solving societal conflicts resemble genetic algorithms (Holland 1995, p 69f). Both types of default rules form typically heuristic programs and can be improved by learning. Contrary to the genetic-biological evolution, societal evolution is regimented to an increasing part by man-made, genetically effective rules and societal default rules of which we are conscious and which are linguistically expressed.

All modern examples and models which instruct us how to solve societal conflicts and realize the solutions share, besides their being rule-bound, a common structural property with all other evolutionary biological processes; it is the statistico-causal backbone of their networks. The linking of statistico-causal pairs, CEP's (cause–effect pairs), in mente, as the units of change to empirical, causal, linear, and nonlinear complex networks, also their step-wise technical (artificial) replication and realization can be expressed by series (batches) of default rules which are stored in our memory systems for further use, as already discussed.

Default rules express linguistically the statistico-causal pairs in "if…then…" form. Such CEP's are: actions → reactions; moves → countermoves; decisions → realizations; strategies → realizations. When different CEP's are causally concatenated (see below), they form the basic sequential structure of evolutionary trajectories in causal nets, which can be described by series of default rules.

14.2 The Statistico-Causal Nature of Cognitive Thinking, the Evaluative and the Memory Function of Our Neuronal Brain and Its Protosemantics

After having outlined the evolutionary "macrosocietal" dynamic of solving societal conflicts and the realization of their solutions within democratic populations, we will return to the microsocietal and brain-physiological foundations of how the human brain may solve individual conflicts, realize solutions, and create new solutions of societal conflicts. Since two decades, since the brain-physiological turn, we know that the traditional theory of knowledge "without the human brain" is a truncated theory. There is another reason why the traditional theory of knowledge has to be rewritten: According to the author's view, the protosemantic function of the human brain, the representation of the external happenings of the world onto our brain's memory is missing in most philosophical theories of knowledge. Protosemantic meaning is a prestage of linguistic meaning; it functions very well in many "speechless" mammals. The mammalian brain has the ability to develop, store, and use either consciously or unconsciously non-linguistic meaning and preference. This is an intermediary protosemantics in the mammalian/human brain. How linguistic semantic meaning and relatively invariant reference between linguistic symbols and external objects and happenings come about is well known. But how our brain manages this without language proper is not yet known. Yet there exist protosemantic non-linguistic relations between the sensed episodes and happenings in the external world and their representations onto the internal, lowest neuronal memory system$_1$. P. Churchland rejects the traditional direct representations of the external world unto our language as a mere dogma of analytic "philosophy without brain" and calls it "sentence crunching". The protosemantics proposed here may serve as the missing cognitive link which can fill the gap between the external world and their internal representation (mapping) onto our language (for more details see Leinfellner, 1988a,b). From the societal, historical evolution of the human brain and from the most recent cognitive, brain-physiological, and linguistic research, we know that cognition, evaluations; memory storing, decision making, problem solving, and the realization of decisions and solutions of societal conflicts include a brain-based, evolutive, mental neuronal processing which involves the entire body as well (Damasio 1994, p 9f, 165ff; Basar 1988, pp 397, 491). The direct representation onto memory$_1$ presupposes a non-linguistic, brain-physiological, physical, cognitive protosemantic preferential evaluations. There is no direct representation onto linguistic memory$_2$ (Churchland 1986, p 388). The neuronal cognitive and evaluating brain starts cognition with an almost invariant, one–one, deterministic, cognitive, and evaluative, unconscious representation or mapping process (see Table 14.1) which connects external objects and causal happenings, that is, event pairs $e_{t1} \rightarrow e_{t2}$, $\text{action}_{t1} \rightarrow \text{reaction}_{t2}$, or $\text{stimulus}_{t1} \rightarrow \text{response}_{t2}$, the units of change, to their protosemantic meaning in our lowest, neuronal memory system$_1$, wave mechanically. This physical representation process initiates protosemantic meaning by fixing the invariant primitive reference relations between the dynamic "states" of the memory

Table 14.1 Protosemantic functions

$v(C_{t1} \rightarrow Et_2)$	Stored empirical protosemantic meaning plus its pragmatic evaluation (v)
↑	↑
$(e_{ti} \rightarrow e_{t2})$	Empirical objects and empirical happenings (designata, denotata)

$system_1$ and the invariantly occurring pairs, the CEP's, plus giving them always an evaluative meaning. This is the protosemantic meaning, not yet expressed linguistically. We follow the hypothesis that empirical protosemantic meaning, in the most primitive sense, begins with a one–one, relatively invariant reference relation (sensing) between physical states and the invariant wave patterns of the neuronal brain in our memory $system_1$. At the same time, the represented object, happening will be automatically evaluated with respect to us and/or to our society. Without this primitive meaning and preferences, most games and decision-theoretical models would not work and could not be regarded as dynamic statistico-stochastic reconstructions and evaluations of the intuitive methods of how the human brain solves societal conflicts and even scientific problems.

For this reason, we have to go back to the physical grass roots of the cognitive and evaluative protosemantic functions of our neuronal brain. Memory storing of happenings, of empirical causal networks begins in each case with the cognitive representation of the external, sensed, causal episodes, of the statistico-causal pairs of events, the CEP's, and their statistico-causal concatenations in our memory $system_1$. The representations of temporarily linked events, i.e., of cause → effect pairs ($e_{t1} \rightarrow e_{t2}$), of action → consequence concatenations ($a_{t1} \rightarrow c_{t2}$), etc. onto our neuronal lower-level $memory_1$ system is the first step of cognition; it is always accompanied by an individual evaluation. Both establish the primitive meaning in $memory_1$ according to Basar. These primitive, causally ordered tuples (basic causal pairs = CEP's) are represented and stored unconsciously into neuronal brain-wave patterns, they permit the recognition and afterwards the retrieval from $memory_1$ as internally evaluated episodes at our sense organs, without language. We become aware, but not fully conscious, of the neuronally stored and sensed images when the stored neuronal wave patterns, e.g., sound waves, are retrieved. Primitive retrieval from our neuronal $memory_1$ is succeeded by an "internal" excitement of our sense organs, the outposts of our brain, just like a video recorder plays back the recorded episode on its screen (for more technical details, see Leinfellner 1988a, p 349–354; Tulving 1983, p 169ff).

If these retrievals are repeated during an evolutive, internal processing, we may experience that we and the mammals have used the same protosemantical meaning successfully without using linguistic semantics. To every empirical statistico-causally linked pair of events ($e_{t1} \rightarrow e_{t2}$) = CEP corresponds invariantly (protosemantically) a neuronally represented and stored pair of evaluated episodes ($C_1 \rightarrow E_2$).

This is only a simplified, traditional sketch of how the brain-based perception by neuronal representation and storage may work. But the empirical dynamic memory

is not a static storage shelf. From the standpoint of the brain-physiological approach as represented by Basar and his group, protosemantic memory storing begins with an incoming, sensed wave pattern, e.g., of light waves emanating from a happening. What our sense organs receive are wave patterns; these wave patterns are conducted via dendrites to the brain and superimposed on the oscillating brain waves. The resulting superpositions are stored; they are dynamic "memes" or carriers of memories (Scientific American, vol. 283, n. 4, October 2000). They stay in the memory system_1 as superimpositions as long as it is active, oscillating, or alive. For example: In the case of seeing, representations begin with a mapping of empirical, relatively invariant wave patterns (our sense impressions when we see) on neuronal wave patterns; they can be made visible on screens as evoked sensory potentials, EPs, as curves with specific frequencies (Basar 1980, 1988; Leinfellner and Köhler 1998). Brain-wave models (Freeman 1988, p 30; Basar 1988, p 30; Adey 1989, p 157; Leinfellner 1988a, p 349) make important assumptions: (i) Besides the slow communication process via dendrons, dendrites, and their synapses, there exists a fast communication between the neurons of the sense organs and the brain by transneural wave communication which lasts no longer that 1/300 of a second (the P 300 wave). (ii) Perceptions, emotions, evaluations, memories, thoughts, and computations are encoded in specific patterns of brain waves. (iii) Cognitively sensed empirical episodes or happenings which recur frequently are external wave patterns (for example of light or sound waves). These are represented and encoded one-one almost invariantly onto specific brain-wave patterns by superposition on the oscillating, dynamic memory system_1 (Basar 1988, pp 45, 397); this compares to Tulving's episodic memory (Tulving 1983, pp 28, 134f). (iv) The neurons of the brain are not only interconnected in their neuronal networks by dendrons, dendrites, and their synapses, but also by internal electromagnetic waves (Basar 1989, p 47) where millions of neurons serve as emitters and others as receivers. (v) This can be observed and measured by evoked potentials (EPs), event-related potentials (ERPs), and endogenous potentials on computer screens (Basar 1988, p 30). Perceived images are sensed as incoming, for example, optically invariant wave patterns. These patterns are mapped unto the electrodynamically oscillating, neuronal carrier waves; by quantum mechanical superposition, they produce invariantly swinging internal wave patterns. Physically, they are superimposed on the oscillating neuronal waves and stored as invariantly swinging wave packets, as evoked potentials (wave patterns) which possess a specific invariant frequency. Thus we get a new physical explanation and physical underpinnings of protosemantic almost one-one relations between the invariant wave patterns_n of the memory system_1 and the empirical, incoming (sensed) wave patterns_e which have been perceived. The subscripts "e" and "n" will denote empirical and neuronal wave patterns, respectively. The representation follows the principles of the superposition of incoming (sensed) wave patterns_e on continuously oscillating neuronal carrier waves_n; this is similar to the "superhet" principles of radio receivers and transmitters. Incoming optical and auditory perceptions are physical wave patterns_e of light, of sounds, etc. with a specific frequency. They are received and represented as invariant neuronal brain-wave patterns_n and stored in this dynamic form in the neuronal memory level_1.

These internal wave patterns$_n$ can be observed via electrodes through or on the skull; their characteristic form and frequencies can be compared and measured as evoked potentials on computer screens. Even invariant thoughts and computations are encoded as dynamically swinging specific brain-wave patterns$_n$; they surpass by far the 10^{14} to 10^{15} storage capacity of our neuronal brain as it has been described in older theories. In a nutshell: When superimposed on the dynamic, perpetually oscillating biological neuronal networks, the incoming, perceived external optical wave patterns$_n$ are stored dynamically in our memory system$_1$, but not by static memes. "Dynamic" means "at least as long as the neuronal brain waves oscillate" or "as long as we live". These adaptive, cognitive, electrodynamic representations of the empirical, statistical networks$_e$ onto the lowest neuronal memory$_1$ have not been fully included into philosophical or cognitive theories of knowledge. In the threefold memory system, the cognitive, neurophysical, genetically-based representation rules regiment not only the representation unto the first memory level$_1$ but also, by special mapping rules, the immediately following representation of the first, lower, unconscious level$_1$ onto the second, higher memory level$_2$. This higher level$_2$, Tulving's semantic-linguistic level, stores the first, episodic, memory level$_1$ in the form of semantic, linguistically encoded symbols of level$_2$ of our brain-based memory$_2$ system (Tulving 1983, p 24). When the wave patterns encoded at level$_1$ are sent to, and innervate, via the Broca and Wernicke centers, the muscles of our tongue, they produce the spoken language, for example, when we describe the external world. This may explain the consciousness of the human brain by the duplication effect between the external image and the internal, linguistically described copy. Duplication means that, in our memory system, we have at our disposal the originally sensed world and the spoken or written representation of the external world simultaneously; and we can compare them to see whether they are similar or different. Consciousness appears by this kind of linguistic, representational duplication effect; this resembles the situation when we walk with open eyes through a city and compare it almost simultaneously with the map, vice versa. We have two worlds at our disposal, the empirically sensed and the stored, memorized world. Awareness, on the other hand, is a not so distinct a duplication when we see a house and compare it with the imagined house stored in our memory$_1$. We compare a real incoming picture$_e$ with an already stored, imagined picture$_n$ in our memory system$_1$, when we speak of awareness. Awareness needs no additional linguistic representation, but can be loaded (and encoded) into the symbols, words of memory$_2$. This is possible, since the time elapsed between seeing an object or an (action → reaction)-pair and the uttering of its name is so short: it takes only 1/300 of a second (Basar 1988, p 47).

But that is not all there is to it: In a similar dynamic representation process, our threefold dynamic memory system maps the linguistically encoded information from the memory system$_2$ via language automatically onto society's memory system$_3$. In the course of societal evolution, individual linguistic information simply overflows into the third collective, cultural, linguistic memory system$_3$, Popper's and Eccles' "third world" (Eccles 1989, p 16). This linguistic information overflow and storage, then, is the product of a long societal evolution; it is enforced, for

example, in present democratic societies. The capacity of the cultural memory$_3$, the written language in the works of literature and of science, which is stored in libraries and artificial networks or storage systems, is almost limitless (Eccles and Popper, 1997, p 6f; Leinfellner 1984, p 268). This is the way how information is transmitted from generation to generation by education and learning.

The cultural-scientific, brain-based memory$_3$ literally connects our brains by the external use of languages, by the exchange of spoken and written information. Today, the internet multiplies information globally and more effectively than ever.

14.3 Creativity, Lotteries, and the Combinatorial Role of Evolutionary Randomizers and Bayesian Learning

Most people believe that there are economic laws which enable us to make economic predictions, just like we make predictions in the traditional natural sciences. But the cultural, the societal, the economic, and the political evolution are not governed alone by deterministic laws, but by uncertain and risky evaluations and expectations of future gains and losses of individual and collective welfare, which are expressed by default rules. This is the reason why economists and stockbrokers abandon predictions in favor of expected, possible evaluations of what will be beneficial or detrimental for them, the individuals on the market and, at the same time, for the entire society provided it is democratic. But since risky expectations are caused by, and have to take into account, the output of randomizers, and since each output has an expected value, their expected values can be regarded as lotteries. Interestingly, most people like to win in lotteries and are familiar with playing in lotteries; to bet against random events is, in a sense, the oldest human expertise (Savage). There are far-reaching similarities between commercial and evolutionary lotteries; but there are also differences. For example, in horse lotteries we use in a Bayesian way our past experience, i.e., we can learn from our past. Whoever buys a lottery ticket and hopes to win a prize, is actually gambling under risk; without knowing it, most humans are betting specialists from birth. But a person who is given a lottery ticket is not gambling, for he risks nothing. Therefore, to partake in lotteries will become a simple empirical measure of our risk attitude when we face non-computable random events ex ante. A game of mere skill, on the other hand, is not gambling. But anybody who bets and risks something, either to win or to loose, is gambling. This is exactly the way in which we face the future course (evolutionary trajectories) of our societal evolution and the causal impact of possible random events. But there is no need to share Gould's pessimistic view of evolution.

When individuals or groups hold tickets and play in a lottery, this is, according to Gould, unrelated to the state of their bodies, to their bodily fitness, or to their knowledge. This is true of commercial lotteries only (Gould 1989, p 306). The doomsdayers and doomsayers forget that, so far, we have been unbelievably lucky gamblers in the evolutionary lottery, the reason being that we possess a memory and learn scientifically and practically from the past how to cope with random events.

Since the dawn of mankind, then, we know how to adapt ourselves to, and exploit, random events. By this cryptic verb "exploit" we mean that random events need not always lead to catastrophes. As our entire evolution proves, it depends on us, on our scientific and technical know-how, to use it for our advantage, for self-organization, and for creative outputs. This is, in a nutshell, a new interpretation of Darwin's dangerous idea. We owe our mental, our scientific, etc. creativity, acquired during the course of societal coevolution, to positively biased evolutionary randomizers. It is our scientific and technological creativity which, in democratic societies, can bias the evolutionary randomizers in our favor and has increased humanity until today. Our scientific and technical creativity has the greatest chance to protect us from future catastrophes, for example from asteroids hitting our planet. Our free will may act as a societal, egoistically biased randomizer and plague democracies in the form of dictators; but today we know better than ever how to get rid of them.

It is important and also good to know that evolutionary or temporal lotteries differ from commercial lotteries and lotteries in computers (Machina 1987). All of them possess random devices or randomizers and produce random events or random numbers, combinatorial series of digits, generated with no apparent logical order. Classical randomizers work with infinite results and combinations; but in evolutionary theory they have to be "renormalized". This resembles the renormalization in quantum physics which removes infinite nonsensical answers. Evolutionary randomizers yield, firstly, finite outcomes; secondly, they are empirically "biased", as our evolution confirms; and an increasing number of them can be biased in our favor by statistic scientific and technological know-how. They were not favorably biased for those species which are already extinct. They are biased favorable for us because, during the course of our evolution, evolutionary randomizers depended conditionally on previously favorable randomizers. Thus, series of evolutionary randomizers became biased under the impact of innumerable, forever changing partial causes which were either favorable or unfavorable for species and their societies. Likewise, internal neuronal and mental randomizers produce solutions S_i which are not equiprobably distributed, unlike random numbers or commercial randomizers. For example, in the course of our evolution, internal evolutionary and evolutive randomizers deviated from the Gaussian mean in our favor. Any evolutionary and evolutive randomizer is simply biased for a certain species by its whole evolutionary past; this begins, curiously enough, with the breaking of the primordial, equiprobable symmetry of the false vacuum by the Big Bang (Freeman 1988, p 20).

Commercial lotteries are isolated happenings whose (normal) randomizers produce always an equally probable outcome. Therefore, commercial lotteries do not evolve. But, since evolutionary lotteries come in interdependent series where one lottery depends conditionally on preceding "already played" lotteries, societal evolution can be seen as a series of interdependent advantageous lotteries. Series of lotteries, l_i, then, can form complex lotteries whose value V is $V(l) = v(\alpha_1 l_{t1}, \alpha_2 l_{t2}, \ldots, \alpha_n l_{tn})$, i.e. the value sum of the previous simple lotteries: $V(l) = v(\alpha_1 l_{t1} + \alpha_2 l_{t2} + \cdots + \alpha_n l_{tn})$, where each single lottery's value depends statistico-causally on the preceding lotteries. The evolutive processing of the solutions of consecutive societal conflict solutions $l_1, l_2, \ldots, l_n$ becomes equivalent to

a complex evolutionary lottery whose value V is $V(l) = v(\alpha_1 l_{t1}, \alpha_2 l_{t2}, \ldots, \alpha_n l_{tn})$. The value of a complex lottery is known ex post, just like the value of a single lottery. In commercial lotteries, their (normal) randomizers should follow the rules of classical probability calculus; but not the randomizers in evolutionary lotteries. Kolmogoroff's randomizers produce equiprobable outcomes, just like the roll of a single die. Evolutionary randomizers were often empirically biased in our favor; hence they deviate from classical probability calculus. They clearly violate the classical axiom of the independence of probabilities, at a more fundamental empirical level than in utility theory (Machina 1987). The Nobelist Allais' results, which became famous under the name "Allais paradox", demonstrate how our inborn risk attitude may influence our decisions and our practical solutions of societal conflicts. Both attitudes, risk loving and risk averting, "skew" the Gaussian normal distribution of expected values; but the deviations can be computed by the third movement (derivation) of the Gaussian normal distribution curve (Allais, Hagen). Risk averting "skews" the equal distribution to the left, risk loving to the right. That is exactly how evaluating humans can influence individually and collectively the course of their evolution and create biased randomizers: when they iterate their societal conflicts and successfully realize their societal solutions. Serial evolutive solutions violate the independence axiom of classical probability calculus when they maximize the individuals' and their societies' security, stability, and welfare. In this special case, our inborn risk attitudes of being either risk averting or risk loving influence the course of societal evolution. Particularly when these attitudes aggregate within populations, they can gain a tremendous impact on the future course of societal evolution. Risk loving means to gamble again and again and to try to change our future; risk averting, to be traditional, to maintain stability, but be defenseless when evolutionary randomizers prove unfavorable for us. There is only one recipe: If the future looks bright, avoid any risks; if not, take a risk to improve it. This is, at the same time, one statistico-causal way to practically influence evolutionary randomizers.

A simple example of a favorable evolutionary randomizer is a roulette table which is unevenly balanced in favor of the gamblers or the casino. Then the outcome of gambling will not be distributed equiprobably. To give an example of a simple prototype of an evolutionary randomizer: Instead of rolling just one die, we roll knowingly two dice at the same time. We know that rolling one die – the most-used example in classical probability theory – will yield only one number 2 with an equiprobable outcome of 1/6. We know that the simultaneous outcome of a pair of dice is a random number between 2 and 12. But the numbers do not have equal probabilities. The probability of rolling a 3 is twice that of rolling a 2, since a 3 can be achieved by rolling either a [1 + 2] or a [2 + 1], while a 2 can be achieved only by rolling two 1's. For a single die, all outcomes are equally possible; but it is hard to know, according to the physicist Guth, which properties of a nascent universe, if any, should be taken as analogous to the roll of a single die, or could be analogous to the traditional probability calculus (Güth 1997, p 250). Since millennia, our evolutionary roulette tables are favorably skewed, since no big cosmic or other catastrophe has disrupted our human evolution. Just like the

right-skewed distribution of risk-friendly and the left-skewed of risk-averting people deviate from the risk-neutral symmetrical distribution curve of their expected utilities (Leinfellner 1989, p 87), the different environments and the different histories produced by societal evolution are man-made and favorably "skewed" by randomizers until today. The deviations thus produced are slight but additive; they were favorable for us, our environment and our democratic societies. An example are the randomizers produced by the "green" political movements in European democracies. Our evolutionary randomizers work till today; they are far away from the symmetrical, equally probable, traditional distributions. Within the last 300 years, certain randomizers have favored democratic societies. But that is no guarantee that the randomizers will continue to be favorable; this depends, to an increasing extent, on our research in societal evolution – evolutionary lotteries can become lotteries of life and death (Gould 1989, p 306); but our scientific-technological knowledge can influence their outcomes to a far greater extent than we normally assume (Ruelle 1991, p 24).

14.4 The Heuristic Scheme of Human Evolutive Creativity as Inductive Gambling with Randomizers

Chance alone is the origin of every innovation, of all creation in the biosphere. This central concept of modern biology is no longer one among other conceivable hypotheses. It is today the sole conceivable hypothesis, the only one that squares with observed and tested facts. And nothing warrants the supposition or the hope that on this score our position is likely ever to be revised (Monod 1970, p 127).

Brain-based evolutive thinking and learning include cognition and evaluations, memory storage, and evolutive, internal randomizers (Pinker 1998, p 224f). But how does creativity function when societal conflicts have to be solved, for example by creating new culturefacts (mentifacts, sociofacts, artifacts, and technifacts)? The same holds for innovations, or partial creations and improvements, of culturefacts or methods. Here, like in all creative mental processes, mental randomizers and our simultaneous evaluations of the outcomes of mental lotteries play a leading role. They enable a new way of expected evaluations in case we don't know anything and have to search for a solution never used before, using ex ante probabilities; they also enable the realization of new solutions of social conflicts. According to Penrose (1994, pp 26, 154), Kauffman (1993, pp 174, 228), Ruelle (1991, p 5), Basar (1988, p 47), and Freeman (1988, p 28), internal neuronal randomizers are strange attractors, since they produce a vast number of expected and possible solutions, each of them with a certain value for us, in short: a lottery.

Each kind of evolution uses special evolutionarily or evolutively biased randomizers. Tunneling in physics, mutations and the genetic drift, and neuronal randomizers can initiate new products; neuronal randomizers initiate the creation of new and better adapted sociofacts, customs, and societal innovations, etc., and, by the following evolutive processing, adaptations of culturefacts to any selective

change. These mental, neuronal randomizers are strange combinatorial or chaotic attractors (Ruelle 1991, p 64; Kauffman 1993, p 178); but only they can initiate the creation of new mentifacts in a way that is similar to, but more complex than, the biological creation of species. Neuronal randomizers produce finite random combinations, a set S_i. Each single combination is submitted to our evaluation, like in a lottery. Our evaluations change the combinations to mental and empirical evolutive lotteries whose outcomes may contain new, creative, old, optimal and not optimal, fantastic, chaotic mentifacts, artifacts, sociofacts, and technifacts. When we do not know exactly how to solve a conflict, when past experience cannot help us, then it is a fact of life that we can produce optimal, creative solutions only with lotteries.

There are no counterarguments to the explanation of self-organization as an evolutionary, and creativity as an evolutive, process; they differ just as to their empirical interpretation. Internal randomizers function often within immense populations, for example neurons, as Minsky has said. Here they are seen as the primordial, initial, and blind source, possible prestages of any mental creations. To repeat: The mental creativity of our evolutive intelligence is based on evolutive neuronal randomizers which initiate rule-bound, randomlike combinations of solutions. The solutions can surpass all received traditional ones, since they may be imagined, possible and impossible, creative, etc. But since the randomizers are restricted and biased by past randomizers, the selection of the solutions is empirically bounded. In the case of conflict solving, the sets S_i contain a "rule-bound random" mixture of fantastic, prima facie causal, but also of acausal combinations of probable old and new solutions. In a next step, each set S_i is subjected to an evolutive screening processing in our memory systems in the form of a new evolutionary lottery. Evolutive screening removes at first all traditional, old, and useless solutions, then all those which do not have a statistico-causal backbone. Thus we get the set S_2 of potential or possible solutions. We then allot to each solution in S_2 a certain value; therefore, set S_2 will form a lottery, too. Our intellectual and practical efforts and our search for solutions are like tickets; we pay automatically to partake in serial mental lotteries. Since the probabilities and expected values of the winnings are known, any further evolutive processing (screening) will mean singling out the optimal solutions S_3 from the set S_2 by another evolutive lottery. Thus we regard S_3 as a new lottery, which separates from S_4 the empirically realizable solutions. We get the set S_4, again a lottery of all realizable alternatives of S_4. Since we do not know in advance which solution will win or be the creative one, we can regard, until we have "won", the series of lotteries as a compound lottery. The last "winning" set S_i may contain one or more creative solutions, a best, a second best, a third best, etc. If we have no luck and there are losses after a certain number of trials (lotteries), we may stop processing, for example when the costs are too high.

We may then start anew. Therefore, each successful evolutive, creative processing is a series of evolutive lotteries with winnings where the randomizers are favorably biased. In this case, we may end with new, creative solutions. Thus we obtain smaller and smaller sets till we end up with a set S_j, the ultimate, optimal, realizable, democratically acceptable, and creative solution(s) of a societal conflict, where $S_i > Sj$. The ultimate set S_j is the end product of the iteration of the evolutive

processing and the gambling for creative and realizable solutions of societal conflicts in the series $S_1 > S_2 > S_3 > \ldots > S_i > \ldots > S_j$ where the sign ">" means "greater in numbers". Of course, this evolutive processing and partaking in evolutionary lotteries may stop, break down, and begin again.

The expected yield of such a creative evolutive screening processing, can be formulated statistically by a version of Drake's Equation for a series of evolutive lotteries where N^j is the expected number of the creative and, at the same time, successful, optimal solutions which should increase the survival and the welfare of individuals in their democratic societies. This equation is not computable in advance, only ex post; but it may serve as a guess of the chances and risks of striving or not striving for creativity. The f's are probabilistic transition functions. The probability P that we may achieve a creative solution by evolutive processing is:

$$P = f_1 N f_2 f_3 f_4 f_5$$

This formulation comes pretty close to Boden's definition of creativity as "thinking the impossible" (Boden 1990, p 31f). For the sake of simplicity, we begin with the first step and end with the fifth function (l refers to causal solutions):

f_1 = the fraction of those solutions which are statistico-causal = l_1
N^1 = the number of solutions in the set S_1 which are statistico-causal
f_2 = the fraction of all solutions S_2 in the set S_1 which are optimal solutions = l_2
f_3 = the fraction of optimal solutions S_3 = l_3
f_4 = the fraction of optimal realizable solutions in democracies = l_4
f_5 = the number of optimal, democratically accepted solutions which conform to the Human Rights = l_5
N^j = the ultimate optimal, democratically accepted and empirically realizable creative solutions

There can be more steps than five, of course. The Drake Equation (Barrow and Tipler 1986, 586f) estimates the possible expected percentage of final creative solutions. In the case of P. Ehrlich, his chance has been incredibly high: 0.00165%. The question is nevertheless whether Ehrlich would have ever started his experiments if he would have known this probability.

We may put together all the five creative mental processes and get a compound lottery. Instead of playing each lottery l_i separately, we can play or regard all as a complex series of single lotteries: $V(l) = \alpha_1 v_1(l_1), \alpha_2 v_2(l_2), \ldots, \alpha_n v_n(l_n)$ with an ex-post probability of winning, as mentioned above, in which the gambler or researcher had luck. Any creative evolutionary and evolutive processing goes through this gambling as our examples will illustrate, with the empirical testing of realizations as the last lottery. Routine repetitions, automatic replications, serial fabrications, or commercial realizations are not evolutionary lotteries.

14.5 Examples of Creativity

An example is the creation of salvarsan by the chemist P. Ehrlich (1854–1915; Nobel Prize 1908), the founder of chemotherapy; chemotherapy has improved longevity in democracies from 58 to 78 years today. P. Ehrlich's goal was to invent a remedy against an epidemic illness, syphilis. He began with a randomly chosen, yet biased lottery l_1 of chemical compounds which would possibly kill the spirochetes, the cause of this deadly illness. This lottery of thousands upon thousands of therapeutics offered him to be possible cures. The evolutive screening began by playing a lottery l_1 between the, both possibly therapeutical, chemical, organometallic compounds and the organoarsenic compounds (Bäumler 1997, p 193 ff). The latter won because he evaluated them as having a better chance to cure the disease. In this lottery, the randomizer was already biased by previous ones, for example by Paracelsus' guesses. This lottery l_1 narrowed down the lottery l_1 to the set l_2 of organoarsenic compounds, e.g. Arsatecin (n. 306), Spirasyl (n. 418) and Atoxyl. The result of his evolutive processing and Bayesian scientific learning was that the compounds should contain inbuilt arsenic, a deadly poison; but the poison should be deadly only for the spirochetes, not for humans. Ehrlich used an internal randomizer, which he called "side-chain theory" (ibid., p 124). All possible aromatic compounds with arsenic in their side-chains should act as "magic bullets" which could be fired on the spirochetes to kill them (ibid., p 162). This anticipates today's fight of antibodies against the antigens of microbes in order to save the body's immunity and health. Too bad for Ehrlich: At his time nobody understood his magic-bullet theory; but he continued his evolutive screening, in spite of obstacles and opposition. After using intuitively the heuristic, evolutive processing by realizing (synthesizing) randomly about 300 possible and new chemotherapeutics all containing arsenic in their side-chains, he checked each one empirically on infected mice. He almost gave up when none of the first hundred trial runs cured the mice. But Ehrlich learned from the trials that a lottery l_3 of l_2 fared better than any other group: the benzene organoarsenic compounds. Now the search for a remedy really turned into a series of lotteries where the research costs were analogous to the price we pay for lottery tickets, and the probability of winning became slowly higher. Ex post, we know it: It was 0.00165%. Of course, without luck, as he often said (ibid., p 246), or biased randomizers, as we say, it could have been lower or nil. Soon he found out that a specific lottery l_3 for benzene-arseno compounds worked slightly better than all previously synthesized and tested compounds which he had randomly selected in the beginning. He continued to gamble, synthesized more than 600 chemical compounds from the lottery l_3 and he played 605 lotteries without winning! But he learned from the negative trials; in his last lottery, the 606th, he won "first prize". The 606th compound, later called "salvarsan", cured the world-wide disease (ibid., p 193). We note that the randomizers of each lottery depended on preceding lotteries, as well as on improved evolutive processing and learning. It is not necessary to emphasize the socioethical success of his creation.

Another well-known case is the chemist Kekulé (1829–1896) who pondered month after month about the problem of how six carbon atoms and six hydrogen

atoms could be concatenated to yield the desired formula of benzene, C_6H_6. His brain worked unconsciously and consciously, day and night, creating randomly possible and impossible structural formulas of benzene. One evening, when he almost fell asleep in front of his fireplace, he had a daydream: There were six mice and each mouse bit into the tail of another mouse, thereby forming a whirling ring. For Kekulé, this became a ring of six carbon atoms. Being a scientist, he immediately tested his new formula of the benzene ring. He had luck, and became the founder of organic, aromatic chemistry, the chemistry of life.

It is obvious that creativity works even on the unconscious level. O. Loewi (1873–1962), professor at the University of Graz, discovered in a dream the physiological function of the neurotransmitter acetylcholine which he called "vagusstoff". After waking up, he immediately wrote his result down on a piece of paper, but in the morning he could not decipher his writing. Luckily, the dream recurred. This time Loewi left his bed, went into his laboratory and began to test the dreamed hypothesis. For this discovery, Loewi received the Nobel Prize in 1936. In 1940, the Nazis confiscated the prize money and forced him out of Graz.

One lesson to be learned is: Whenever we gamble for creation and truth or begin to play the "creative evolutive lottery", we should never forget to store each of our past historical experiences. Contrary to Hegel's view that history teaches us that we did not learn anything from it, evolution will continue as long as we learn our lessons by evolutive screening and processing, and storing all past solutions and empirical realizations of all our problems and societal conflicts as living history.

If any last lottery and probing are negative, we can stop the creative process and begin the gambling a new with another initial evolutive randomizer. Repeating the process with a new randomizer, new evolutive processing, and playing new evolutive lotteries in mente and in practice may be the only heuristics to arrive at creative solutions – this we know today. It was Ehrlich's conviction that for being creative we need only for G's: "Glück" – luck; "Geduld" – patience; "Geschick" – know-how or skill; and "Geld" – money to support scientific research. It is four G's, since in Ehrlich's mother-tongue all four words begin with a "G" (ibid., p 246).

*Acknowledgments** I am indebted to J. Götschl for dozens of conversations at the Ludwig Boltzmann Institute for Science and Research, University of Graz. I profited from a graduate seminar on evolution at the University of Rome. I am also indebted to E. Basar and A. Carsetti. I benefited from discussions with M. Allais, J. Harsanyi, E. McClennen, M. Machina, B. Munier, A. Rapoport, R. Selten, J. Nida-Rümelin, B. Skyrms, and M. Wuketits. Last and most important: thanks to my wife Elisabeth Leinfellner for clarifying my thinking and improving the text.

References

Adey WR (1988) Electromagnetic field interactions in the brain. In: Basar 1988
Arthur WB (1999) Complexity and the economy. Science 284
Axelrod R (1984) The evolution of cooperation. Basic Books, New York
Bak P (1996) How nature works. Springer, New York
Barrow JD, Tippler F (1986) The anthropic principle. Oxford U.P., Oxford

Basar E (1980) EEG-brain dynamics. Elsevier, Amsterdam
Basar E (ed) (1988) Dynamics of sensory and cognitive processing by the brain. Springer, Berlin
Bäumler E (1997) Paul Ehrlich. Edition Votzel, Frankfurt
Boden MA (1990) The creative mind: myths andmechanisms. Basic Books, London
Bunge M (1980) The mind–body problem. Pergamon, Oxford
Churchland PS (1986) Neurophilosophy. MIT Press, Cambridge
Damasio A (1999) The feeling of what happens. Harcourt Brace, New York
Damasio A (1994) Decartes' error. Putnam's Sons New York
Eccles J (1989) Evolution of the brain: creation of the self. Routledge, London
Eigen M, Schuster P (1979) The hypercycle: a princip1e of natural self-organization. Springer, Berlin
Feigenbaum JM (1978) Quantitative universality for a class of nonlinear transformations. J Stat Phy 19
Freeman WJ (1988) A watershed in the study of nonlinear neural dynamics. In: Dynamics of sensory and cognitive processing by the brain. Springer, Berlin
Freeman WJ (1988) Nonlinear neural dynamics in olfaction as a model for cognition. In: Basar 1988
Götschl J (ed) (1993) Revolutionary changes in understanding man and society. Kluwer, Dordrecht
Götschl J (1988) Wissenschaftlicher Fortschritt und Bedingungen für Humanitätsgewinn. Zeitschrift für Wissenschafisforschung 4
Gould St. (1989) Wonderful life. Norton, New York
Güth A (1997) The inflationary universe. Addison-Wesley, Reading
Güth W (1992) Spieltheorie und ökonomische (Bei) Spiele, Berlin
Haller R, Stadler F (eds) (1988) Ernst Mach: Werk und Wirkung. Pichler Tempsky, Wien
Helbing D (1995) Quantitative Sociodynamics. Kluwer, Dordrecht
Holland JH (1992) Adaptation in natural and artificial systems. MIT Press, Cambridge
Holland JH (1995) Hidden order. Addison-Wesley, Reading, MA
Kauffman St. (1993) The origins of order. Oxford U.P., New York.
Kratky K (ed) (1989) Systemtheorie und Reduktionismus. Edition S, Wien
Leinfellner E (1992) Semantische Netze und Textzusammenhang. Lang, Frankfurt
Leinfellner E (1994a) Die Negation im monologischen Text: Textzusammenhang und Foregrounding. Folia Linguistica 25
Leinfellner E (1994b) The broader perspective of negation. J Lit Semantics 23
Leinfellner W (1984) Evolutionary causality and theory of games. In: Concepts and approaches in evolutionary epistemology. Kluwer, Dordrecht
Leinfellner W (1985) Reconstruction of Schlick's psychosociological ethics. Synthese 64
Leinfellner W (1988a) The brain-wave model as a protosemantic model. In: Basar 1988
Leinfellner W (1988b) Physiologie und Psychologie: Ernst Machs Analyse der Empfindungen. In: Ernst Mach: Werk und Wirkung. Pichler Tempsky, Wien
Leinfellner W (1989) Holismus, Reduktionismus und die Theorie dynamischer Systeme. In: Systemtheorie und Reduktionismus. Edition S, Wien
Leinfellner W (1995) Soziale Intelligenz und Rationalität. Zeitschrift für Wissenschaftsforschung 9/10
Leinfellner W (1997) Empiristische Bemerkungen zu Harsanyis Modell 'Games with Incomplete Information'. Zeitschrift für Wissenschaftsforschung 11/12
Leinfellner W, Köhler E (eds) (1998) Game theory, experience, rationality. Kluwer, Dordrecht
Lumsden CH J, Wilson EO (1981) Genes, mind and culture. Harvard U.P., Cambridge
Machina MJ (1987) Expected utility analysis without the independence axiom. Econometrica 50
Monod J (1970) Le Hasard et la nécessité. Seuil, Paris
McClennen EF (1998) Rethinking rational cooperation. In: Game theory, experience, rationality. Kluwer, Dordrecht
Penrose R (1994) Shadows of the mind. Oxford U.P., Oxford
Pinker M St (1998) Wie das Denken im Kopfentsteht. Kindler, München
Popper K, Eccles JC (1997) The self and its brain. Springer, Berlin

Ruelle D (1991) Chance and Chaos. Princeton U.P., Princeton
Selten R (1988) Game theory, experience, rationality. In: Game theory, experience, rationality. Kluwer, Dordrecht
Schuster P (1983) Replicator dynamics. J Theor Biol 100
Schuster P, Hofbauer J, Sigmund K (1979) A note on evolutionary stable strategies and game dynamics. J Theor Biol 81
Smith JM (1982) Evolution and the theory of games. Cambridge U.P., Cambridge
Sigmund K (1993) Games of life. Oxford U.P., New York
Tulving E (1983) Elements of episodic memory. Oxford U.P., Oxford
Weibull JW (1995) Evolutionary game theory. MIT Press, Cambridge
Wuketits FM (ed) (1984) Concepts and approaches in evolutionary epistemology. Kluwer, Dordrecht
Wuketits EM (1993) Verdammt zur Unmoral? Piper, München
Wuketits FM (1997) Soziobiologie. Spectrum, Heidelberg

Chapter 15
The Emergence of Mind: A Dualistic Understanding

Antonella Corradini

15.1 Emergentism as Monism and Its Critics

The aim of this essay is to show that emergentism in the philosophy of mind should be understood as a dualistic position. Before exposing my thesis I would like to say something about emergentism. It is a philosophical movement that was initiated in Great Britain in the first quarter of the twentieth century by thinkers such as S. Alexander (1920), C. Lloyd Morgan (1923), C.D. Broad (1925) and others. From a methodological viewpoint, emergentism strives to safeguard the autonomy of the so-called special sciences. It also supports an image of reality as structured into hierarchical levels of increasing complexity. According to British Emergentism, there are properties of complex systems, the *emergent* ones, that cannot be reduced to those of less complex systems. The concept of irreducibility can be traced back at the ontological level by and large to the concept of *non-deducibility*. By saying that a property of an emergent system, for example liquidity, is non-deducible, we mean that the belonging of that property to the emergent system cannot be logically deduced from the laws governing lower-level components, that is to say the atomic micro-structure. This implies that the theory which describes the properties at the lower-lever is *incomplete* as regards the properties occurring at the higher-level.

In spite of the British Emergentists' commitment to non-reductivism they have all been in favour of ontological monism. This allows us to better understand why in the present-day debate emergentism in philosophy of mind has often been assimilated to non-reductive physicalism. Both positions are supposed to have in common a commitment to a monistic materialistic ontology, though combined with the claim that higher-level properties, such as the psychological ones, are not reducible to the physical basic properties. Jaegwon Kim, one of the most resolute advocates of the similarities between emergentism and non-reductive physicalism, goes so far as to declare the latter as a form of emergentism, and to see in the recent success of non-reductive physicalism a renewal of the emergentistic atmosphere of the 20s

A. Corradini (✉)
Catholic University of Milan Largo A. Gemelli, 1 20123 Milan, Italy
e-mail: antonella.corradini@unicatt.it

A. Carsetti (ed.), *Causality, Meaningful Complexity and Embodied Cognition*,
Theory and Decision Library A 46, DOI 10.1007/978-90-481-3529-5_15,

and 30s of past century (Kim 1992, p. 121). As a matter of fact, Kim's fervour in assimilating the two views mainly aims at making a trenchant criticism of both of them. Let us thus turn to a short analysis of Kim's argument against emergentism and non-reductive physicalism.

Besides the just mentioned principles of physical monism and of the irreducibility of higher-level properties, emergentism and non-reductive physicalism would share on Kim's construal two further principles, that is to say the "Physical Realization Thesis" and "Mental Realism". The Physical Realization Thesis says that "all mental properties are physically realized; that is, whenever an organism or system instantiates a mental property M, it has some physical property P such that P realizes M in organisms of its kind" (1993, p. 344). As regards Mental Realism it corresponds to the thesis that "mental properties are real properties of objects and events ... not fictitious manners of speech" (1993, p. 344). The main consequence of the reality of mental properties is, according to "Alexander's dictum", that they have their own causal powers. This idea fits perfectly in the emergentistic frame. Emergentists, in fact, typically maintain that each emergent level of reality is endowed with specific causal powers that can be exerted at the same level of complexity, but also from the higher levels towards the lower ones (for this reason the epistemologist David Campbell later dubbed this form of causation "downward causation").

Kim's claim, however, is that both emergentism and non-reductive physicalism are committed to downward causation, as this is entailed by the basic tenets of both views (1993, p. 350). According to the Causal Realization Principle, "if a given instance of S occurs by being realized by Q, then any cause of this instance of S must be a cause of this instance of Q (and of course any cause of this instance of Q is a cause of this instance of S)" (1993, p. 352). This principle implies that same-level causation is possible only if a causal action is exerted upon the physical realization basis of the property to be instantiated and this sort of causation is downward causation. However, does this combination of downward causation with "upward determination" make downward causation plausible? Or, alternatively, does downward causation make sense within the conceptual frame of physicalism? In his influential 1999 essay on emergence, Kim employs powerful argumentative tools to give a negative answer, that downward causation does not make sense in a physicalistic context. Kim's criticism addresses first downward causation in its reflexive synchronic variety. Kim wants to show that this position leads to causal circularity. I quote: "... how is it possible for the whole to causally affect its constituent parts on which its very existence and nature depend? If causation or determination is transitive, doesn't this ultimately imply a kind of self-causation, or self-determination – an apparent absurdity?" (1999, p. 28).

The second variant of downward causation, diachronic reflexive downward causation, which is scrutinized by Kim, is still a form of reflexive causation, since the emergent property causally influences the underlying microstructure. However, such a causation does not display the antinomic and circular character of synchronic causation. In fact, the emergent property M causes at t the whole W's acquiring the new property Q at t + Dt, but W's having Q is not part of W's microstructure at t. Downward causation in its diachronic version is thus a coherent notion. Still, in the

light of Kim's conception of emergence, it is void of significance. Kim shows that the causal activity of the emergent mental property M in producing the physical property P_- is redundant, as P_- can be simply caused by P, the subvenient physical basis of M. To argue in favour of the independent causal role of M we must resort to a further positive argument, which in Kim's view has not yet been provided.

Kim's verdict, therefore, is that downward causation, even in its most plausible version, is incompatible with physicalism. It can still have a place in science and in philosophy, provided we are ready to give it up as an ontological category and to consider it as a way of describing the world, which, yet, is a purely physical world (Kim 1993, 1999).

Neither emergentism nor non-reductive physicalism can accept these conclusions, which undermine the plausibility of both positions. But are these conclusions unavoidable? They are so only if the mental level is *determined* by the physical one, that is to say if the underlying basis is not only a necessary but also a *sufficient* condition of the emergent property. However, the thesis of upward determination is not a part of any scientific discipline, but is a mere assumption, which turns out not to be true if the emergent quality is such in virtue of its *not being wholly dependent* on its realization basis. Indeed, it is plausible to maintain that higher-level mental functions, even if they presuppose the activation of the neuro-physiological level –, for there is no thought without brain –, cannot be produced by their neuro-physiological basis alone. If this is the case, then higher-level mental functions are able to influence the brain's activity, that is to say, they are sufficient conditions for it (according to downward causation), while their neuro-physiological basis is insufficient for generating higher-level mental functions, although it represents the necessary condition of them.

Against this view of the micro–macro relationship can be objected that it leans towards dualism, whereas both emergentists and non-reductive physicalists are as reluctant to embrace dualism as they are to endorse reductive physicalism. From the historical point of view it is surely right to say that emergentists were no dualists. However, I shall try to show that the emergentistic view, though not non-reductive physicalism, finds its most natural collocation in a dualistic framework.

In non-reductive physicalism downward causation is a derivative concept. In fact, Kim obtains it by showing that it is implied by same-level causation, which, in its turn, is implied by upward causation. But the implication from same-level causation to downward causation holds only under the condition that upward determination holds. Therefore, non-reductive physicalism is committed to downward causation insofar as it is a form of physicalism. Kim applies the same scheme to emergentism, but in this case his strategy is not justified, since for emergentists downward causation is not a derivative notion, but a primitive one, which lies at the very heart of their view. It is the utmost expression of the emergentistic thesis of the irreducibility of higher-level properties, thus it cannot be thought as disjointed from the non-explainability thesis. The fact that the explainability of the mental by the physical *does not hold* for emergentism undermines any project – like Kim's – to give a physicalistic interpretation of downward causation. While non-reductive physicalism is tied to upward determination and is compatible with

the explainability of the mental by the physical and with Kim's idea of downward causation, in the case of emergentism the non-explainability of the mental by the physical and the correlated concept of downward causation are hardly compatible with upward determination.

Among the many differences between emergentism and non-reductive physicalism that I cannot mention due to lack of space, I still would like to address one that pertains the structure of their respective theories. Both views are often said to imply a "property-dualism" in virtue of their claim of the irreducibility of higher-level properties. However, the sort of property-dualism implied by non-reductive physicalism is profoundly different from the emergentistic one. The former, in fact, views mental properties as situated at two different levels, the abstract and the concrete ones. At the abstract level, properties are exclusively defined by their formal role in producing the behavioural output and are not committed to any ontological position, being thus in principle also compatible with dualism. But, at the concrete level, mental properties are implemented by physical states and, as we know, they are determined by them. Mental properties, therefore, are distinct from the physical ones only at the abstract level but, once implemented, they are token-identical with them. Dualism of properties in emergentism does not involve in principle abstract mental states and considers mental properties, as far as they exist, as concrete properties token-different from their physical bases. This idea generates some tension within emergentism understood as a form of monism, but it can become wholly coherent by disavowing monism itself and by putting emergentism into a dualistic framework.

15.2 Emergentism as Dualism

In recent years some attempts have been made to develop emergentistic models which repudiate the original monistic tenets of British Emergentism and display more or less marked "dualistic" features (O'Connor 1994, 2000a,b, 2003; O'Connor and Wong 2005, 2006, O'Connor and Jacobs 2003; Humphreys 1997a,b; Hasker 1999, 2008). In this part of my essay I shall address the main theses put forward in one of these models, O'Connor's one and discuss it critically in the light of the results achieved in the previous part of the paper.

Aiming at laying out a strong ontological concept of emergence, in several essays Timothy O'Connor characterizes emergent properties as "non structural" properties. He defines structurality as follows: "A property, S, is structural if and only if proper parts of particulars having S have properties not identical with S and jointly stand in relation R, and this state of affairs is the particular's having S" (O'Connor and Wong 2005, p. 663). An emergent property is defined by contrast as the property of a composite system that is wholly nonstructural, and emergentism is defined as the view according to which there are basic, non structural properties had by composite individuals (p. 664). The view supported, at least in the 2005 essay, is property-dualism, according to which mental properties are token-distinct from the microphysical ones (p. 664).

But, how to figure out the relationship between these two different sorts of properties? O'Connor complains that the relationship is often conceived as synchronic, static and formal, due to the contemporary tendency to assimilate emergentism to non-reductive physicalism and, as a consequence, emergence to the concept of synchronical supervenience. Rather, the relationship of micro-level structures and macro-level emergent properties should be viewed as dynamic and causal. In fact, the causal action of the underlying properties is needed to explain the occurrence of emergent properties at a given level of complexity. Yet, emergent properties have causal powers which are irreducible to those of the micro-level structure and which exert at their turn an influence on lower-level and/or same-level entities (p. 665).

O'Connor's claims about the causal relationship between macro- and micro-level are in my opinion the most critical aspects of his proposal. On the one hand he defends the typical emergentistic doctrine of the existence of a downward causation. Given the non-structurality of emergent properties, "their causal influence does not occur via the activity of the micro-properties which constitute (them)"; rather they bear their influence "in a direct 'downward' fashion on the object's microstructure" (O'Connor 2003, p. 5). On the other hand, however, O'Connor also maintains that emergent properties, as everything that occurs, depend on the causal dispositions of the fundamental physical properties (p. 7). The tension existing in O'Connor's thought on this matter can be well illustrated by the passage where he examines the criticism of epiphenomenalism levelled by Kim at emergent properties (O'Connor and Wong 2005, p. 668). He emphasizes that an emergent system is not causally closed as regards its purely physical aspects and that emergent properties are thus not epiphenomenal. But, immediately after making this claim, he writes: "Consistent with this, it is true in an emergentistic scenario that everything that occurs rests on the complete dispositional profile of the physical properties prior to the onset of emergent features. For the later occurrence of any emergent properties are contained (to some probabilistic measure) within that profile, and so the effects of the emergent features are indirectly a consequence of the physical properties, too". Now, it is hard to agree with O'Connor about the consistency of downward causation with the "Causal Unity of Nature Thesis", as he names the just mentioned thesis (2003, p. 7).

A way out of this difficulty can perhaps be found in O'Connor's response to Kim's criticism of downward causation. Though conceding that "the distinctive potentialities of emergent properties do stem indirectly from the total potentialities of the basic physical properties", he adds that "… they do not determine the emergent effects (or fix the emergent probabilities) *independently* of the causal activity of those emergents" (O'Connor and Wong 2005, p. 670). What does this sentence precisely amount to? The only coherent meaning I can give to it is that the potentialities of the basic physical properties are necessary but not sufficient conditions of the causal powers of emergent properties. But this has two consequences which are not compatible with O'Connor's picture of emergence. First, the "Causal Unity of Nature Thesis" is no more valid; second, the fundamental question arises about where the special causal powers had by emergent properties stem from. As they do not entirely derive from the potentialities of the basic physical properties, they must be

rooted in a different dimension of reality. This implies that an unambiguous reading of O'Connor's previous sentence brings us to a more explicit form of dualism than that allowed by the author himself.

O'Connor also devotes one of his essays (O'Connor and Jacobs 2003) to the examination of a controversial issue, that is to say whether emergent dualism is likely to be a variant of property dualism or if it also acknowledges the emergence of whole individuals. He moves from the consideration that human beings are endowed with mental states which confer on them a unity as thinking biological substances. This functional unity of persons as wholes implies their particularity, which does not derive from the particularity of their parts, but is primitive. In the same way, the essential properties of a person are also primitive, since they cannot be reduced to those of her fundamental parts. Thus, O'Connor accepts emergent individuals, but by "individual" he means the composite system itself, with its distinctive particularity and its distinctive holistic features. He does not allow, instead, the emergence of a mental substance, whose acceptance would lead to a kind of substance dualism. O'Connor's rejection of substance dualism holds both for the case where, after having emerged, the emergent individual is ontologically independent from the physical substrate and for the alternative case where it continues to depend on it. In the first case, "a radical kind of creation ex nihilo" is required, for which there are "no remotely plausible candidate instances". As far as the second case is concerned, O'Connor objects that the natural emergence of an individual wholly distinct from the body is implausible and runs against the best empirical evidence (pp. 548–549). However, is the "emergent composite view of human persons" (p. 553) able to account for emergence understood in a strong ontological sense?

To ask this question, we must recall what we have pointed out in the previous paragraph about the origin of emergent properties and their causal powers. The problem with substance monism lies in the fact that the origin of emergent properties cannot be merely physical, because a physical structure is not sufficient for justifying the emergence of non-physical, mental properties. As I am not a Platonist, I believe – as O'Connor does – that emergent properties exist and can exert their causal powers only as instantiated properties. But, differently from O'Connor, I think that they cannot be instantiated in a mere physical substrate. Hence, they must be instantiated in a substrate which is ontologically independent from the body. It must be stressed that only an entity endowed with ontological independence is able to guarantee that its inherent forces can really exert their causal powers. If the mental substance were ontologically dependent on the body, in fact, it would have to borrow its causal powers from the body itself, so that a substance dualism with ontological dependence of the mental substance on the body would not be a much better option in this respect than a substance monism. The fact that the non-material mental substance is ontologically independent from the body, however, does not implies that it is wholly independent from it. Unlike Descartes, and similarly to Aquinas, emergent dualism *does* require a sort of dependence of the mental substance on the body, that is to say, a *functional* dependence. The mind needs an external structure, the body, in order to perform its own functions, such as perceiving, thinking, reasoning or deliberating (see on this Aquinas, *Summa Theologiae*, I, q84, a7).

By postulating a non-material, mental substance ontologically independent but functionally dependent on the body, am I subject to the first objection raised by O'Connor against substance dualism, according to which this theory requires a *creatio ex nihilo*? I take this objection to be a strong one, because, if valid, two unwelcome consequences would follow from it. First, emergent dualism would cease to be emergent, since an act of creation would render the process leading to emergence simply redundant. Emergent dualism would lose its distinctiveness from traditional dualisms "which postulate a special divine act of creation as the origin of the soul" (Hasker 2008, p. 13). But, still worse, while in traditional dualisms the notion of a *creatio ex nihilo* by God is perfectly coherent, the same does not hold for emergent dualism. What is in fact a *creatio ex nihilo*? It amounts to put into existence a particular endowed with ontological independence. But, under the supposition that an emergent individual is "an individual that *comes into existence* as the result of a certain configuration of the brain and nervous system, but which is *not composed* of the matter which makes up that physical system" (Hasker 2008, p. 13), a *creatio ex nihilo* in an emergentistic scenario is simply impossible. The emergence of a mental individual, in fact, cannot be a creation of the material basis, because empirical causes are able to modify the properties of an already existing substance, but they are *not* able to bring a new substance into existence!

Yet, however strong O'Connor's objection may be, it does not affect my own position. Substance dualism with ontological independence of the mind implies an impossible *creatio ex nihilo* only under the condition that the processes from which the mind emerges are merely material processes. Thus, this criticism can be countered if the development of the mental substance is traced back not only to material components, but also to a distinctive, non-material dimension of reality, endowed with ontological independence and existing from the very beginning of the emergent process. Such a dimension is the origin of the potentiality of development of the mental substance, which becomes actualized at the moment in which the biological structure reaches the necessary degree of complexity. Emergent dualism champions the idea of a *co-evolution* of mind and body, at the ontogenetic as well as at the phylogenetic level, on whose basis the realisation of non-biological potentialities is induced by the development of the biological structure, which, in its turn, is afterwards affected by the causal activity of the conscious mind (see on this Hasker 2008). Moreover, it is worth mentioning that the process of actualization of the mental substance also implies its particularization, its being the mind of a specific human individual. As we have just seen, the actualization of the mind is induced by a biological process of high complexity, but increasing complexity is also a sign of increasing individualization, so that my position does not face the problem of having to explain why a certain mental substance exerts its causal powers exclusively on its brain and not on somebody's else brain (this as a response to Kim 2001).

The point of view I am here sketching out could however be subject to the second objection that O'Connor makes against substance dualism, that is to say that it forces us to contemplate "a composite physical system's giving rise, all in one go, to a whole, self-contained, organized system of properties bound up with a distinct

individual". The implausible consequences of this idea, as applied to human beings, would lie in the fact that "at an early stage of physical development, a self emerges having all the capacities of an adult human self, but most of which lie dormant owing to immaturity in the physical system from which it emerges" (O'Connor and Jacobs 2003, p. 549). I confess that this objection puzzles me, in particular as regards the alleged lack of accordance of emergent substance dualism with empirical evidence. What a developmental psychologist observes concerning the developmental history of a child is the appearance at a certain stage of her development of mental capabilities, whose complexity and sophistication gradually increase, together with the concomitant maturing of the physical structure. This empirical state of affairs – it seems to me – may be interpreted equally well both by an "emergent composite view" and by an emergent dualistic view of the human being. In other words, accordance with empirical evidence is not the benchmark on whose basis a confrontation among both positions has to take place. The merits of my variant of emergent dualism are to be found first of all at the conceptual level. My proposal explains the emergence of the mental substance without resorting to any *creatio ex nihilo*, and also accounts for its ontological independence from the biological structure. In so doing, it guarantees that the mental substance has autonomous emergent powers that it can exert in a downward fashion on the body. Moreover, due to the mind's functional dependence on the body, my proposal, unlike Cartesian dualism, accounts for the existence of correlations of all mental states with brain states. As we know, neuroscientific research attests the detailed dependence of mental functions on brain functions and the existence of a systematic network of mind–brain correlations, so that at this stage of neuroscientific advancement no dualistic theory can afford to be ill at ease with such empirical data.

Other forms of emergent substance dualism meet the criterion of accounting for mind-body correlations (such as Hasker 1999). I submit that, together with these, my proposal deserves a closer look.

References

Alexander S (1920) *Space, time, and deity*, 2 vol. Macmillan, London

Beckermann A, Flohr H, Kim J (eds) (1992) *Emergence or reduction? Essays on the prospects of nonreductive physicalism.* Berlin-New York, de Gruyter

Bedau M, Humphreys P (eds) (2008) *Emergence: Contemporary readings in science and philosophy*, Cambridge, Mass: The MIT Press

Broad C.D (1925) *The mind and its place in nature*, London, Routledge and Kegan Paul

Hasker W (1999) *The emergent self*, Ithaca and London, Cornell University Press

Hasker W (2008) Emergent dualism: a mediating view of the nature of human beings, manuscript; Italian translation: "Dualismo emergente: una prospettiva di mediazione sulla natura degli esseri umani". In Lavazza A (ed.) *L'uomo a due dimensioni*. Bruno Mondadori. pp. 240–55

Humphreys P (1997a) How properties emerge, *Philosophy of Science* 64:1–17.

Humphreys P (1997b) "Emergence, not supervenience." Philosophy Sci 64 (Proceedings) pp. S337–S345

Kim J (1992) 'Downward causation' in emergentism and nonreductive physicalism. In: Beckerman

A, Flohr H, Kim J (eds) Emergence or reduction? Essays on the prospects of nonreductive physicalism. de Gruyter, Berlin, New York, pp. 119–138
Kim J (1993) The nonreductivists's troubles with mental causation.. In: Kim J (ed) Supervenience and the mind: selected philosophical essays. Cambridge University Press, Cambridge, pp. 336–357
Kim J (1999) Making sense of emergence. Philos Stud 95, pp. 3–36
Kim J (2001) Lonely souls: causality and substance dualism, pp. 30–43. In: Corcoran K (ed) Soul, body, and survival. Cornell University Press, Ithaca and London
Kim J (2006a) Being realistic about emergence. In: Clayton P, Davies P (eds) (2006) The re-emergence of emergence. Oxford University Press, Oxford
Kim J (2006b) Emergence: core ideas and issues. Synthese 151(3):347–354
Kistler M (ed) (2006) New perspectives on reduction and emergence in physics, biology and psychology, synthese. Special Issue 15/3
Lloyd Morgan C (1923) Emergent evolution. Williams & Norgate, London
McLaughlin BP (1992) The rise and fall of British emergentism, pp. 49–93. In: Beckerman A, Flohr H, Kim J (eds) Emergence or reduction? Essays on the prospects of nonreductive physicalism. de Gruyter, Berlin, New York
Mill J St (1973) A system of logic. In: Collected works of John Stuart Mill, vol VII. Toronto University Press, Toronto
O'Connor T (1994) Emergent properties. Am Philos Quart 31:91–104
O'Connor T (2000a) Persons and causes. Oxford University Press, Oxford, New York
O'Connor T (2000b) Causality, mind, and free will. Philos Perspect 14:105–117
O'Connor T (2003a) Groundwork for an emergentist account of the mental. Progress Complexity Inf Des 2(3.1):1–14
O'Connor T, Jacobs JD (2003) Emergent individuals. Philos Quart 53:540–555
O'Connor T, Wong HY (2005) The metaphysics of emergence. Nous 39:658–678
O'Connor T, Wong HY (2006) Emergent properties. In: Zalta EN (ed) Stanford encyclopedy of philosophy (October 2006 Edition)
Stephan A (1992) Emergence. A systematic view on its historical facets, pp 25–48. In: Beckerman A, Flohr H, and Kim J (eds) Emergence or reduction? Essays on the prospects of nonreductive physicalism. de Gruyter, Berlin, New York
Stephan A (2005) Emergenz. Von der Unvorhersagbarkeit zur Selbstorganisation. mentis, Paderborn

Chapter 16
Doing Metaphysics with Robots

Domenico Parisi

Science and philosophy are both rational attempts at understanding reality but they are attempts of a different nature. A crucial difference is that scientific theories are supposed to generate specific predictions that match reality as we systematically, and possibly quantitatively, observe it with our naked senses or aided by instruments, whereas philosophical theories are normally supported only by arguments and are evaluated only through analysis and discussion. Science and philosophy were born together in ancient Greece and were initially fused together. However, beginning from the seventeenth century, science has separated itself from philosophy due to the mathematical formulation of its theories and the adoption of the experimental method as a more powerful and precise way of empirically observing reality. But this has been true for the natural sciences, for physics, chemistry, and biology, not for the sciences that study human behaviour and human societies. The cognitive and social sciences deal with more complicated and elusive phenomena and neither mathematically formulated theories nor the experimental method can normally be applied to them. As a consequence, while the natural sciences are today clearly separated from philosophy and we ask scientists, not philosophers, to tell us about nature, this is not true for the sciences that study human beings. Their theories, like philosophical theories, are expressed in words, not mathematical symbols, and therefore they tend to be vague, ambiguous, and unable to generate uncontroversial empirical predictions. The experimental method is not only a very powerful empirical method but, since it can be equally used in physics, chemistry, and biology, it constitutes a powerful instrument of dialogue and unification among the natural sciences. In contrast, the cognitive and social sciences use many disparate empirical methods, and this is a serious obstacle to dialogue and integration and prevents these sciences from offering us a unified picture of human behaviour and human societies. All these factors explain why the sciences of human behaviour and human societies are much less advanced and much less able to make cumulative progress

D. Parisi (✉)
National Research Council 44, Via S. Martino della Battaglia, 00185 Rome, Italy
e-mail: domenico.parisi@istc.cnr.it

A. Carsetti (ed.), *Causality, Meaningful Complexity and Embodied Cognition*,
Theory and Decision Library A 46, DOI 10.1007/978-90-481-3529-5_16,

compared with the natural sciences, and why in these disciplines there is perpetual conflict among different "schools", like what we see in philosophy.

The advent of the computer is likely to change all this. Until now science has studied reality using two 'arms': the empirical observation of reality and the formulation of theories that try to explain what is observed. The computer makes it possible to use a third 'arm': the reproduction of reality in artefacts. The artefacts are simulations and robots or collections of robots. If an artefact behaves like some aspect or phenomenon of reality, we can claim that the principles we have followed in constructing the artefact are the same principles that govern that aspect or phenomenon of reality, and therefore we have understood that aspect or phenomenon of reality. Simulations and robots are a new way of expressing scientific theories. Traditionally, scientific theories are expressed either in words or using the symbols of mathematics. A computer simulation or the control system of a robot is a theory expressed as a computer program. This forces the researcher to formulate his or her theory in a precise and unambiguous way because, otherwise, the theory cannot be expressed in a computer program or in the control system of a functioning robot. Furthermore, theories expressed as computer programs or as the control system of a robot necessarily generate a great number of very detailed and uncontroversial empirical predictions since the results of the simulation or the behaviours of the robot are the empirical predictions of the theory incorporated in the program or in the control system of the robot.

Computer simulations are increasingly used in all scientific disciplines, especially when the phenomena of interest are too complicated to be treated with the traditional tools of mathematics, which can happen even in physics. However, computer simulations represent a decisive novelty for the cognitive and social sciences for at least three reasons: (1) they may replace the traditional verbally formulated, and therefore vague and ambiguous, theories of these disciplines; (2) they generate a great number of uncontroversial empirical predictions; (3) they constitute a single, unified, research methodology that can be equally used in all cognitive and social sciences, from psychology to anthropology, from sociology to economics, from political science to history. And, even more crucially, computer simulations can unify the study of human beings and the study of nature because, when one is trying to simulate the behaviour of an individual human being or of a collection of human beings, one can start from biology (genetics, the neurosciences) and arrive to behaviour and social phenomena, all in one and the same simulation. From this point of view robots are especially important because they make entirely clear how the physical body of human beings and the physical interactions of the body with the external environment are crucial for understanding human behaviour at both the individual and social level.

The importance of the artificialist approach for the cognitive and social sciences – understanding human beings and their societies by reproducing them in artefacts – can be summarized in the following way. The adoption of the artificialist approach can play the same role for the cognitive and social sciences that mathematical theories and the experimental method have played for the natural sciences, by allowing the cognitive and social sciences to become as autonomous from philosophy as the

natural sciences and to make the same type of cumulative progress which characterizes the natural sciences. Unlike what has been so far, in the future we will ask scientists, not philosophers, to tell us about human beings in much the same we way as today we ask scientists, not philosophers, to tell us about nature.

But the relationship between science and philosophy will be changed by the adoption of the artificialist approach in other ways. Science will be involved in tasks that traditionally have been philosophical tasks. The stronger cognitive and social sciences that will emerge with the adoption of simulations and robotics as research tools will try to ask questions that traditionally have been philosophical questions: What is art? What is religion? What is mathematics? What is science? Even: What is philosophy? But these questions will not be formulated as "what" questions. They will be formulated as "How" questions. Not "What is art, religion, mathematics, science, philosophy?", but "How can we construct an artificial system which exhibits art, religion, mathematics, science, philosophy?"

The new cognitive and social sciences will also ask the most fundamental and classical of all philosophical questions: What is reality? In other words, they will be involved in doing metaphysics, where metaphysics is the attempt to describe reality in its most fundamental and general aspects. (Following Strawson (1959), this can be called "descriptive metaphysics" and contrasted with "revisionary metaphysics", which proposes to change our view of reality to make it better. Strawson identifies Aristotle and Kant as descriptive metaphysicians and Plato and Berkeley as revisionary metaphysicians.) But the new artificialist cognitive and social sciences will do metaphysics in a different way compared to philosophers. Philosophers do metaphysics through conceptual analysis, reasoning, imagination, the proposition of ideas and theories, and discussion with colleagues. Their work, as always in philosophy, takes place entirely through the medium of language: all they do is speak and listen, write and read. The new cognitive and social scientists will do metaphysics in a different way: by constructing robots. The metaphysics described by them will be the metaphysics of the robots that they will construct, reality as the robots know and understand it. Robots are physical artefacts, whether they are simulated in a computer or physically realized, and this is very important because the knowledge that any organism has of reality depends on the organism's body, its external morphology of size and shape and its internal structure of organs and systems, and on the nature of its sensory and motor organs. A robot is a simulation of the body of an organism and of its sensory and motor organs. By constructing robots, and by comparing robots with different bodies and different sensory and motor organs, one can do "comparative metaphysics", trying to identify what general view of reality develops in each type of robot and comparing these different views. What we should do is construct a robot which is like us, with our body, our sensory and motor organs, our type of brain, our type of communication system (language), our social life, and then construct a robot which is unlike us, with a different body (much bigger or much smaller, with a different shape, etc.), different sensory and motor systems, a different type of brain, a different type of social life, or no social life at all. If, after we have constructed robots of both types, we can tell what is the general view of reality of the two types of robots and we can describe how the two views of reality

are different, this can help us understand what is our general view of reality, our "metaphysics". Or, since simulations and robots can be used to study not only real "reality" but also possible "reality", we can construct robots that do not resemble any animal that actually exists or has existed on Earth, or robots that live in an artificial environment which in different from the environment which exists on Earth, and determine what is their general view of reality. In other words, we can do not only "comparative metaphysics" but also "experimental metaphysics", determining how the metaphysics of an organism changes as we manipulate the organism's various properties.

In his book titled *Individuals. An Essay in Descriptive Metaphysics* (Strawson 1959), has proposed that one way of doing (philosophical) metaphysics is to work by contrast, by imagining alternative views of reality possessed by organisms with alternative sensory systems compared with ours. Strawson tries to describe what our general conception of reality would be if, instead of having all the senses that we do have, we had only the sense of hearing and, therefore, we would live in a world made only of sounds. What would be our general view of reality, and how would it be different from the view of reality that we actually have because, in addition to hearing sounds, we see things, we feel them when we make contact with them, we have taste, smell, and the proprioceptive sense that tells us where the different parts of our body are at any particular time?

Doing metaphysics with robots has a number of advantages with respect to doing metaphysics as Strawson and other philosophers do it, that is, using analysis, imagination, and language. What are the advantages?

If one does metaphysics simply by imagining alternative worlds, for example human beings who have only the sense of hearing but not all the other senses, there is always the risk of ignoring aspects of humans beings that are important if we want to reconstruct the general properties of the human conception of reality. For example, when he imagines his creatures living in a world made only of sounds, Strawson implicitly assumes that these creatures speak and think like normal human beings. But can one speak and think like a normal human being in a world which is made only of sounds? Is it not also necessary to see and, even more importantly, to touch and manipulate things so that these things can have a certain stability and permanence and we can refer to them with words? In a world of only sounds, would a distinction between nouns and verbs, which is so central in human language, make sense? These are questions which do not have an easy answer but an advantage of doing metaphysics by constructing robots is that it becomes impossible to take anything for granted since everything in our robots, every characteristics possessed by them or existing in their environment, must be carefully introduced by the researcher in the simulation in order to be able to observe its consequences. Doing metaphysics using robots forces me to make all my assumptions explicit, and tells me, in a way which is noncontrovertible because it is mechanical, which are, and which are not, the consequences of my assumptions.

That the pure reasoning and imagining of philosophers have limitations that constructing robots does not have is demonstrated by another example of philosophical "imagination", an example better known than Strawson's creatures that live in a

world of only sounds. In his *Symposium* Plato tells us that, in the beginning, every human being was made up of two individuals physically united by their shoulders. This being made up of two individuals had four legs, fours arms, and two faces looking in opposite directions. And it was androgine, half male and half female. Then Zeus, to punish them for their desire to escalate the Olympus and become similar to the gods, divided each of these imaginary beings into two individuals, by separating them by their shoulders and by making each individual either male or female. This philosophical imagination is used by Plato to explain love, that is, what pushes every human being to fuse with another human being, thereby recreating the primordial unity. But Plato's androgine does not stand on its feet (almost literally), that is, it cannot work, and this would become completely clear if one tried to construct the androgine as a robot. How could the androgine robot of Plato move in space? One of its two parts would tend to go in one direction and the other in the opposite direction, so that the androgine would be unable to move. The robotic version of Plato's androgine would teach us another interesting lesson, which would go against Plato's idea of love as a universal principle. In the same way as the two parts of the androgine would dispute between them with respect to the direction in which to move, once they have been separated by Zeus and have become two human beings as we know them, the two human beings would dispute between them on the many things that create conflict between human beings as much as love pushes them to unite. So a robotic version of Plato's androgine would give a better account of human beings as guided by both love and conflict.

A second advantage of doing metaphysics by constructing robots derives from the fact that robots do not have only sensory organs but they also have motor organs and they are physical objects with a given size and a given shape. A long philosophical tradition attributes to our senses, and especially to our visual sense, the sole responsibility for the construction of our knowledge of reality, and ignores the role played in this construction by our body and by the movements of our body. According to this tradition, knowledge is passive in that it derives from the information that our senses provide us about reality, while the movements with which our motor organs respond to sensory information and modify external reality, or at least the manner in which external reality arrives to our senses, tend to ignored. Our knowledge of reality is assumed to be independent of the characteristics of our body, of its size and shape, and of the nature and spatial disposition of our sensory and motor organs. Both assumptions are mistaken. The knowledge that organisms have of reality is more dependent on the actions with which the organisms respond to sensory information and act on reality, than on this sensory information as such. The body of the organism, with its particular physical characteristics, has a great importance for determining the type of knowledge that the organism constructs about the environment in which the organism exists and with which it interacts physically. Doing metaphysics by constructing robots is advantageous because it forces us to consider the role of the body and of the movements of the body in determining the manner in which human beings conceive of, and know, reality. By constructing robots we can more easily see how we put different objects in the same category not because they are similar from a sensory point of view but because we respond to them with the

same action(s), how knowledge of where things are in space is knowledge on how to reach things with our eyes, hands, or feet, how counting is always counting only our actions, how time is counting our actions in time, etc. Unless we recognize the crucial role of our actions, and of the body that accomplishes these actions, in the definition of reality, we will describe an imaginary or superficial metaphysics.

The third advantage of doing metaphysics with robots is that robots necessarily presuppose a biological view of knowledge and reality. This is clear if we ask what is a robot. A robot is a physical artefact which behaves like a living organism, either an animal or a human being. It is an artificial living organism. Therefore, robotics is necessarily very close to the scientific discipline which studies living organisms, biology. Biology conceives every organism as the result of an adaptive process, as possessing characteristics that have been shaped by the particular evolutionary history that the species to which the individual organism belongs has gone through. Therefore, also the knowledge and general conception of reality possessed by any given species of organisms is the result of this particular evolutionary history. Adaptation means to change in such a way that the species becomes more able to survive and reproduce in its environment, with the changes reflecting the nature of the environment as perceived and acted upon by the particular species. Therefore, knowledge is incorporating, embodying, the environment. We do not call "knowledge" all the changes in an organism's body that are the result of adaptation but only those changes that take place in a particular part of the organism's body: the organism's brain. And we call knowledge not only the changes in the brain that occur at the evolutionary scale in the population of organisms as a result of selective reproduction and genetic mutations but also the changes that occur at the scale of an individual's life as a result of the individual's experience in the individual's particular environment. In today robotics robots tend not to be 'programmed' by the researcher but to be evolved at the population level and developed at the individual level (Parisi et al. 1990; Nolfi and Floreano 2001).

When we do metaphysics what we actually do is describe the particular adaptive pattern of one particular species of organisms, Homo sapiens. Doing metaphysics by constructing robots makes this entirely clear. And this is advantageous because it introduces a useful comparative approach that considers different species of organisms and different views of reality (which explains why Strawson finds it useful to imagine a species like us but with only acoustic sensors), and because it creates a relativistic attitude towards our conception of reality. It shows us that what we call "metaphysics" is only one among many existing conceptions of reality, those possessed by other species of animals, while this is normally not recognized because the conception of reality that we try to describe when we do metaphysics is the conception of reality of the species that does the description.

The fourth and last advantage of doing metaphysics with robots is methodological. Philosophers do metaphysics as they do everything else: by using words. What is in this respect the advantage of robotics? Words have meanings which are ill-defined and vary from one person to another person. And of course it is no solution to try to make their meaning more explicit and precise by defining them in terms of other words. In contrast, robots are physical objects, and they transform our ideas

into physical objects. What they are and what they do can be observed publicly, both in the robot's external behaviours and in the robot's internal workings that determine those behaviours, and can be measured and experimentally manipulated, observing the consequences of our manipulations. A robotic metaphysics is a scientific metaphysics. It is metaphysics as done by science. (Fellous and Arbib (2005) contrast the philosophical and the robotic approach to the study of emotions in an imaginary dialogue between Bertrand Russell and Thomas Edison.)

We conclude this paper by noting that doing metaphysics with robots has some similarity to Kant's transcendental philosophy. Kant believed that if we want to understand what is reality in its most general characteristics and why reality has the general characteristics it has, what we should do is determine how reality can be known by us, what there is in us that makes it possible for us to know reality. Similarly, if we do metaphysics by constructing robots, what we do is construct a robot which is like us and then we examine what is reality for the robot we have constructed. The general nature of reality for the robot derives from the principles we have followed in constructing the robot and, if these principles are the same principles that govern us, we have discovered why reality looks to us the way it does. Like Kant, we are trying to identify the general characteristics of reality by looking at us, not at reality. The difference with respect to Kant is that we look at us by trying to reproduce us in an artefact.

This is an important difference, and not only because of the four general advantages of doing metaphysics by constructing robots that we have discussed in the paper. With respect to Kant there are two specific advantages (Ferraris 2004). Kant can be interpreted, and has been interpreted, as licensing an idealistic, subjectivist, account of reality according to which reality is all in the mind, which appears to be an unwelcome conclusion. This idealistic reading of reality is not possible if we adopt a robotic approach to metaphysics and a biological, adaptivist, view of reality. Knowledge for any given species of organisms, is the result of the evolutionary history of adaptation to the environment of that species of organisms, and this history is all but purely mental and subjective. The view of reality possessed by the organism is entirely objective in that it is the only one that allows the organism to survive and reproduce and, therefore, it is "forced" on the organism, not chosen by the organism. The other advantage of a robotic version of Kant's transcendental philosophy is that Kant tends to have a conceptualist conception of knowledge, a conception according to which reality is known by us only in so far as we map sensory "intuitions" into concepts, where concepts, in practice, are words or, more exactly, meanings of words. This intellectualistic conception of knowledge is extraneous to the robotic approach to metaphysics. Knowledge, that is, adaptation, can manifest itself in all sorts of manners, and knowledge mediated by concepts and by language is only of one of them. Robots that replicate animals (or even plants), not human beings, also have knowledge, because they also are adapted, but they do not have language. Hence, their "metaphysics" is not linguistic and conceptual. Humans have language, and this may influence *some* of their knowledge of reality, but not all their knowledge of reality. Their ancestors did not have language, and no part of the past completely disappears in biology.

References

Fellous JM, Arbib MA (2005) "Edison" and "Russell": definitions and inventions in the analysis of emotions. In: Fellous JM, Arbib MA (eds) Who needs emotions. The brain meets the robot. Oxford University Press, Oxford

Ferraris M (2004) Goodbye Kant! Milan, Bompiani

Nolfi S, Floreano D (2001) Evolutionary robotics. The biology, intelligence, and technology of self-organizing machines. MIT, Cambridge, MA

Parisi D, Cecconi F, Nolfi S (1990) Econets: neural networks that learn in an environment. Network 1:149–168

Strawson PF (1959) Individuals. Methuen, London

Chapter 17
Knowledge Construction, Non-Standard Semantics and the Genesis of the Mind's Eyes

Arturo Carsetti

From an informational point of view, life can be characterised in terms of a concrete answer to three difficult questions: "how is information generated?", "how is information transmitted?" and "how is information assimilated?". With respect to this last interrogative, we have immediately to realise that the assimilation-process of external information implies the existence of specific forms of determination at the neural level as well as the continuous development of a specific cognitive synthesis. Actually, information relative to the system stimulus is not a simple amount of neutral sense-data to be ordered, it is linked to the "unfolding" of the selective action proper to the optical sieve, it articulates through the imposition of a whole web of constraints, possibly determining alternative channels at, for example, the level of internal trajectories. Depth information grafts itself on (and is triggered by) recurrent cycles of a self-organising activity characterised by the formation and a continuous *compositio* of multi-level attractors. The possibility of the development of new systems of pattern recognition, of new modules of reading will depend on the extent to which new successful "garlands" of the functional patterns presented by the optical sieve are established at the neural level in an adequate way. The aforementioned self-organising activity thus constitutes the real support for the effective emergence of an autonomous cognitive system and its consciousness. Insofar as an "I" manages to close the "garland" successfully, in accordance with the successive identification of specific attractors and the actual intervention of meaning selection, thereby harmonising with the ongoing "multiplication" of mental processes at the visual level, it can posit itself as an adequate grid-instrument for the "vision-reflection" on behalf of the original Source of itself, for its self-generating and "reflecting" as *Natura naturata*, a Nature which the very units (monads) of multiplication will actually be able to read and see through the eyes of the mind.

If we take into consideration, for instance, visual cognition we can easily realise that vision is the end result of a construction realised in the conditions of experience. It is "direct" and organic in nature because the product of neither simple mental

A. Carsetti (✉)
University of Rome "Tor Vergata", V. Columbia n.1, 00199 Rome, Italy
e-mail: carsetti@uniroma2.it

A. Carsetti (ed.), *Causality, Meaningful Complexity and Embodied Cognition*,
Theory and Decision Library A 46, DOI 10.1007/978-90-481-3529-5_17,

associations nor reversible reasoning, but, primarily, the "harmonic" and targeted articulation of specific attractors at different embedded levels. The resulting texture is experienced at the conscious level by means of self-reflection; we actually sense that it cannot be reduced to anything else, but is primary and self-constituting. We see visual objects; they have no independent existence in themselves but cannot be broken down into elementary data. Grasping the information at the visual level means managing to hear, as it were, inner speech. It means first of all capturing and "playing" each time, in an inner generative language, through progressive assimilation, selection and real metamorphosis (albeit partially and roughly) and according to "genealogical" modules, the articulation of the complex semantic apparatus which works at the deep level and moulds and subtends, in a mediate way, the presentation of the functional patterns at the level of the optical sieve.

What must be ensured, then, is that meaning can be extended like a thread within the file, identifying a "garland"; only on the strength of this construction can an "I" posit itself together with a sieve: a sieve in particular related to the world which is becoming visible. In this sense, the world which then comes to "dance" at the level of the eyes of my mind is impregnated with meaning. The "I" which perceives it realises itself as the fixed point of the garland with respect to the "capturing" of the thread inside the file and the genealogically-modulated articulation of the file which manages to express its invariance and become "vision" (visual thinking which is also able to inspect itself), anchoring its generativity at a deep semantic dimension. The model can shape itself as such and succeed in opening the eyes of the mind in proportion to its ability to permit the categorial to anchor itself to (and be filled by) intuition (which is not, however, static, but emerges as linked to a continuous process of metamorphosis). And it is exactly in relation to the adequate constitution of the channel that a sieve can effectively articulate itself and cogently realise its selective work at the informational level. This can only happen if the two selection processes (operating respectively within an ambient "meaning" and an ambient "incompressibility") meet, and a *telos* shape itself autonomously so as to offer itself as guide and support for the task of both capturing and "ring-threading". It is the (anchoring) rhythm-scanning of the labyrinth by the thread of meaning which allows for the opening of the eyes, and it is the truth, then, which determines and possesses them. Hence the construction of an "I" as a fixed point: the "I" of those eyes (an "I" which perceives and which exists in proportion to its ability to perceive according to the truth). What they see is a generativity in action, its surfacing rhythm being dictated intuitively. What this also produces, however, is a file that is incarnated in a body that posits itself as "my" body, or more precisely, as the body of "my" mind: hence the progressive outlining of a meaning, "my" meaning which is gradually pervaded by life.

Vision as emergence aims first of all to grasp (and "play") the paths and the modalities that determine the selective action, the modalities specifically relative to the revelation of the afore-mentioned semantic apparatus at the surface level according to different and successive phases of generality. These paths and modalities thus manage to "speak" through my own fibres. It is exactly through a similar self-organising process, characterised by the presence of a double-selection mechanism, that the mind can partially manage to perceive (and assimilate) depth information in

an objective way. The extent to which the network-model succeeds, albeit partially, in encapsulating the secret cipher of this articulation through a specific chain of programs determines the model's ability to see with the eyes of the mind as well as the successive irruption of new patterns of creativity. To assimilate and see, the system must first "think" internally (at the iterative level) the secret structures of the possible, and then posit itself as a channel (through the precise indication of forms of potential coagulum) for the process of opening and anchoring of depth information. This process then works itself gradually into the system's fibres, via possible selection, in accordance with the coagulum possibilities and the meaningful connections offered successively by the system itself.

The revelation and channelling procedures thus emerge as an essential and integrant part of a larger and coupled process of self-organisation. In connection with this process we can ascertain the successive edification of an I-subject conceived as a progressively wrought work of abstraction, unification, and emergence. The fixed points which manage to articulate themselves within this channel, at the level of the trajectories of neural dynamics, represent the real bases on which the "I" can graft and progressively constitute itself. The I-subject can thus perceive to the extent in which the single visual perceptions are the end result of a coupled process which, through selection, finally leads the original Source to articulate and present itself as *true* invariance and as "harmony" within (and through) the architectures of reflection, imagination, computation and vision, at the level of the effective constitution of a body and "its" intelligence: the body of "my" mind. These perceptions are (partially) veridical, direct, and irreducible. They exist not in themselves, but, on the contrary, for the "I", but simultaneously constitute the primary departure-point for every successive form of reasoning perpetrated by the observer. As an observer I shall thus witness *Natura naturata* since I have connected functional forms at the semantic level according to a successful and coherent "score".

In accordance with these intuitions, we may tentatively consider, from the more general point of view of contemporary Self-organisation theory, the network of meaningful (and "intelligent") causal "programs" living at the level of our body as a complex one which forms, articulates, and develops, functionally, within a "coupled universe" characterised by the existence of a double selection. This network gradually posits itself as the real instrument for the actual emergence of meaning and the simultaneous, if indirect, surfacing of an "acting I": as the basic instrument, in other words, for the perception of real and meaningful processes, of "objects" possessing meaning, aims, intentions, etc.: above all, of objects possessing an inner plan and linked to the progressive expression of a specific cognitive action.

The mind considered as an "intelligent" network which develops with its meaning articulates as a growing neuronal network through which continuous restructuring processes are effected at a holistic level, thus constituting the indispensable basis of cognitive activity. The process is first of all, as stated above, one of canalisation and revelation (*in primis* according to specific reflection procedures) of precise informational (and generative) fluxes-principles. It will necessarily articulate through schemata and attractors which will stabilise within circuits and flux determinations. In this sense the mind progressively constitutes itself as a self-organising observ-

ing device in the world and of the world. When, therefore, the model-network posits itself as a 'I-representation' (when the arch of canalisation "reaches completion"), and observes the world-Nature before it, it "sees" (and computes) the world in consonance with the functional operations on which its realisation was based, i.e. according to the architecture proper to the circuits and the patterns of meaning which managed to become established. The result is Nature written in mathematical formulae: Nature read and seen *iuxta propria principia* as a great book (library) of functional and operational forms by means of symbolic characters, grammatical patterns and specific mathematical modules.

From a general point of view, at the level of the articulation of visual cognition, we are actually faced with the existence of precise forms of co-evolution. With respect to this dynamic context, we can recognise, at the level of the aforementioned process of inventive exploration, not only the presence of modules of self-reflection but also the progressive unfolding of specific fusion and integration functions. We also find that the *Sinn* that embodies in specific and articulated rational intuitions guides and shapes the paths of the exploration selectively. It appears to determine, in particular, by means of the definition of precise constraints, the choice of a number of privileged patterns of functional dependencies with respect to the entire relational growth. As a result, we are able to inspect a precise spreading of the development dimensions, a selective cancellation of relations and the rising of specific differentiation processes.

We are faced thus with a new theoretical landscape characterised by the successive unfolding (in a co-evolutive context) of specific mental processes submitted to the action of well-defined selective pressures and to a continuous emergence of depth information. This emergence, however, reveals itself as canalised by means of the action of precise constraints that represent the end product of the successive transformation of the original *Gestalten.* (Actually, the *Gestalten* can "shape" the perceptual space according to a visual order only insofar as they manage to act (on the basis of the transformation undergone) as constraints concerning the generative (and selective) processes at work.)

The *Gestalten* constitute first of all the natural forms through which meaning can be enclosed (i.e., realising its thread-like extension) and can modulate its action along the ramparts of its surface "captivity". In this sense, they determine at a primary level the gradual shaping of the structures of the "I" which cannot help but think through forms if it is to self-organise as an ongoing process of vision: if it wishes to perceive veridically, and ultimately posit itself as the fixed point for the process of vision (including, Husserl would add, the vision of the categories themselves).

This "I" as incarnated, embodied mind, gradually becoming "occupied" by meaning while it articulates as life, ultimately reveals itself as the "I" of a body ("my body"), a body that articulates as an autonomous production of forms, the achieved extension of the meaning within the file, and as the world of virtual possibility in the guise and limits of necessity. It acts as the "I" of a body-meaning which, in articulating as "my" body, can posit itself as the source of new creativity. In actual fact, it is this body, intended as an operant form-production allowing for the

inscription of the file within itself, which finally articulates as a guide and support for the activity of ring-threading by conceptual *schemata* proper to the file itself, which determines the rising and the extended articulation of the neural connections at the level of the brain. This is the drawing which is ultimately donated: a drawing for the Other, however. The abstract frame in accordance with which the body progressively disincarnates itself, and which outlines the contours of cerebral connections, is related to the Other and is for the Other. While the body in which the mind is incarnated is my body, the brain through which the body is disincarnated (through simulation) is a brain which serves the intentionality of the Other, progressively inhabited by the meaning of the Other: indeed, it is the Other's brain in that I, as body, simulate it. Its constituting itself as autonomous unit marks and identifies my body–brain's constitution as an objective measuring device in the world and of the world.

Vision by principles can posit itself as a real basis for new creativity in that it is able, through simulation, to confer autonomy on the Other precisely as it is able to sound out its meaning. M.lle de Saint Loup allows the poet to hand himself down to the Other, and the poet's body, the work which is now a cathedral, to reach beyond self into the Other, finding real fulfilment in the achieved narration. The result of this is the possible emergence of a nucleus of new, individual creativity.

Selection is creative because it determines ever-new linguistic functions, ever-new processing units which support the effective articulation of new coherence patterns. The development of this creativity, however, would not be possible without the above mentioned transformation and the inner guide offered by the successful *compositio* of the different constraints in action. On the other hand, the very irruption could not take place if we were not able to explore the non-standard realm correctly, i.e. if we were not capable of outlining adequate non-standard models and continuously comparing, in an "intelligent" way, our actual native competence with the simulation recipes at work.

We can perceive the objective existence of specific (self-organising) models only insofar as we constitute ourselves into a sort of arch or gridiron for the articulation, at the second-order or higher-order level and in accordance with specific selective procedures, of a series of conceptual plots and fusions, a series that determines a radical metamorphosis of our intellectual capacities. It is exactly by means of the actual reflection on the new-generated abstract constructions as well as of the mirror represented by the successful invention of adequate standard (and non-standard) models that I shall finally be able to inspect the realisation of my autonomy, the progressive embodiment of my mental activities in a "new" coherent and self-sustained system.

Meaning can selectively express itself only through (a) the nested realisation, at the abstract level, of specific "fusion" processes, (b) the determination of specific schemes of coherence able to support this kind of realisation, (c) a more and more co-operative and unified articulation at the deep level of the primary informational fluxes, (d) the identification of a model able to reflect and renew within itself as a fixed point the (original) meaning "word". It shapes the forms in accordance with *Sinn* connections, precise stability factors, symmetry choices, coherent contractions and ramified completions. We can partially inspect (and "feel") this kind

of embodiment, at the level, for instance, of "categorial intuition", insofar as we successfully manage to reconstruct, identify and connect the attractors of this particular dynamic process. It is exactly by means of the successive identification and guided *compositio* of these varying attractors that we can manage to imprison the thread of meaning and identify the coherent and becoming texture of the constraints concerning the architecture of visual thinking. In this way we shall finally be able to obtain a first self-representation of our mental activities, thus realising a form of effective autonomy: a representation that exactly concerns the "narration" relative to the progressive opening of the eyes of our mind and the correlated constitution of the *Cogito* and its rules.

* * *

Given a structure $\mathbf{\Psi} = <\mathrm{D}, <\mathrm{D}_n >_{n\geq 1}, <\mathrm{G}^{\Psi}>_{\mathrm{G}\epsilon\mathrm{O.C.}}>$ and a second-order language L', we can distinguish many kinds of relations. For instance we can distinguish: (a) first and second-order relations on the universes of the structure; (b) relations into the universes of the structure; (c) definable relations of the structure using a given language, and so on.

It is important to underline that these kinds of relations which we have just referred to are not always restricted to the category of relations among individuals. In other words, not all of them are first-order relations: in this way we can realise that hidden in the structure, but definable with the second-order language in use, some relations exist that do not appear as relations among individuals but are utilised in order to define first-order relations. On the other hand, we know that in the universes of any second-order structure $\mathbf{\Psi}$ there are only relations among individuals; when the structure is standard all the relations among individuals are in the universes of the structure. As M. Manzano correctly remarks, in standard structures all the n-ary first-order relations on $\mathbf{\Psi}$ are into $\mathbf{\Psi}$.

In this sense, when we are faced with a standard structure the ur-elements are fixed and we cannot "inspect", with respect to the inner relational growth of the structure, the successive unfolding of some specific depth dimensions different from the simple dimension relative to the full exploitation of the "surface power" of the structure itself.

Things are different when we take into consideration structures with non-full relational universes. In order to understand the secret nature of this kind of passage it is useful to examine more carefully the problematics concerning the definition of non-standard models. As is well known, Skolem discovered the existence of non-standard models of arithmetic in the Thirties. At the end of the Forties Henkin utilised non-standard structures in order to prove his famous weak completeness theorem for the theory of types and, at the same time, outlined a non-standard model of $\mathbf{N}^2$.

In order to present the modalities of construction of this kind of non-standard model, let us, first of all, show how to build a non-standard model of the first-order theory of Peano arithmetic ($\mathbf{N}^1$): a very well known model, which results non-isomorphic with the structure $< \mathbb{N}, 0, \mathrm{S} >$, i.e., with the intended model of $\mathbf{N}^1$ in the natural numbers. The construction can also be carried out for the enlarged model $< \mathbb{N}, 0, \mathrm{S}, +, \cdot >$.

Consider the theory $\mathbf{N}^{1*}$ which results from $\mathbf{N}^1$ by adding the individual constant "a" together with the following infinite sequence of axioms, one for each natural number:

$a \neq 0$
$a \neq S0$
$a \neq SS0$
...

It is easy to show that the infinite set of axioms of $\mathbf{N}^{1*}$ will be consistent if $\mathbf{N}^1$ is consistent. Now, by Gödel's incompleteness theorem, any consistent set of first-order formulas has a model. But, the intended interpretation of $\mathbf{N}^1$ in the natural numbers cannot be a model of $\mathbf{N}^{1*}$. Actually, any model Q of $\mathbf{N}^{1*}$ must be a model of $\mathbf{N}^1$ and, at the same time, a model of the new formulas $a \neq 0$, $a \neq S0$, $a \neq SS0$. Therefore, the universe of Q contains non-standard numbers.

We know that for every infinite cardinal, there are at least $2^{\aleph^\circ}$ non-isomorphic models of $\mathbf{N}^1$ of that cardinality. Those models of $\mathbf{N}^1$ that are isomorphic with the intended model of $\mathbf{N}^1$ are its *standard models*. All other models are *non-standard models*.

Now, let us imagine building a particular extension of our non-standard model of the first-order theory of natural numbers, Q, that is a second-order structure out of it capable of presenting itself as a model of $\mathbf{N}^2$. Let us call this structure Q′. It is easy to show that if the structure Q′ were, as required, a model of $\mathbf{N}^2$, it must be non-standard in the second-order sense: i.e., such that each $D_n \subseteq \mathbf{P}D^n$ and $D_m \neq \mathbf{P}D^m$ for at least one $m \geq 1$. Actually, in the universe of Q′ there are non-standard numbers. This means that the set of standard numbers is not in the universe of unary relations of Q′. Thus, the structure Q′ is non-standard in the second-order sense.

When we are in second-order logic, but we make essential reference to non-standard interpretations and allow structures with non-full relational universes, quantification only applies for the sets and relations that are present in the structure. In the general structures of Henkin, for instance, we put in the universes all sets and relations that are parametrically definable in the structure by second-order formulas. In this sense, it is not surprising that the set of standard numbers is not definable by a second-order formula in a structure having non-standard numbers. If we indicate with P. Def. $(\mathbf{\Psi}, L')$ the set of all parametrically $\mathbf{\Psi}$ definable relations on individuals using the language L′, we can say directly that a given frame $\mathbf{\Psi}$ is a general structure iff $D_n = \mathbf{P}D^n \cap \text{P.Def.}(\Psi, L')$.

What it is important to stress again is the fact that hidden in the structure some specific relations exist, some "rules" (second-order relations) that cannot be defined as relations among individuals, but are utilised to define first-order relations (i.e., relations among individuals). As a result, we obtain a particular structure where the n-ary relation universe is a proper subset of the power set of the n-ary Cartesian product of the universe of individuals. So, whereas in the standard structures the notion of subset is fixed and an n-ary relation variable refers to any n-ary relation on the universe of individuals, in the non-standard structures, on the contrary, the notion of subset is explicitly given with respect to each model. Thus, in the case

of general structures the concept of subset appears directly related to the definition of a particular kind of constructible universe, a universe that we can explore utilising, for instance, the suggestions offered by Skolem (cf. his attempt to introduce the notion of propositional function axiomatically) or by Gödel (cf. Gödel's notion of constructible universe).

From a more general and philosophical point of view, we can say that at the level of general structures, the relations among individuals appears as submitted to a bunch of constraints, specifications and rules having a relational character, a bunch that is relative to the model which we refer to and that acts "from the outside" on the successive configurations of the first-order relations. In other words, as we have just remarked, in the universes of any second-order frame Ψ there are only relations among individuals, but it is no longer true that all the n-ary first-order relations on Ψ are into Ψ.

These hidden relations, these particular "constraints" play a central role with respect to the genesis of our models. In particular, let us remark that as a consequence of the action performed by these constraints, the function played by the individuals living in the original universe becomes more and more complex. We are no longer faced with a form of unidimensional relational growth starting from a given set of individuals and successively exploring all the possible relations among individuals, according to a pre-established surface unfolding of the relational texture. Actually, besides this kind of unidimensional growth, further growth dimensions reveal themselves at the second-order level; specific types of development that spring from the successive articulation of the original growth in accordance with a specific dialectics. Such a dialectics precisely concerns the interplay existing between the first-order characterisation of the universe of individuals and the whole field of relations and constraints acting on this universe at the second-order level. As a result of the action of the rules lying at the second-order level, new dimensions of growth, new dynamic relational textures appear. Contemporarily the original universe of individuals changes, new elements grow up and the role and nature of the ancient elements undergo a radical transformation. In this sense, the identification of new growth dimensions necessarily articulates through the successive construction of new *substrata*. The aforesaid dialectics reveals itself as linked to the utilisation of specific conceptual tools: limitation procedures, identification of fixed points, processes of self-reflection and self-representation, invention of new frames by "fusion" of previously established structures, coagulum functions, etc.

Moreover, as we shall see, we have to recognise the presence of specific patterns of selection and differentiation. Discovery procedures, construction processes, coagulum functions, selective pressures act as a chorus of functions in unison in order to shape the varying texture of mental constructions. At the level of this chorus (if successful) *omnis determinatio est negatio*. The plot of limitation procedures and cancellations of relations progressively constitutes itself as the gridiron of an intellectual order capable of allowing for the successive "production" of specific *Gestalten*. If we are able to recognise and follow the secret path of this order, we can finally manage to illuminate the "good" structures and to "read" (and "play") the progressive embodiment of the *Sinn* that selectively determines the real constitution of the events.

What we have remarked until now permits us to understand more deeply the ultimate sense of Henkin's conceptual revolution. As M. Manzano correctly remarks, Henkin arrived to prove the completeness theorem for type theory, "... by changing the semantics and hence the logic. Roughly presented, the idea is very simple: The set of validities is so wide because our class of standard structures is too small. We have been very restrictive when requiring the relational universes of any model to contain all possible relations (where "possible" means in the background set theory used as metalanguage) and we have paid a high price for it. If we also allow non-standard structures, and if we now interpret validity as being true in all general models, redefining all the semantic notions referring to this larger class of general structures, completeness (in both weak and strong senses), Löwenheim–Skolem, and all these theorems can be proved as in first-order logic."[1]

In this sense, in accordance for instance with Németi's opinion, standard semantics is not logically adequate because it does not include all logically possible worlds as models. On the contrary, in Henkin's general semantics many "hidden" possibilities are progressively taken into consideration as possible models. We can have, for instance, models with or without GCH (generalised continuum hypothesis). Things are really different in the case of standard semantics.

This argument can be extended in a significant way. Actually, according to Gödel's incompleteness theorem it is possible to prove that a precise link does exist between non-standard models and formally undecidable propositions. On the other hand, we have just seen how it is possible to outline, according to Henkin's results, a model containing a non-standard number system which will satisfy all of the Peano postulates, as well as any preassigned set of further axioms. We only have to introduce a new primitive "a" and add to the given set of axioms the infinite list of formulas, $a \neq 0$, $a \neq S0$, $a \neq SS0$, ...

By adding a non denumerable number of primitive constants b_i^{ξ} together with all formulas $b_i^{\xi_1} \neq b_i^{\xi_2}$ for $\xi_1 \neq \xi_2$, we may even build models for which the Peano axioms are valid and which contain a number system having any given cardinal. This kind of theoretical construction shows, as we have just said, that no mathematical axiom system can be categorical, unless it constrains its universe of elements to have some specific finite cardinal number.

The conceptual importance of the discovery of non-standard models can be well understood if we try to elucidate a precise dialectical aspect characterising the link between non-standard models and undecidable propositions. Actually, even if Gödel's theorems indicate to us how to build certain formulas which are shown to be true but unprovable, however, as Henkin remarks, there is no general method indicated for establishing that a given theorem cannot be proved from given axioms. Such a method is exactly supplied by the procedures of constructing step by step non-standard models for number theory in which "set" and "function" are reinterpreted. These procedures show us that in order to model thought processes in an adequate way we must explore the non-standard realm on the basis, first of all, of

[1] Cf. Manzano (1996), p. XVI.

the identification of precise fixed points and the tentative definition of new kinds of universes. For example, we know that, in accordance with Mostowski's results, Gödel's famous undecidable proposition can be simply considered as a proposition that characterises the class of natural numbers. If we refer this proposition to a system of non-standard numbers, it will be no longer valid. In this way, we can realise that, along our exploration, we are really driving specific "conceptual" stakes into the ground of an unknown territory and that this exploration articulates in a co-evolutive landscape. At the level of this particular landscape constructing and discovering appear as dialectically interrelated.

It is precisely by means of this exploration process that we can ascertain that a formal system can admit models with a universe of individuals that does not have the order type of natural numbers. Henkin explicitly quotes, as an example, a simple result, every non-standard denumerable model for the Peano axioms has the order type $\omega + (\omega^* + \omega)\eta$, where η is the type of the rationals.

If we make essential reference to non-standard structures, then the set of validities is considerably reduced. At the same time, if we interpret validity as being true in all general models, completeness, Löwenheim–Skolem theorems and other well-known theorems can be proved as in the case of first-order logic. As a matter of fact, the set of validities will coincide with the set of sentences derivable in a calculus, which is an extension of the first-order calculus. However, this kind of reduction will reveal itself as successful only if we are able to explore the non-standard realm in an intelligent and "creative" way and if the arising differentiation processes articulate in accordance with precise coherence patterns and stability factors.

From a general point of view, the limitation theorems are theorems that are based on a precise distinction between theory and metatheory, between language and metalanguage. A formal system can be considered as an "objectified" language and we well know that by means of precise arithmetical procedures the syntactical properties of a given formalised theory T can be expressed in terms of arithmetic predicates and functions.

We have just remarked, for instance, that Gödel's incompleteness theorem concerns a sentence of **Z** (where **Z** is a formal system obtained by combining Peano's axioms for the natural numbers with the logic of type theory as developed in *Principia Mathematica*) which says of the sentence itself that it is not provable in **Z**. However, the existence of such a sentence can be identified only because we are able to arithmetise metamathematics: i.e., to replace assertions about a formal system by equivalent number-theoretic statements and express these statements within the formal system.

In this sense, as we have said before, limitation theorems show that that particular reality (or "essence") represented by "arithmetical truth" is not exhausted in a purely syntactical concept of provability. From a more general point of view, we can directly affirm that in **Z** we cannot define the notion of truth for the system itself. In other words, by constructing a system and then treating the system as an object of our study, we create some new problems, which can be formulated but cannot be answered in the given system. Actually, every sufficiently rich formal system is always submitted to the diagonal argument, an argument that is always present

in the limitation theorems and, in particular, in the Löwenheim–Skolem theorem. Let us show, as a simple consequence of this last theorem, how one can prove that no formalised set theory can give us all sets of positive integers. Let S be a standard system of set theory. Since we can enumerate the theorems of S, we can also enumerate those theorems of S each of which asserts the existence of a set of positive integers. Let us consider now the set J of positive integers such that for each m, m belongs to J iff m does not belong to the m-th set in the enumeration. By Cantor's argument, J cannot occur in the enumeration of all those sets of positive integers which can be proved to exist in S. Hence, either there is no statement in S which affirms the existence of J, or, if there is such a statement, it is not a theorem of S. In either case, there exists a set of positive integers which cannot be proved to exist in S. In other words, the axioms of our formal system cannot give us a representation of all sets of positive integers. It is precisely in this sense that the systems containing these axioms must necessarily admit non-standard models.

Thus, the limitation procedures permit us to identify the boundaries of our intellectual constructions, to characterise, for instance, as we have just remarked, the class of natural numbers. They permit us to "see", once given a specific representation system **W**, that if **W** is normal then every predicate H (the predicates, in this particular case, can be thought of as names of properties of numbers) has a fixed point. They also permit us, for example, to identify an unlimited series of new arithmetic axioms, in the form of Gödel sentences, that one can add to the ancient axioms. Then, we can use this new system of axioms in order to solve problems that were previously undecidable.

We are faced with a particular form of mental "exploration" that, if successful, embodies in an effective construction constraining the paths of our intellectual activity. This exploration concerns the identification of new worlds, of new patterns of relations, the very characterisation of new universes of individuals. We shall have, as a consequence, the progressive unfolding of an articulated process of cancellation of previously established relations and the birth of new development "languages" that are grafted on the original relational growth. As we have said before, this type of mental exploration articulates at the second-order level: it can be reduced however (if successful) at the level of many-sorted first-order logic, by means of well known logical procedures.

In a nutshell, the nucleus of this kind of reduction consists in explicitly showing in many-sorted structures what is implicitly given in second-order or in type theory. According to Post's famous thesis, any law we become completely conscious of can be mechanically constructed. So, we add to the many-sorted language membership relation symbols and to the many-sorted structures membership relations as relation constants. Throughout this reduction process, we simply consider that a second-order structure (or a type theory structure) is basically a peculiar many-sorted structure, since it has several domains. In short, we prove first of all that Henkin semantics and many-sorted first-order semantics are pretty much the same. Then, via Henkin semantics, we establish a form of reduction of second-order semantics to first-order semantics. Second-order logic with the Henkin semantics is, in general terms, a many-sorted logic.

However, we immediately have to emphasise that this kind of reduction does not imply that the secret "reasons" that guide, from within the mental activity, the progressive unfolding of the processes of exploration and invention can be reduced to a first-order mechanism or to a set of pre-established rules.

It is true that insofar as the aforementioned exploration process manages to embody in an effective construction that acts as a bunch of constraints and classification procedures, then we have the possibility to translate this kind of structure in a many-sorted language. But, the actual unfolding of abstract procedures that constitutes the primitive nucleus of the exploration process necessarily articulates (at least) at the second or higher-order level. As a matter of fact, the first result of this very unfolding is the birth of specific (and previously unknown) differentiation processes, as well as the successive appearance of new universes of individuals.

Let us quote Gödel, "P. Bernays has pointed out on several occasions that, in view of the fact that the consistency of a formal system cannot be proved by any deduction procedures available in the system itself, it is necessary to go beyond the framework of finitary mathematics in Hilbert's sense in order to prove the consistency of classical mathematics or even of classical number theory. Since finitary mathematics is defined as the mathematics of *concrete intuition*, this seems to imply that *abstract concepts* are needed for the proof of consistency of number theory.... By abstract concepts, in this context, are meant concepts which are essentially of the second or higher level, i.e., which do not have as their content properties or relations of concrete objects (such as combinations of symbols), but rather of *thought structures* or *thought contents* (e.g., proofs, meaningful propositions, and so on), where in the proofs of propositions about these mental objects insights are needed which are not derived from a reflection upon the combinatorial (space–time) properties of the symbols representing them, but rather from a reflection upon the meanings involved."[2]

In this sense, there must be proofs that are not fully formalisable at a given stage in our mental experience, but that are "evident" to us at that stage on the basis of particular arrangements of limitation procedures, of the successive identification of fixed points, of the utilisation of abstract concepts, of the exploration of new universes of individuals, and so on.

In other words, there are, for instance, proofs of Con (PA) (primitive arithmetic) that require abstract concepts as well as the necessary construction of new elements; concepts, for instance, that are not immediately available to concrete intuition (Hilbert's concrete intuition as restricted to finite sign-configurations). We need, in general, not only rules, but also rules capable of changing the previously established rules. In Gödel's consistency proof, for example, we can directly see that the theory of primitive recursive functionals requires the abstract concept of a "computable function of type t".

Thinking in mathematical terms cannot be completely constrained within the boundaries of the syntax of a specific language. In fact, we would also need to know that the rules of this particular syntactical system are consistent. But in order to

[2] Cf. Gödel (1972) in Feferman et al. (1990), pp 271–272.

realise this, we will, by the second incompleteness theorem, as we have seen before, need to use mathematics that is not captured by the rules in question.

According to Gödel, utilising mathematical reason we are capable of outlining and, at the same time, discovering specific abstract relations that live at the second-order level and that we utilise and explore at that stage in order to define first-order relations. We are faced with a particular "presentation" of the Fregean *Sinn*, a presentation that selectively constrains the paths of our reasoning in a significant way. So, abstract and non-finitary intellectual constructions are used to formulate the syntactical rules. Once again, this is for many aspects a simple consequence of incompleteness results: mental constructions cannot be exhausted in formal concepts and purely syntactical methods. We have, in general, to utilise more and more abstract concepts in order to solve lower level problems.

The utilisation at the semantic level of abstract concepts, the possibility of referring to the sense of symbols and not only to their combinatorial properties, the possibility of picking up the deep information living in mathematical structures open up new horizons with respect to our understanding of the ultimate nature of mental processes. We are actually dealing with a kind of categorial perception (or rational perception) that does not concern simple data (relative to the inspectable evidence), but complex conceptual constructions. And we know that, in Husserlian terms, meaning "shapes" the forms creatively. However, we must immediately remark that categorial perception appears to embody in a realm that is far beyond the limits of Gödel's primitive suggestions, in particular of his primitive Platonist approach. Actually, at the level of the articulation of mental constructions, we are faced with the existence of precise forms of co-evolution. On the one hand, we can recognise, at the level of the aforementioned process of inventive exploration, not only the presence of forms of self-reflection but also the progressive unfolding of specific fusion and integration functions, on the other hand, we find that the *Sinn* that embodies in specific and articulated rational intuitions guides and shapes, in a selective way, the paths of the exploration. It appears to determine, by means of the definition of precise constraints, the choice of some privileged patterns of functional dependencies, with respect to the entire relational growth. As a result, we can inspect a precise spreading of the development dimensions, a selective cancellation of relations and the rising of specific differentiation processes. We are faced with a new theoretical landscape characterised by the unfolding of a precise co-evolution process, by the first articulation, in particular, of specific mental processes submitted to the action of well specified selective pressures, to a continuous "intervention" of depth information determining the successive appearance, at the surface level, of specific *Gestalten*. This intervention, however, could not take place if we were not able to explore the non-standard realm in the right way, if we were not capable of outlining adequate non-standard models and continuously comparing our actual native competence with the simulation recipes. Meaning selection is creative because it determines ever-new symbolic functions, ever-new processing units which support the effective articulation of new coherence patterns. And, it is precisely by means of these new patterns that we shall be able to "narrate" our inner transformation, to become aware of our mental development and, at the same time, to ascertain the objective character of the transformation undergone.

We can perceive the objective existence of abstract concepts only insofar as we transform ourselves into a sort of arch or gridiron for the articulation, at the second-order or higher-order level and in accordance with specific selective procedures, of a series of conceptual plots and fusions, a series that determines a radical transformation of our intellectual capacities. It is exactly by means of the actual reflection on the new-generated abstract constructions that I shall finally be able to inspect the realisation of my autonomy, the progressive embodiment of my mental activities in a "new" unitary system. At the level of Skolem's conception, for instance, ideas such as countability and uncountability are inherently relative: our belief that the power set of the natural numbers, $\mathbf{P}(\omega)$, is uncountable is correct but must be understood relative to our own current viewpoint; from the point of view of another "observer", this set may in fact be considered as countable. Actually, we can identify some sets as uncountable only in the sense that there does not exist within the model a bijection from the natural numbers onto the sets; it is perfectly possible, however, that such a bijection exists outside the model. So, for an internal observer living within the model a specific set can appear as uncountable, on the contrary, for an external observer the very set seems countable. From a more general point of view, we well know that there are some powerful characterisations of the system of natural numbers within an ambient set theory: according to Skolem's point of view, these set-theoretic characterisations are all relative. An internal observer, for instance, can find that in his world there is just one "system of natural numbers" satisfying Peano's second-order postulates. An external observer, however, can easily realise that this particular system is in fact non standard, containing infinite unnatural numbers. What it is important to underline in this context, is the role played by the different observers and by the successive identification of the different ontologies. Things are even more complicated if we postulate, for instance, the existence of a circular link between the different observers in a co-evolutive ambient: the ontologies will undergo continuous changes. Then, according to this line of thought, we can easily realise the importance of the progressive constitution at the co-evolutive level of the mind's eyes and the role played, with respect to this genesis, by the successive conceptual exploration of non-standard models.

* * *

With respect to this frame of reference, Reality presents itself as a set of becoming processes characterised by the presence-irradiation of a specific body of meaning and by an inner (iterative) *compositio* of generative fluxes having an original character. These processes then gradually articulate through and in a (partially-consistent) unifying development warp with internal fluctuations of functional patterns. It is this functional, self-organising and "irradiating" warp, in the conditions of "fragmentation" in which it appears and is reflected at the interface level through the unfolding of the canalisation process, that the network-model progressively manages to reconstruct and replicate within itself as regards its specific functional aspects, ultimately synthesising and reflecting it into an operating architecture of causal programs. In this way, it is then possible to identify a whole, complex "score" which will function as the basis for the reading-reconstruction of

the aforementioned functional warp. However, to read–identify–represent the score will necessarily require the contemporary discovery-hearing of the underlying harmony. Only the individual capable of representing and tuning the work as living harmony, and the score as silent object, will actually be able to depict him/herself as "I" and as subject. This individual will then not only be able to observe objects, but will itself be able to see the observing eye, modeling those objects. The "I" able to portray itself as such will be able to rediscover the root of the very act of seeing, positing itself as awareness and as the instrument allowing the emergence of the "thinking I", and, conjointly, of the metamorphosis of the original meaning.

It is thus through the continuous metamorphosis of the network that new Nature can begin to speak, and Reality can channel itself (*in primis* as regards the external selection), in accordance with its deep dimension, ultimately surfacing and expressing as an activity of synthetic multiplication, i.e. as a form of operating generativity at the level of surface information and as a "thinking I" able to reflect itself in (and through) the work outlined by the network-model.

It is the face-texture of the effected reconstruction which provides the guidelines for the I's edification; and indeed the "thinking I" which gradually surfaces reflects itself in the constructed work, thereby allowing the effective emergence of an "observer" which reveals finally itself as a cognitive agent able to observe the Nature around him in accordance with the truth, i.e. we are actually faced with the very multiplication of the cognitive units. The system is thus able to see according to the truth insofar as it constitutes itself as an "I" and as consciousness, i.e. in proportion to the extent it can "see" (and portray-represent) its own eye which observes things.

In this sense vision is neither ordering, nor recognising, nor pure comparison, nor, in general, simple replica, but is above all a reading-reconstruction of the (becoming) unity of the original body of meaning (with operating self-reflection): a process of progressive identification and assimilation of this unity in terms of an adequate texture of self-organising programs able to portray itself as such, a process which becomes gradually autonomous and through which, via selection, in a renewed way and at the surface level, Reality can canalise the primary modules of its own complex creative tissue: i.e. surfacing as generativity and nesting as meaning. The better the reconstruction, the more adequate and consistent the canalisation: the system will function ever more sophisticatedly as a reflecting and self-organising filter. As a matter of fact, in parallel to this an "observer" will progressively arise through the narration and the methodical verification of the distinctions relative to the functional forms managing to move at the level of the unitary and cohesive articulation of the self-organising programs. As narration and synthesis, the "I" posits itself as autonomous and as the increasingly adequate mirror of a precise "metamorphosis": namely, the metamorphosis proper to an intelligent network which grows into autonomy. The mirror is image-filled at the moment of selection, when new emergence can simultaneously come about and "eyes" can then open and see both things and their meaning.

The "thinking I" which surfaces and the meaning which emerges thus fuse in the expression of a work which ultimately manages to articulate and unite itself with the awareness-*Cogito* and the ongoing narration: an observer thus joins a work

acting as a filigree. The resulting path-Via can then allow real conjunction of both function and meaning. The result will be not merely simple generative principles, but self-organising forms in action, creativity in action, and real cognitive multiplication: not a simple gestaltic restructuring, but the growth and multiplication of cognitive processes and units, i.e. the actual regeneration and multiplication of original Source according to the truth.

The adequate work of unification-closure of network programs, which joins and encapsulates, at the level of the ongoing emergence and self-reflection, the selection internally operated by meaning according to the living warp-filigree, constitutes the real basis of vision in action. In actual fact it comprises a multiplicity of interconnected works, to each of which is linked a consciousness. In this way the aforementioned unification necessarily concerns the continuous weaving of a unitary consciousness, albeit within the original fragmentation of the micro-consciousness and the divided self.

It is from this viewpoint that vision appears as necessarily related to a continuous emergence, in its turn connected primarily with the progressive articulation of a self-expressing and self-synthesising "I". As the system manages to see, it surfaces towards itself and can, then, identify and narrate itself as an "I", and specifically as an "I" that sees and grasps the meaning of things: in particular the emergence related to the meaning that is concerned with them. At the moment the aforementioned work becomes vision (expressing itself in its completeness), it simultaneously reveals itself as a construction in action and at the same time as the filter and the lynch-pin of a new canalisation through which new Reality can reveal itself unfolding its deep creativity. Meaningful forms will then come into play, find reflection in a work, and be seen by an "I" that can thus construct itself and re-emerge, an "I" that can finally reveal itself as autonomous: real cognition in action.

I neither order nor regiment according to principles, nor even grasp principles, but posit myself as the instrument for their recovery and recreation, and reflect their sedimentation in my self-transformation and my self-proposing as *Cogito*. Actually, I posit my work as the mirror for the new canalisation, in such a way that the new emergent work (the self-organising mirror), if successful, can claim to be the work of an "I" which posits itself as an "added" observer. It is not the things themselves that I "see", then, but the true and new principles, i.e. the meaningful forms in action: the rules-functions linked to their emergent meanings. I thus base myself on the "word" which dictates. Hence the possibility of seeing Nature *iuxta propria principia.*

The world thus perceived at the visual level is constituted not by objects or static forms, but by processes appearing imbued with meaning. As Kanizsa stated, at the visual level the line per se does not exist: only the line which enters, goes behind, divides, etc.: a line evolving according to a precise holistic context, in comparison with which function and meaning are indissolubly interlinked. The static line is in actual fact the result of a dynamic compensation of forces. Just as the meaning of words is connected with a universe of highly-dynamic functions and functional processes which operate syntheses, cancellations, integrations, etc. (a universe which can only be described in terms of symbolic dynamics), in the same way, at the level of vision, I must continuously unravel and construct schemata; must assimilate and

make myself available for selection by the co-ordinated information penetrating from external reality. Lastly, I must interrelate all this with the internal selection mechanisms through a precise "journey" into the regions of intensionality.

The resulting global determination will present itself as something "perceived" insofar as it will reveal itself as linked to precise postulates of meaning: it will thus emerge as a scene (a scene for an I-subject), and the single processes of determination as meaningful observers or as objects, actions, etc. which populate the scene and which result as encapsulated in observation systems. The I-subject will recognise itself through the co-ordinated action of these observation systems; it will mirror itself in the "pupils" of these very systems to the extent that it will be recognised as the primary factor of their recovery as autonomous units.

In this sense, Nature is the very (original) opening of the process of determination. It presents itself as a dynamic system of meaningful processes in action; the "method" in its turn must offer real instruments in order to feed and coagulate the self-organising growth and the articulated unfolding of these very processes. On the other hand Nature can be also considered as a body-system of meaning that cannot be occupied. Hence the possibility to consider Nature contemporarily as both "irruption" and emergence, as deep information that hides itself with the ever-new emergence of postulates of meaning (*Natura Naturans*); to this emergence will correspond the progressive "surfacing" of ever-new constraints and rules at the generative level.

Vision is partially objective and veridical – veridical mainly since, through the effected selection and canalisation, it appears anchored to the revelation of the original creativity, to the actual unfolding and opening of the maximal determination. It does, indeed, seem able to unravel its inner creativity in accordance with its message, thus providing a coherent filter for the realisation of an adequate biological canalisation. It is namely veridical since there can only be objective vision if the enacted simulation and inscription emerge as truth and posit themselves as the basis for new creation: in this way the eyes of the "flesh" will coincide with the eyes of the mind.

Vision, in this sense, is the process of inscription, reconstruction, assimilation and reduction realised in the conditions of double selection in accordance with the truth. It appears necessarily moulded by the mathematical forms and the modules which determine and shape it; in particular it articulates as a coupled pattern of emergence and irruption, thus finally constituting itself as the vision of an "I" which manages to establish its full autonomy. As unfolded I–*Cogito*, I see primarily developmental processes articulated tri-dimensionally and originally possessing meaning. Vision thus appears as the embodiment of the method relative to the process of canalisation of the generative fluxes – principles in action: if adequate, it constitutes the way in order to partially permit the real unfolding of the deep information content in accordance with different and successive levels of complexity: it articulates each time through the reading – individuation of that particular co-ordinated series of functional closures, i.e. that specific chain of fixed points that is necessary for the coherent unfolding and encapsulation of the *Sinn* according to its original virtuality.

References

Atlan H (1992) Self-organizing networks: weak, strong and intentional, the role of their underdetermination. La Nuova Critica 19–20:51–71

Carnap R, Bar Hillel Y (1950) An outline of a theory of semantic information. MIT, Tech. Rep. N.247

Carsetti A (1993) Meaning and complexity: the role of non-standard models. La Nuova Critica 22:57–86

Carsetti A (2000) Randomness, information and meaningful complexity: some remarks about the emergence of biological structures. La Nuova Critica 36:47–109

Carsetti A (ed) (2000) Functional models of cognition. Self-organizing dynamics and semantic structures in cognitive systems. Kluwer Academic, Dordrecht

Carsetti A (ed) (2004) Seeing, thinking and knowing. meaning and self-organisation in visual cognition and thought. Kluwer Academic, Dordrecht

Chaitin G (1987) Algorithmic information theory. Cambridge University Press, Cambridge

Feferman S et al. (eds) (1986, 1990, 1995) Kurt Gödel: collected works, I, II, III. Oxford University Press, Oxford

Gaifman H (2000) What Gödel's incompleteness result does and does not show. J Philos 9708:462–470

Grossberg S (2000) Linking mind to brain: the mathematics of biological intelligence. Notices AMS 47:1358–1374

Henkin L (1950) Completeness in the theory of types. J Symb Logic 15:81–91

Hintikka J (1970) Surface information and depth information. In: Hintikka J, Suppes P (eds) Information and inference. Reidel, Dordrecht, pp 298–330

Hoffman DD (1998) Visual intelligence: how we create what we see. W.W. Norton, New York

Kanizsa G (1980) Grammatica del vedere. Il Mulino Bologna

Kohonen R (1984) Self-organization and associative memories. Springer, Berlin

Kozen D. et al. (eds) (1982) Logic of programs. Springer, Berlin

Manzano M (1996) Extensions of first-order logic. Cambridge University Press, Cambridge

Németi I (1981) Non-standard dynamic logic. In: Kozen D et al. (eds) Logic of programs. Springer, Berlin

Talmy L (2000) Toward a cognitive semantics. MIT, Cambridge

Van Dalen D (1983) Logic and structure. Springer, Berlin

Wang H (1974) From mathematics to philosophy. Routledge & Kegan Paul, New York

Author Index

Subject Index