UNDERSTANDING COMPLEX SYSTEMS

Springer: COMPLEXITY

Nancey Murphy
George F.R. Ellis
Timothy O'Connor
Editors

Downward Causation and the Neurobiology of Free Will

Springer

Springer Complexity

Springer Complexity is an interdisciplinary program publishing the best research and academic-level teaching on both fundamental and applied aspects of complex systems - cutting across all traditional disciplines of the natural and life sciences, engineering, economics, medicine, neuroscience, social and computer science.

Complex Systems are systems that comprise many interacting parts with the ability to generate a new quality of macroscopic collective behavior the manifestations of which are the spontaneous formation of distinctive temporal, spatial or functional structures. Models of such systems can be successfully mapped onto quite diverse “real-life" situations like the climate, the coherent emission of light from lasers, chemical reaction-diffusion systems, biological cellular networks, the dynamics of stock markets and of the internet, earthquake statistics and prediction, freeway traffic, the human brain, or the formation of opinions in social systems, to name just some of the popular applications.

Although their scope and methodologies overlap somewhat, one can distinguish the following main concepts and tools: self-organization, nonlinear dynamics, synergetics, turbulence, dynamical systems, catastrophes, instabilities, stochastic processes, chaos, graphs and networks, cellular automata, adaptive systems, genetic algorithms and computational intelligence.

The two major book publication platforms of the Springer Complexity program are the monograph series “Understanding Complex Systems" focusing on the various applications of complexity, and the “Springer Series in Synergetics", which is devoted to the quantitative theoretical and methodological foundations. In addition to the books in these two core series, the program also incorporates individual titles ranging from textbooks to major reference works.

Editorial and Programme Advisory Board

Understanding Complex Systems

Founding Editor: J.A. Scott Kelso

Future scientific and technological developments in many fields will necessarily depend upon coming to grips with complex systems. Such systems are complex in both their composition - typically many different kinds of components interacting simultaneously and nonlinearly with each other and their environments on multiple levels - and in the rich diversity of behavior of which they are capable.

The Springer Series in Understanding Complex Systems series (UCS) promotes new strategies and paradigms for understanding and realizing applications of complex systems research in a wide variety of fields and endeavors. UCS is explicitly transdisciplinary. It has three main goals: First, to elaborate the concepts, methods and tools of complex systems at all levels of description and in all scientific fields, especially newly emerging areas within the life, social, behavioral, economic, neuroand cognitive sciences (and derivatives thereof); second, to encourage novel applications of these ideas in various fields of engineering and computation such as robotics, nano-technology and informatics; third, to provide a single forum within which commonalities and differences in the workings of complex systems may be discerned, hence leading to deeper insight and understanding.

UCS will publish monographs, lecture notes and selected edited contributions aimed at communicating new findings to a large multidisciplinary audience.

Nancey Murphy, George F.R. Ellis,
and Timothy O'Connor (Eds.)

Downward Causation and the Neurobiology of Free Will

Springer

Editors

Prof. Nancey Murphy
Department of Philosophy
Fuller Graduate Schools
135 N. Oakland Ave.
Pasadena, CA 91182
USA
E-mail: nmurphy@fuller.edu

Prof. Timothy O'Connor
Department of Philosophy
Indiana University
1033 E. Third St., Sycamore Hall 026
Bloomington, IN 47405-7005
USA
E-mail: toconnor@indiana.edu

Prof. George Ellis
Mathematics Department
University of Cape Town
Private Bag
Rondebosch 7701
South Africa
E-mail: george.ellis@uct.ac.za

ISBN 978-3-642-03204-2 e-ISBN 978-3-642-03205-9

DOI 10.1007/978-3-642-03205-9

Understanding Complex Systems ISSN 1860-0832

Library of Congress Control Number: Applied for

Typeset & Cover Design: Scientific Publishing Services Pvt. Ltd., Chennai, India.

Printed in acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

The nature of our understanding of free will in the light of present-day neuroscience is becoming increasingly important because of remarkable discoveries on the topic being made by neuroscientists at the present time, on the one hand, and its crucial importance for the way we view ourselves as human beings, on the other. It also has major implications for our understanding of ethical responsibility and hence for public policy and legal issues.

This book arises out of a workshop held in California in April of 2007, which was chaired by Dr. Christof Koch, and arranged by Dr. Mary Ann Meyers, with funding from the John Templeton Foundation. It was unusual in terms of the breadth of people involved: they included physicists, neuroscientists, psychiatrists, philosophers, and theologians. This enabled the meeting, and hence the resulting book, to attain a rather broader perspective on the issue than is often attained at academic symposia. The book is further enriched by chapters from some who were not present at the meeting (Sarah-Jayne Blakemore, David Hodgson, Owen D. Jones, Alicia Juarrero, Hakwan C. Lau, Dean Mobbs, and Emmanuelle Tognoli) that contribute especially to the important topic of how neuroscience is already impinging on legal issues.

The editors wish to express their appreciation to the Templeton Foundation, to Christof Koch, all of the participants, including three who are not represented among the book's contributors: Itzhak Fried, Güven Güzeldere, and Daniel Wegner. We thank our editor at Springer, Dr. Thomas Ditzinger. Particular gratitude goes to two people: Mary Ann Meyers, without whose tireless persuasion this book would not have materialized, and Susan Carlson Wood, who has edited and converted our many and various typescripts into camera-ready copy for publication. (We also thank Fuller Theological Seminary for making Susan's assistance available to us.)

Pasadena, California
March 2009

Nancey Murphy
George F.R. Ellis
Timothy O'Connor

Contents

Part III: Broader Understandings of Volition and Consciousness

Part IV: Human Implications of the Debate

1

Introduction and Overview

Nancey Murphy

School of Theology
Fuller Theological Seminary
Pasadena, CA 91182
nmurphy@fuller.edu

Summary. This chapter provides an overview of some of the history of debates regarding free will, and concurs with several authors who claim that the philosophical discussions have reached a stalemate due to their focus on a metaphysical doctrine of universal determinism. The way ahead, therefore, requires two developments. One is to focus not on determinism but on reductionism; the other is to attend to specific scientific findings that appear to call free will into question. The chapter provides an introduction to the topics of reductionism, emergence, and downward causation, and then surveys the works of Daniel Wegner and Benjamin Libet, which have been taken to show the irrelevance of conscious will in human action. It summarizes the chapters comprising the rest of the volume, and then offers a reflection on the achievement of the work as a whole – in brief, a critique of free-will skeptics based on human capacities such as meta-cognition and long-term planning, which allow agents to exert downward control on neural processes and behavior. It ends by highlighting, in light of Alasdair MacIntyre's work on moral responsibility, an important additional factor involved in creating the possibility for freedom of choice, namely the possession of abstract symbolic language.

Keywords: voluntary action, bottom-up causation, downward causation, top-down causation, emergence, free will, symbolic language, Benjamin Libet, Alasdair MacIntyre, self-transcendence, complex dynamical systems, Daniel Wegner.

1 Historical Debates

This section provides context for the chapters that follow. First I comment on the state of philosophical discussions of free will. Second, I trace some of the history of the development of concepts of reductionism, on the one hand, and of emergence and downward causation, on the other. Finally I introduce some of the recent research in cognitive neuroscience and psychology which many have taken to call free will into question, and to which many chapters in this volume respond.

1.1 The Stalled Free-Will Debate

Although philosophers tend to speak of "the free-will problem" this is misleading in at least two ways. First *the will* is a concept with a history in the West, approximately from Augustine's fifth-century account of the hierarchical faculties of the soul to Gilbert Ryle's critique of "the myth of volitions" (Ryle 1949, pp. 62–69). Although I agree with Ryle that there is no such thing as a will and that we would be better off speaking in terms of voluntary versus involuntary actions, I shall continue to use the conventional terminology here.

Second, it is misleading to speak of *the* free-will problem. Over the centuries, philosophers and theologians have debated a number of problems that share a family resemblance. Ancient Greek dramatists explored the role of fate. In the early Christian era two problems arose: First, if God had predestined some humans to be saved, is this reconcilable with anyone's freely choosing to obey the will of God? The second problem is whether human freedom is reconcilable with divine foreknowledge. This topic is still hotly debated. Yet another problem, prominent in the behaviorist era, was the question of social or other environmental determinism. Today, challenges are taken to come from particular sciences: physics, genetics, or neurobiology. What all of these have in common is that they are in one way or another opposing *some* concept of human freedom to *some* concept of determinism.

In fact, much of the philosophical literature today speaks of determinism *tout court,* that is, a metaphysical assumption of total causal determinism of present events by events in the past. And if all events are determined by prior causes, then must not human choices themselves be determined by prior causes? Thus, current philosophical literature is structured by the compatibilist-incompatibilist distinction: free will either is or is not compatible with determinism. There are two questions, then: is the causal determinist thesis true? and if so, is free will possible? Compatibilists say that determinism may well be true, but it is a conceptual error to suppose that this rules out free will. Libertarians say that free will is incompatible with determinism, but that determinism does not hold universally. (Timothy O'Connor, in chap. 10 below, will provide a richer account of current positions on free will.)

Galen Strawson, in his article on free will in the *Routledge Encyclopedia of Philosophy*, sees little chance of progress in settling this issue: "The principal positions in the traditional metaphysical debate are clear. No radically new option is likely to emerge after millennia of debate" (Strawson 1998, 3:749). Similarly Louis Pojman concludes a particularly lucid overview of the problem of determinism and free will with a confession of ignorance: "I do not know the answer to this enigma.… [a] paradox which has, since the dawn of reflective thought, perplexed the very best minds" (Pojman 1987, p. 416).

There may nonetheless be a way forward in this debate by recognizing the particular origins of the universal determinist thesis, and then by asking whether it is still justified. It was a reasonable worry in light of early modern physics. The success of Newtonian physics led Laplace to formulate a determinist worldview, which entailed that the movements of human bodies were also governed by the laws of physics. This worldview, of course, was called into question by the predominance of indeterministic interpretations of quantum physics. Bernard Berofsky writes that contemporary determinists base their position on the assumption that, for each event there is some theory or system of laws such that the occurrence of that event is derivable from those laws together with initial conditions (Berofsky 1995, p. 197). However, much has changed with regard to the concept of *laws of nature* during the modern period. The concept began as a metaphor: God has laws for human behavior and for nonhuman nature. While it was thought that nature always obeyed God's laws, many presumed that God could change or override his own laws. By Laplace's day the laws of nature were thought to be necessary. But today, with multiple-universe cosmologies and reflection on the anthropic issue (why does the universe have laws and constants, from within a vast range of possibilities, that belong to a *very* small set that permit the evolution of life?), there is much room, again, to imagine that the laws of our universe are contingent.

Jeremy Butterfield argues that the only clear sense to be made of determinist theses is to ask whether significant scientific theories are deterministic. This is more difficult than it first appears, however. It may appear that the determinism of a set of equations is simply the mathematical necessity in their transformations and use in predictions of future states of the system. One problem, though, is that "there are many examples of a set of differential equations which can be interpreted as a deterministic theory, or as an indeterminate theory, depending on the notion of state used to interpret the equations" (Butterfield 1998, p. 38). Second, he points out, even if a theory is deterministic, no theories apply to actual systems *in* the universe because no system can be suitably isolated from its environment. The only way around this problem would be to take the whole universe as the system in question. If the idea of a theory that describes the relevant (essential, intrinsic) properties of the state of the entire universe and allows for calculation of all future states is even coherent, it is wildly speculative.

These considerations make it appear that progress with regard to "the problem of free will" is more likely to come from examining the implications of particular developments in current science. This volume will focus largely on perceived threats from neuroscience and psychology. However, despite recent developments in physics, many still take the presumed causal closure of physics to pose a threat to free will. This threat is addressed here not by questioning determinism at some levels of physics, but instead by calling into question the modern assumption of reductionism, that is, the view that in the hierarchy of complex systems, all causation is bottom-up and ultimately, therefore, from the level of physics. To this issue I now turn.

1.2 Reductionism, Emergence, and Downward Causation

1.2.1 Reductionism

There have been a variety of interrelated reductionist programs promoted in modern science and philosophy. One is methodological reductionism, the view that the proper way to do science is to analyze or decompose an entity or system into its parts, and then to study the behavior of the parts. The enormous success of this approach to science inspired other reductionist theses. There is epistemological or theoretical reductionism, which is the assumption that the laws or theories of higher-level sciences can and should be reduced to the next level below, and ultimately to physics. This was the goal of many twentieth-century philosophers and scientists. Carl Hempel and Ernst Nagel worked out the most elegant theories regarding the nature of scientific explanation. The phenomena of any scientific field should be deducible from strict, deterministic scientific laws and theories (Hempel 1965). And ideally higher-level theories would be explained by reducing them to lower-level theories (Nagel 1961). This sort of reduction has not turned out to be possible in more than a few instances.

Philosophers have defined ontological reductionism as the thesis that higher-level entities are *nothing but* the sum of their parts. However, this thesis is ambiguous; it can describe two distinct positions. One is the view that as one goes up the hierarchy of levels, no new kinds of nonphysical "ingredients" need to be added to produce higher-level entities from lower. No vital force or entelechy must be added to get living beings from nonliving materials; no immaterial mind or soul is needed to get consciousness. A much stronger thesis is that only the entities at the lowest level are *really* real; higher-level entities – molecules, cells, organisms – are only composite structures (temporary aggregates) made of atoms.

This stronger form of ontological reductionism was combined with other assumptions of early modern physics that together entailed causal reductionism: the thesis that all causation ultimately derives from the behavior of the atoms alone and thus, in the hierarchy of complex systems, all causation is "bottom-up." An additional assumption, which early modern physicists derived from Epicurean atomism, is that the atoms are not affected by their interactions with one another or by the composites of which they are a part. By analogy it was then assumed that, in all higher-level systems, the parts unilaterally determine the behavior of the whole, and are not affected by their relations to one another or to the whole.

1.2.2 Emergence and Downward Causation

The most significant criticisms of causal reductionism fall into three stages: an early emergentist movement (from approximately 1920–1950); the exploration of the concept of downward causation or whole-part constraint (beginning in the

1970s); and, currently, an account of causation that combines both downward causation and emergence.

The idea of emergence was proposed in the philosophy of biology as an alternative both to mechanist-reductionist accounts of the origin of life and to vitalism. The vitalists claimed that in order to get life from inorganic matter something like a vital force needed to be involved. Emergentists, such as Roy Wood Sellars, argued that the increasingly complex organization, as one ascends the hierarchy of systems, accounts for the appearance of new kinds of entities with causal powers that cannot be reduced to physics. The organic emerges from the physical; so too do the levels of the mental or conscious, the social, the ethical, and the religious or spiritual.

Sellars claimed that reductive materialism overemphasizes “stuff” in contrast to organization. Wholes are not mere aggregates of elementary particles. The concept of matter needs to be supplemented by concepts of *integration, pattern*, and *function* (Sellars 1970). With hindsight we can see that Sellars and some of the other emergentists were exactly right; however, their arguments did not prevail against the reductionist philosophers of science.

In the 1970s psychologist Roger Sperry and philosopher Donald Campbell both wrote specifically about downward (or top-down) causation. On some occasions Sperry wrote of the properties of the higher-level entity or system *overpowering* the causal forces of the component entities, which rightly raised worries regarding the compatibility of his account with adequate respect for the basic sciences.

In Donald Campbell’s work there is no talk of overpowering lower-level causal processes, but instead a nonmysterious account of a larger system of causal factors having a *selective* effect on lower-level entities and processes. His example is the role of natural selection in producing the remarkably efficient jaw structures of worker termites. He argues that all processes at the higher levels are restrained by and act in conformity to the laws of lower levels, including the levels of subatomic physics; the achievements at higher levels require for their implementation specific lower-level mechanisms and processes. Explanation is not complete until these micromechanisms have been specified. However, biological evolution encounters laws, operating as selective systems, which are not described by the laws of physics and inorganic chemistry. Downward causation occurs when natural selection operates through life and death at a higher level of organization; the laws of the higher-level selective system determine in part the distribution of lower-level events and substances. Description of an intermediate-level phenomenon is not completed by describing its possibility and implementation in lower-level terms; its presence, prevalence, or distribution will often require reference to laws at a higher level of organization as well (Campbell 1974).

While the concept of downward causation has been used extensively in the sciences in the past generation, it appears that little was written on it by philosophers until the idea was taken up in philosophy of mind in the 1990s. Robert Van Gulick made an important contribution by spelling out in more detail an account based on selection. The reductionist’s thesis is that the causal roles associated with the classifications employed by higher-level sciences are entirely derivative from the causal roles of the underlying physical constituents. Van Gulick argues that even though the events and objects picked out by higher-level sciences *are* composites of physical constituents, the causal powers of such an object are not determined solely by the physical properties of its constituents and the laws of physics. They are also determined by the *organization* of those constituents within the composite. And it is just such patterns of organization that are picked out by the predicates of the higher-level sciences.

These patterns have downward causal efficacy in that they can affect which causal powers of their constituents are activated. "A given physical constituent may have many causal powers, but only some subsets of them will be active in a given situation. The larger context (i.e. the pattern) of which it is a part may affect which of its causal powers get activated.… Thus the whole is not any simple function of its parts, since the whole at least partially determines what contributions are made by its parts" (Van Gulick 1995, p. 251).

Such patterns or entities are stable features of the world, often in spite of variations or exchanges in their underlying physical constituents. Many such patterns are self-sustaining or self-reproducing in the face of perturbing physical forces that might degrade or destroy them. Finally, the selective activation of the causal powers of such a pattern's parts may in many cases contribute to the maintenance and preservation of the pattern itself. Taken together, he says, these points illustrate that "higher-order patterns can have a degree of independence from their underlying physical realizations and can exert what might be called downward causal influences without requiring any objectionable form of emergentism by which higher-order properties would alter the underlying laws of physics. Higher-order properties act by the *selective activation* of physical powers and not by their *alteration*" (Van Gulick 1995, p. 252).

Van Gulick has also helpfully related the variety of current emergentist theses to anti-reductionist theses. Given that causal reductionism is the concern of this book, it is appropriate to consider the best current work on the emergence of new *causal* capacities as one ascends the hierarchy of the sciences. I believe that the best account so far is that developed by Terrence Deacon (see Deacon 2007).

Deacon distinguishes three types or levels of emergence. There is no emergence in mere aggregates, though an aggregate does have some sorts of global properties. For example, the weight of a volume of liquid is a simple addition of the weights of its molecules. The important difference between an aggregate and a system is that in a system it is *relational* properties of the constituents (as opposed to primary or intrinsic properties) that constitute the higher order. In such cases additional configurational and distributional information is needed to account for the higher-order properties. Deacon includes here the viscosity of liquids, turbulence in large bodies of water, and typical feedback systems such as a thermostatically controlled heating system. This he calls first-order emergence. In Juarrero's terms (chapter 5 below), the relations among components impose constraints on the system. Because fluctuations in such systems are dampened out across time it is possible to give (rough) reductionistic accounts of their behavior.

Second-order emergence occurs when there is symmetry breaking or the *amplification* of a fluctuation rather than dampening. Systems in which this occurs are nonlinear; their history matters. There are simpler and more complex versions of such systems. The simpler sort is self-organizing, in that higher-order patterns selectively constrain the incorporation of lower-order constituents into the system or select among possible states of the lower-level entities (this is Van Gulick's point, as well). More complex second-order emergent systems are also autopoietic: they change the lower-order constituents themselves. Examples of the simpler sort are the Bénard phenomenon (the development of orderly convection rolls in a heated liquid), a thermostat that amplifies rather than dampens feedback, and the development of a snowflake. An autocatalytic cycle is of the more complex sort, in that the system manufactures some of its own components. All life involves second-order emergence of the more complex sort.

Deacon distinguishes between first- and second-order (as well as third-order) emergence in terms of what he calls "amplification logic" or "the topology" of causal processes. In systems without emergence, global properties are all produced bottom-up (or by means of *local* interactions with boundaries – e.g., a water molecule constrained by the surface of the container). In first-order emergent systems there is "nonrecurrent" causal architecture: a simple bottom-up and top-down relation in which global properties of the system (e.g., density of components) makes a difference to the relations among components and thus to the behavior of the whole system.

Second-order systems have more "tangled" or "recurrent" causal architecture as a result of the amplification of lower-level fluctuations. This amplification changes the total state of the system in a way that makes a decisive difference for the future development of the system. This can lead to new orders of complexity. Deacon's second-order emergent systems are the simplest of those that Juarrero describes as being driven by context-sensitive constraints: what happens before changes the probabilities for future behavior of the components.

Third-order emergence involves the interaction among three levels and appears (naturally) only in the biological realm. Here a variety of second-order forms emerge, and are selected (constrained) by the environment, but in such a way that a *representation* of its form is introduced into the next generation. The simplest example is the evolutionary process. The micro-level (the genome) in interaction with the organism's environment, directs the construction of the organism (the mid-level), whose reproductive fate is determined top-down by the environment (top level). The preservation of information regarding the organism's success in the environment is the means by which a relatively stable population of successful organisms can be produced, within which future fluctuations appear. Some of these may be amplified (preserved and reentered into the system) by means of interaction with the environment, thus enabling the appearance of still higher degrees of complexity. Deacon describes such systems as exhibiting recurrent-recurrent causal architecture: over time, a two-stage process of emergence occurs that results in downward causation not just from top to mid-level, but from top to bottom (environment to genome).

From Van Gulick's and Deacon's accounts we can see that evading causal reductionism requires the recognition that higher-level entities and systems have emerged (evolved) from lower, and that these entities can be somewhat independent of the causal processes of their constituents, thereby manifesting new, higher-level causal capacities. The sort of organization and selection of lower-level causal processes that Van Gulick describes calls for new concepts, and, in fact, represents something like a paradigm change across the sciences. This is the shift from thinking in mechanistic terms to thinking in "systems" terms. This is the point of departure for essays in Part One of the present volume.

1.3 Is Conscious Will an Illusion?

There are two areas of research that have stimulated the authors in this volume. One is that of Benjamin Libet, the other is Daniel Wegner's. Libet's research will be described only briefly here, since there are reports on this topic throughout the book. Libet's research began with the finding by Kornhuber and Deecke that the performance of "self-paced voluntary actions" was preceded by a slow electrical change, recorded on the scalp, called the readiness potential (RP) or *Bereitschaftspotential* (Kornhuber & Deecke 1965). Libet devised a method for measuring the relations among the RP, subjective feelings of volition, and action. Subjects in his studies were told to flick their wrists "at any time they felt the urge or wish to do so." These acts were to be performed "capriciously, free of any external limitations or restrictions" (Libet 1999, p. 49). He used an EEG to measure the RP and an EMG to record muscle movements. He asked the subjects to report when they were first aware of the wish or urge to act.

Averaging across numerous trials, Libet found that the RP preceded the action by approximately 550 milliseconds, and the wish to act occurred approximately 200 milliseconds before the muscle movement. The significance, according to Libet, is that the volitional process is initiated unconsciously, whereas in the traditional view of conscious will, "one would expect the conscious intention to appear before, or at the onset of, the RP, and thus command the brain to perform the intended act" (Libet 1999, p. 49).

Libet recognizes that his research appears to have negative implications regarding free will, but has done further research showing that subjects can veto the action after feeling the urge to act. He locates free will in this veto power.

Wegner's research is reported in his book *The Illusion of Conscious Will* (Wegner 2002). He distinguishes two ways of talking about conscious will: as a *feeling* of voluntariness or of doing something on purpose, and as "a force of mind, a name for the causal link between our minds and our actions" (Wegner 2002, p. 3). He draws from a variety of sources to show that the feeling of conscious will does not always correlate with will in the second sense. Thus, the feeling cannot be a veridical perception of that which causes the action, and we need an alternative theory of where this feeling comes from. He proposes that we come to attribute causal agency to our thoughts in the same way we attribute causality in other domains. When A regularly precedes B and there is no other apparent cause, we take

A to cause B. We also have a tendency to project intentions and agency on both animate and inanimate beings. In the case of perceiving our own agency there is the additional factor that the thought or intention is consistent with the act. “For the perception of apparent mental causation, the thought should occur before the action, be consistent with the action, and not be accompanied by other potential causes” (Wegner 2002, p. 69).

Wegner brings together reports of research on the role of consciousness and other factors in behavior with surveys of some of the stranger phenomena from the history of the human race, all supporting his contention that the feeling of conscious will does not always correlate with the true causes of action. One category of evidence is cases where it is highly likely that people are in fact the causes of their own actions, but they experience the acts as being controlled by some other source. These include instances of automatic writing, spirit possession and mediumship, table-turning, and so on.

Wegner’s second sort of cases involve the feeling of will when causation is absent. These include Libet’s research on readiness potential, as well as research with subjects whose right and left hemispheres have been severed. When such patients are prompted to act on the basis of information presented only to the right hemisphere, and then are asked why they did it, the left (verbal) hemisphere quickly makes up a reason.

We are led to expect by the title of Wegner’s book that it will show conscious agency to be an illusion, but in fact the book is, in the first instance, about something else – the *feeling* of conscious agency. This is an important topic in itself: it is important to recognize that the feeling can be distinguished from the real thing and studied productively by psychologists and neuroscientists. But what, if any, implications does this have for the age-old philosophical question about human responsibility – in Wegner’s words, “do we consciously cause our actions, or do they happen to us?” This question is explored by a number of authors in this volume.

2 Overview of the Volume

In this section I shall sketch the contents of each of the chapters, noting some of the relations among them, but shall save for the final section a brief synthetic account of the conclusions of the book.

Christof Koch, in chapter 2, “Free Will, Physics, Biology, and the Brain,” sets up the problems to which this volume responds. Are we humans conceited in believing that we alone (or perhaps with other higher animals) can escape the iron law of cause and effect? His focus is on libertarian free will, which he defines in terms of having been able, in exactly the same circumstances, to have chosen differently than one did. Without some degree of such freedom, he says, cherished beliefs, institutions, and cultural practices are in jeopardy, particularly our assignment of moral and legal responsibility.

While Koch recognizes the variety of factors that are seen to limit our choices – prior actions of our own, family history, cultural context, genetics, and neurobiology – much of the focus of his chapter is on physics. He traces the changes from the Newtonian-Laplacian clockwork image of the universe through the variety of developments that have shown the impossibility of predicting much of the future, including Henri Poincaré's recognition of what, since Edward Lorenz, has been called deterministic chaos. There is also the "noise" present at the molecular level that comes from jostling motion. None of this, however, defeats the claim that physics determines all events.

Koch then turns to the question of whether genuine indeterminacy at the quantum level has any significance for free will. He considers but does not endorse the views of Roger Penrose and others who link consciousness in one way or another to quantum effects. He considers the possibility that quantum fluctuations in the brain could be amplified by deterministic chaos and thereby lead to behavioral choices. This would mean that some choices were not predictable, but he concludes that this would not be what is wanted by way of free will. He considers, also, the argument made by John Eccles and Karl Popper to the effect that a nonmaterial mind could determine the outcome of otherwise indeterminate quantum events in the brain. This would fit the definition of libertarian free will if it were workable, but it founders for lack of an account of *how* it could work – the problem unsolved since Descartes's day of mind-body interaction.

Koch concludes that there is ample evidence that nervous systems display noise, random activity, and that even simple organisms such as fruit flies behave in spontaneous, unpredictable ways, but notes that the majority of neuroscientists do not believe that quantum indeterminacy is relevant here.

Koch's chapter also introduces the aforementioned neurological and psychological studies that are currently seen to cast doubt on free will: Libet's measurement of the readiness potential occurring prior to subjects' conscious intention or urge to act, and Wegner's showing that people sometimes have the experience of agency when they are not in fact acting, and sometimes lack it when they do act. Koch concludes from this body of research that the conscious mind does not cause the action; it is more like a marker for voluntary action, "an afterthought." The actual workings of the sense of agency – why we choose as we do – is opaque, hidden from conscious access. There is a fundamental mystery of how bio-electrical activity in a restricted part of the brain gives rise to these experiences of agency. He raises the question of what are the neuronal correlates of willful conscious experience. These questions regarding the locus of the experienced sense of agency, as well as the sources of movement itself, will be addressed below by Mark Hallett and Sarah-Jayne Blakemore.

William Newsome, in "Human Freedom and 'Emergence,'" also states clearly the problems addressed in this volume: how can we reconcile the causal character of our scientific worldview with traditional belief in free will; and if we cannot, what then becomes of the presuppositions of our legal system? He notes that most religious traditions also presuppose free will. In addition, Newsome argues that, whatever the difficulties, scientific conclusions cannot be taken to undermine free will, since the practice of science itself *also* presupposes that scientific judgments are more than the inevitable outcome of atomic, molecular, and cellular interactions in the brain.

Newsome agrees with Koch that solution to the free-will problem will not be found in quantum indeterminacy; rather than looking to the bottom of the hierarchy of complexity, Newsome begins our presentation of resources based on the concepts of *emergence* and *downward causation,* which endow complex systems with a degree of behavioral autonomy. In organisms, he claims, this autonomy can be regarded as meaningful choice.

Newsome notes that unicellular organisms are often cited as examples of emergent systems since they exhibit an enormous number of phenomena that go well beyond the capacities of their parts. The complexity of even such simple organisms, however, is so great as to make it impossible to show *how* it is that their behavior *fails* to violate the laws of physics and chemistry. Therefore he turns to the much simpler example of artificial neural networks. In one sense we know everything there is to know about how they work. Multiple layers of computing units are linked hierarchically so that the behavior of lower levels influences each unit in the layer above. The strengths of influences are governed by "weights," which, by means of a backpropagation network, are gradually adjusted to produce the desired output at the top. Yet in another sense the programmer usually does *not* know how the system works, in that it is not possible to tell *how* the problem was solved.

Newsome proposes these networks as a model of an emergent system – it can do remarkable things that its components cannot. And we know that it does so without any causal gap at the bottom. The relevance of this "toy" example is in showing that a complex system with the ability to learn possesses the autonomy to discover solutions to problems that cannot be derived from lower-level descriptions. The key feature is that the information embedded in higher organizational levels is the most important locus of control of the system.

This model suggests that brains will not be understood in terms of their components, because at certain levels of functioning the primary drivers of the system will be the logical rules that apply to the higher levels of the system. In the case of humans, this includes symbolic reasoning and, especially, the ability to reason recursively about our own reasoning, in interaction with our environment.

George Ellis, in his chapter titled "Top-Down Causation and the Human Brain," presupposes Newsome's account of emergent complex systems and expands on his explication of the role of downward causation. The possibility of downward causation depends on the fact that the hierarchy of complexity is made up of whole-part relations. True complexity at higher levels depends on modularity: the system will be composed of quasi-independent modules interacting with one another in a network, allowing for "encapsulation," that is, information hiding, abstraction, and inheritance. Another term Ellis uses is "coarse graining": details of the lower-level components of the modules become irrelevant for higher-level functioning. The car mechanic does not need to know the physics of the metals with which he works. In addition to the properties of the modules themselves, it is the set of relations, particularly functional relations, that are crucial for creating the complex system.

That downward causation takes place can be shown by changing higher-level variables and determining that lower-level variables then change in a reliable way. Top-down causation is ubiquitous in physics, chemistry, and biology because the outcome of lower-level interactions is always determined by context. Ellis describes five types of top-down causation, showing in each case *how* the higher-level variables affect the lower. He begins with the simplest: algorithmic top-down causation. This occurs when high-level variables have causal control of lower-level dynamics through the structuring of a system such that the outcome of a process depends on the higher-level structural, boundary, and initial conditions. An example is algorithmic computational procedures in a digital computer on the basis of initial data. The algorithms (stored in high-level programs) determine the machine code that then determines the low-level switching of transmitters.

Ellis's second level of downward causation is via nonadaptive information control. An example here is a thermostatically controlled heating system. The behavior of the components is determined by the higher-level goal via a feedback system. This is downward causation because the goals are only expressible in terms of the system as a whole, and cannot be expressed in terms of the characteristics of the lower-level entities that make up the system.

The third level is adaptive selection, in which entities at the lower level display variation, and those that are better suited to their environment are selected and survive, while other variants disappear. This corresponds to Donald Campbell's example of the top-down processes resulting in the effective jaw structure of worker termites and ants. This can be thought of as a generalized feedback loop with a meta-purpose, because, unlike the nonadaptive feedback-control system, the selection criteria can develop over time to adapt to new contexts. One of Ellis's examples is the training of artificial neural networks, the illustration central to Newsome's chapter. Another example is the process of adaptive selection called neural Darwinism. Neural connections are formed and tuned on the basis of higher-level fitness criteria that guide the brain in response to environmental interactions.

The fourth level of top-down causation Ellis calls adaptive informational control. Here there is not merely evolution of the goals that govern selection, but a system capable of learning and of anticipating future outcomes. This, in turn, allows for switching of the goals themselves. It is exhibited in animals that are capable of switching, for example, from the pursuit of a drink of water to running from a predator. Note that here Ellis is describing the causal interactions in what Deacon calls a third-order emergent system.

Finally, only humans display intelligent top-down causation. This is enabled by symbolic representation, allowing for conscious selection of goals, many based on abstract entities such as theories and values. Language allows information to be stored and selected for relevance, for precise prediction of outcomes of complex actions. Ellis notes that while we do not yet fully understand the neural processes involved, we know that this sort of top-down agency must be taking place or else science itself would be impossible. He notes also the importance of social roles and other cultural resources, especially the values that provide for ethics, aesthetics, and meaning.

Ellis considers what conditions are necessary for downward causation without causal overdetermination at the lower levels of the hierarchy. He says that we observe “causal slack” in lower-level systems when they are open systems. He makes Van Gulick’s point that downward causation acts not by overpowering lower-level processes but by selecting among lower-level entities and processes, but argues that the indeterminacy at the micro-level is also a necessary condition. Ellis ends by noting the extent to which the complexity of causal processes now recognized in science tends to draw us toward something like Aristotle’s fourfold account of causation. Besides the efficient cause, we also need to consider the materials involved, the structures of complex systems, and the effects of systems with goals (teleology).

Ellis has pointed out that systems in which quasi-independent modules interact are partially decoupled from lower-level causal processes. The more complex such systems, the more they come to have their behavior determined by variables pertaining to the system level. **Alicia Juarrero**’s chapter, “Top-Down Causation and Autonomy in Complex Systems,” pursues the means by which systems achieve greater degrees of autonomy from fundamental energetic forces. Her focus is on complex dynamical systems. These differ from aggregates in that mere aggregation does not affect the character of the components. These systems are open and far from equilibrium. They display a unique balance of integration, cohesion, and robustness at the global level, and at the same time, differentiation and multiple realizability at the component level.

Juarrero attributes the cohesion of complex dynamical systems to “context-sensitive constraints.” Her distinction can be illustrated by the difference between throwing a die and playing a card game. Previous throws (context) have no bearing on the outcome of the next throw. In contrast, in a card game the sequence of cards already dealt does constrain the probability of a particular card being dealt next. Context-sensitive constraints create higher degrees of order by making elements of the system interact in such a way that their behavior is dependent on one another’s and on what went on before. Once the probability that event B will happen is altered by the presence or interaction with A, the two have become systematically and therefore internally related. When this happens a global structure, AB, has emerged, defined by conditional probabilities. These constraints integrate previously independent parts into a unified whole that incorporates the record of its history, is embedded in its environment, and possesses emergent properties. Context-sensitive constraints exist in metabolism, language, neurophysiology, and chemistry.

One of the simplest examples of such a system is the Bénard phenomenon. As liquid is heated, the uncoordinated movements of molecules suddenly shift to an ordered pattern of convection rolls. It begins with the amplification of a fluctuation (cf. Deacon’s second-order emergence) and persists because, once each water molecule is captured in the dynamics of the cells, it is no longer related only externally to the other molecules. Its behavior is constrained by the global structure into which it is caught up. It is no longer the intrinsic properties of the molecule that matter, it is its relations to the other molecules.

Juarrero argues that in the course of evolution we see the development of systems in which the higher level becomes increasingly autonomous and self-directed as its capacity for constraint, modulation, and regulation is brought further and further inside, modularized, and additionally decoupled from energetic exchanges (cf. Ellis). Since levels are screened off from each other, new levels of dynamical organization involve the appearance of new capabilities at the top level, and an enlarged phase space with more degrees of freedom than the sum of its constituents'.

Living systems are autopoietic, that is, they construct themselves by creating the constraints that control the matter-energy flows that make the self-organization possible. The simplest of autopoietic systems are autocatalytic cycles, in which the process selects the molecules that participate in its continued coherence. Chemical complexity creates and preserves itself through a natural selection process whose fitness criterion is the persistence of the whole. In so doing, the system can affect its own environment – altering the chemical concentrations outside. This selection according to the goal of the system is an instance of Ellis's downward causation via adaptive selection. An autopoietic system thus exhibits greater self-determination than a dissipative system.

The development beyond the level of chemistry requires the emergence of "dynamical decoupling." By this Juarrero means the production of a new type of functional component, such as DNA, that serves as a record of earlier functions and guarantees replication, while other components carry out metabolic regulation (cf. Deacon's third-order emergence).

The final step toward autonomy occurred with the appearance of the frontal cortex – another means of recording the history of the system in order to guide its future development. Consciousness, self-consciousness, and symbolic language allow humans to possess a higher degree of autonomy from their environment and from energetic forces. Juarrero takes this maximal autonomy to constitute free will. Functional, informational, symbolic, and representational processes operate as formal (not efficient) causes providing second-order context-sensitive constraints that temporally span the onset and terminus of behaviors. This is how conscious intentions can operate as structural causes of meaningful human actions.[1]

Both Ellis and Juarrero have pointed out that complex systems depend on the coupling among relatively autonomous modules. Juarrero claims that context-sensitive constraints represent couplings that are "Goldilocks-like": not too tight, not too loose, so as to allow the same microstructure to participate in different complex dynamics, both synchronically and diachronically. **Scott Kelso** and **Emmanuelle Tognoli**, in their chapter "Toward a Complementary Neuroscience: Metastable Coordination Dynamics of the Brain," pursue this issue. They focus on

[1] The capacity, created by symbolic language, to evaluate records of past behaviors and their consequences and to formulate representations of future behavior is such a leap beyond mere records of the past that I believe we should describe it as fourth-level emergence.

the nature of the interplay between the whole and the parts as expressed through the concept of coordination dynamics, a form of coupling among relatively autonomous modules "which reconciles the well-known tendency of brain regions to express their autonomy with the tendency of those regions to work as a synergy," the first feature being the bottom-up aspect of the dynamics (based in the internal structure of local modules) and the latter the top-down aspect (based in the links between the modules). The proposed mechanism whereby this happens is through coordination between nonlinear coupled oscillators, which is a form of binding between discrete dynamical units, thereby forming a temporary larger emergent entity. This enables a complementarity between larger wholes and their constituent parts. In tightly coupled cases, the brain is locked into one or other such emergent higher-level state, characterized by the phase relations between its parts; in loosely coupled cases, the different parts operate more or less independently, so top-down causation is minimal. Intermediate between these cases are metastable states where the coupling is neither too tight nor too loose, so that shifts between temporary dynamical bindings can occur, such as phase transitions in physics, but here corresponding to a change in the state of the mind. This is potentially related to the way decisions are made in the brain in a context-sensitive manner.

The chapter focuses on the nature of such metastable states in brain functioning, where they allow a flexibility of response that can be adaptive in nature ("instability in this view is a selection mechanism picking out the most suitable brain state for the circumstances at hand"). Thus, this is a specific mechanism whereby the kinds of dynamics discussed by Ellis and Juarrero can be realized: "a delicate balance between integration (coordination between different areas) and segregation (expression of individual behavior) is achieved in the metastable regime." A useful aspect of this chapter is its illustration of how simplified quantitative models can illuminate the nature of dynamical behavior, even in systems as complex as the brain.

Part Two of the book considers in detail the experiments of Libet and Wegner, with their possible threatening implications for free will, and also gives surveys of what is known about the neural correlates of voluntary movement.

In chapter 7, "Physiology of Volition," **Mark Hallett** helpfully distinguishes between the brain systems that are likely to be involved in the actual initiation of movement and those involved in the conscious sense of our own agency. In the latter case, there needs to be the sense of willing the action to occur, properly related with the perception that the movement took place. Insight into this process comes in part from studies of subjects with neurological disorders, such that either the process of movement initiation itself is aberrant or the linkage between movement generation and perception of agency is faulty. There are cases of involuntary movement in which the patient believes that the movement was voluntary, and there are also cases in which the ability to initiate action is lost.

Hallett develops a model to represent the normal relations among volition, action, and perception of agency, along with suggested neural correlates. Movement begins with motivation, and this leads to planning of a movement. While he has concluded that there is no evidence identified for free will as a force in the generation of movement, he points out that it may be misleading to look for *an* initial event of willing since the brain is always working and providing actions; thus the relevant question is why the particular action that occurred was selected. When selected, the action can be executed. The perceptual component is alerted to upcoming movement from both planning and execution modules by feedforward signals. The sense of agency is generated by a match between volition and movement feedback.

Motivation is associated with limbic and prefrontal regions of the brain. Studies show that selection of the action to perform depends on different regions, depending on whether the choice is which action to choose (supplementary motor area [SMA]) or when to move (dorsolateral prefrontal cortex [DLFPC]). It is likely that movement is initiated in mesial motor areas and premotor cortex. The movement command then goes to primary motor cortex. Corollary discharges appear to come from the SMA and dorsal premotor cortex (PMd) to parietal areas, and these may be responsible for the sense of volition. Parietal and frontal areas maintain a relatively constant bidirectional communication. It is likely that this network of structures includes the insula. The sense of agency comes from the appropriate match of volition and movement feedback, likely also centered in the parietal area.

Hallett describes a number of experiments following up on Libet's research, and concludes that the phenomena he has identified are well supported. He also considers both implications and criticisms of this body of research. One issue is the question of when the decision to move one's finger was actually made. One might argue that the *relevant* (free) decision occurred when the subject initially agreed to participate in the study. Another issue is the nature of subjective perception of time and events, and the relation of that perception to events in real time. Subjective timing of events that are felt to occur prior to the movement may be influenced by the movement itself.

Chapter 8, "How We Recognize Our Own Actions," by **Sarah-Jayne Blakemore** further explores the role of the relation between perception of our own actions and the sense of agency. One way in which the brain predicts the consequences of movement is by means of a "forward model" that uses the efference copy of motor commands to predict the sensory consequences of a movement. With little or no discrepancy between the predicted and actual sensory consequences, the movement is classified as self-produced.

One significant factor is the time between the movement and the sensory stimuli. If the experimental situation is organized so that the sensory stimulation is delayed by 100 to 300 milliseconds after the (presumed) action, there is a decreased sense of agency. A series of experiments suggests that the cerebellum is involved in signaling the discrepancy between predicted and actual consequences of movements.

Damage to the parietal lobe is associated with loss of control and awareness of action, for example, in confusing one's own hand movements with those of another agent. Thus, the parietal lobe is hypothesized to be involved in both maintaining and updating the internal bodily states that issue from sensory and motor signals.

Blakemore's forward model of association of action with intention may explain the experience of many schizophrenia patients who mistake actions, thoughts, and emotions of others for their own. It may be that the forward prediction does not reach awareness in these patients. This model also explains some aspects of phantom-limb phenomena: the estimation of the position of the limb is not based solely on sensory information, but also on the stream of motor commands issued to the limb muscles.

Hakwan Lau, in "Volition and the Function of Consciousness," uses Libet's and others' research to raise the question of the role of consciousness in enabling various forms of behavior. It turns out that the best way to investigate this issue may be to consider cases in which consciousness is absent. It is ordinarily assumed that many voluntary actions require conscious effort; without consciousness we would only be able to perform simple actions akin to reflexes. It turns out, however, that there are complex acts that can be performed without what would have been thought to be essential conscious information; instead they can be performed on the basis of unconscious information.

Lau considers Libet's research, which appears to show that the experimental behavior is initiated prior to consciousness; he notes that Libet's own solution, the fact that one can veto the act before it occurs, does not in fact solve the problem of the role of consciousness, since a variety of studies have shown that subjects' perception of the time of conscious urge or intention[2] is often biased so as to appear earlier. That is, the urge *appears* to the subject to have occurred farther in advance of the action than records of brain waves determine (cf. Hallett). This calls into question whether the subjects in fact have enough time to consider the veto. Lau agrees with Wegner that our awareness of intention may be constructed after the fact, and its timing may be manipulated by contextual factors.

Lau next considers situations in which conscious deliberation seems to be needed to avoid certain types of action, for example, completing a word that begins with the letter "d" but avoiding the word "dinner." A peculiar result is that if the excluded word is presented subliminally, subjects tend to produce it with a higher frequency than chance. This indicates that they have in fact received the information but are not conscious of it. These and other studies indicate that inhibition of action indeed requires consciousness. However, here is where the methodological challenge arises. The studies are designed to "knock out" conscious awareness of the relevant stimuli, but they are confounded by the fact that conscious stimuli are stronger and longer-lasting stimuli. So the difference in performance may not be specifically due to the lack of conscious awareness, but rather merely to the difference between weak and strong signals. A potential explanation of the word-exclusion experiment, then, could be that when the excluded word is masked, the signal is too weak for use in conscious control of behavior, but strong enough to have a priming effect.

So is it the case that a more complex task involving top-down cognitive control requires consciousness? An example of such control is our ability to inhibit a

[2] O'Connor (chap. 10) will point out the problematic consequences of using these terms interchangeably.

typical behavioral response (answering the phone) under specific circumstances (being a guest in someone's home). Experimental results here are also ambiguous. The experimental situation requires subjects to perform different tasks (judge whether a word is one or two syllables, versus whether it is concrete or abstract) depending on one of two prior visual cues. If the opposite cue is presented below the conscious threshold before the visible cue, this impairs performance. So it appears that unconscious information can influence more complex cognitive tasks as well. Lau concludes that future studies need to be designed to distinguish between the effects of consciousness per se and signal strength.

The chapters in **Part Three** of the book respond to the research described in Part Two, in various ways calling into question the relevance of Libet's and Wegner's studies to the topic of free will, or interpreting them within the broader context of human behavior and experience.

Timothy O'Connor begins chapter 10, "Conscious Willing and the Emerging Sciences of Brain and Behavior," with an overview of philosophical positions on the nature of free will. He then points out a number of conceptual confusions that tend to support the cases of those he calls the free-will skeptics. In his overview of the empirical findings used by the skeptics he includes (1) confabulation – the tendency of brain surgery patients and research subjects to claim that they had reasons for movements that were clearly caused by external agents; (2) Libet's and colleagues' research; (3) clinical disorders involving misattribution of agency; and (4) Wegner's psychological studies.

To diagnose some of the conceptual confusions he detects, O'Connor distinguishes seven distinct concepts related to agency: (1) minimally voluntary action, which corresponds with one's desires or intentions but unfolds automatically; (2) consciously forming an intention to act, either immediately or later; (3) feeling an urge or desire to perform an act; (4) beliefs concerning one's own actions; (5) beliefs concerning the causal impact of one's basic actions; (6) the experience of willing an action; and (7) a general sense of authorship, a general and persistent sense of being the owner of one's actions. Using these distinctions it is possible to show that some of the empirical findings do not in fact pose a threat to free will.

O'Connor reinterprets instances of confabulation not as illusory experiences of will but as unremarkable instances of our occasional penchant for forming false memories in order to produce coherence with others' expectations. What is termed a false sense of agency (e.g., causing a person to become ill by thinking negative thoughts) actually falls under the category of holding a false *belief* about the causal effects of one's basic acts. Automatisms of various sorts do serve to show the distinction, already emphasized by Hallett, between the *experience* of willing an action and its actual execution, but the existence of automatic behaviors provides no evidence against free will; we could not survive if we needed to intend and consciously monitor all of our behavior.[3]

[3] R.F. Baumeister and K.L. Sommer have estimated that consciousness plays a causal role in as little as five percent of our daily behavior, implying that 95 percent is automatic (Baumeister & Sommer 1997).

Finally, O'Connor questions the relevance of Libet's research to free will, given the peculiar position of the subjects, who have already (freely?) agreed to a predefined action type (cf. Hallett), but have beentold *not* to preplan the timing of the act, and rather to wait for the urge, desire, wish, intention to act. This sets up the subjects to be passive observers of their own experience, and it should not be surprising if there is unconscious neural activity prior to this anticipated urge or desire.

Having concluded that the research referred to above fails to defeat our assumption of free agency (a necessary assumption, by the way, for understanding oneself to have *chosen* to engage in scientific research relevant to the issue of free will – cf. Newsome), O'Connor argues that the research *does* point to the need for fine-tuning philosophical models of free will. First, philosophers tend to argue for idealized conceptions of free will. The pervasiveness of automaticity shows that the responsibility for much of what we do is at best "inherited" from the few directly free choices that we make. In addition, philosophical concepts of free will need to be adjusted by taking into consideration the varying degrees of consciousness we have of that which moves us to act. And perhaps the most important factor in attributing our actions to our own intentions is the degree to which our motives are the product of our own past choices.[4]

O'Connor has emphasized the importance of our ability to be aware of our desires, beliefs, and total motivational structure in determining the degree of our freedom. **Evan Thompson**, in "Contemplative Neuroscience as an Approach to Volitional Consciousness," focuses precisely on the factors that increase this sort of self-awareness. The study of consciousness by cognitive neuroscientists assumes our ability to report accurately our own experience. Such reports depend on meta-awareness – conscious awareness of our first-order conscious experiences. As Lau and Hallett have noted, self-reports requiring introspection are subject to various biases (shifts in experienced timing of events). Thompson adds that our attention tends to shift rapidly, and we are usually unaware of this attentional instability. In addition, the very process of attending to and reporting on experience tends to change its character or edit its content.

Because the difficulties just noted are likely to confound scientific studies of consciousness, Thompson and others have developed the specialization termed contemplative neuroscience. The rationale is based on the fact that experienced contemplatives have trained themselves to attend to and control their own mental processes. Thus, they provide important subjects for neuroscientific research. Volitional consciousness offers an important test case for such research. There has been little sustained investigation of the phenomenology of volition by neurophenomenologists, that is, scientists who combine first-person phenomenological investigation, second-person phenomenological interviews, and third-person behavioral and neurophysiological measures. Thompson and colleagues employ this method to study Theravada Buddhists, whose practice is particularly relevant to research on volition in that it involves training in the ability to notice intentions

[4] I shall elaborate on these points in sec. 3.

and volitions as they arise and consciously to choose whether to act on them. Without such training the volitions usually lead automatically to action.

As do Juarrero and Kelso, Thompson takes conscious states to be embodied in large-scale dynamical patterns of temporally coordinated neural activity across selective brain regions. Measures of electrical brain activity have found distinctive patterns in advanced meditators compared with novices, not only during meditation but also in a resting state before meditation. This suggests that meditation may induce not only short-term changes in neural activity but long-term changes as well. In addition, the self-reported "clarity" of adepts' meditative states correlated closely with high-amplitude gamma activity in frontal regions.

Thompson argues that these brain patterns and correlated states of consciousness are *emergent* in that they are metastable systems of neural behavior that arise spontaneously, given the local couplings among components and the way those couplings are globally constrained. He interprets volition, as does Kelso, in terms of the person's ability to either stabilize or destabilize such an entangled system, and hypothesizes that contemplative mental training creates new types of global order parameters for the neural coordination dynamics underlying various processes.

In chapter 12, titled "Free Will and Top-Down Control in the Brain," **Chris Frith** adds to the understanding of the ability to control one's own focus of attention. Lau has already introduced the concept of top-down cognitive control; Frith contrasts this with bottom-up control, by which he means acting in accordance with all of the forces that happen to be impinging on the person at the time. He defines free will as top-down control, the ability to act (somewhat) independently of all impinging forces.

Frith takes as his first example the well studied capacity for selective attention, which is hypothesized to be achieved by one of two mechanisms. One is a bottom-up process of free competition among stimuli in which the strongest stimulus wins. The second is a top-down process by which the competition is biased in advance in favor of a particular type of stimulus. The neural processes involved in bottom-up processing depend on the fact that the action of one sensory channel inhibits all of the others so, as signals pass through the central nervous system, stronger channels gain strength and weaker channels are ultimately shut down. By this means, only the strongest signal survives to drive behavior and to reach conscious awareness.

In studying top-down control, subjects are told to pay attention to only a certain type of stimulus. As perceived psychologically, this requires effort to refrain from responding to the nontargeted stimuli. At the physiological level this effort correlates with increased activity in areas associated with targeted stimuli, for example, V4 if instructed to attend to color. Bottom-up and top-down processing relate to feedforward versus feedback connections. Bottom-up processing in the psychological sense always maps onto feedforward connections; top-down processes usually but not always map onto feedback connections, for example, from frontal cortex to sensory regions.

Further insight into the neural underpinning of voluntary action comes from studies such as Libet's, in which the action is prescribed but the subject chooses the time, and from Frith's own research, in which the time is specified but the subject chooses between two possible actions. In both cases the dorsolateral prefrontal cortex and the anterior cingulate cortex are activated. Frith inquires whether these two regions should be thought of as the "top" from which top-down control of action originates. He says that this is not farfetched, since these areas are more developed in humans than in animals, and severe damage here leaves patients "slaves to stimuli."

Frith agrees with Hallett that the source of willed action is not an *ex nihilo* act of will, but rather a choice among alternatives. These actions are presented by stimuli; choice is a matter of "sculpting response space," that is, of inhibiting all but one action. When there remains a conflict between demands for two responses that cannot be carried out simultaneously, the anterior cingulate cortex takes precedence over the dorsolateral prefrontal cortex. Frith notes that, given that the choices involved in the research reported so far are rather trivial it is particularly significant that, in studies of moral responses in economic game playing, these same brain regions also turn out to be involved.

At this point, Frith raises a critique of his own model of top-down control. His diagram has a box at the top labeled "goals/plans," and he has been arguing that this box corresponds with the dorsolateral prefrontal cortex and the anterior cingulate cortex. However, the box has only outputs, while there are in fact no brain regions with outputs but no inputs. This leads Frith outside the brain in his quest for the top of the system. As does O'Connor, Frith recognizes the unnaturally circumscribed setting of the subjects in these experiments on "voluntary" behavior. To truly understand the neural bases of free will we need to understand how social factors exert top-down constraints on the brain, and this in turn requires investigation of how brains allow minds to interact.

Sean Spence pursues the role of social interaction in his chapter titled "Thinking beyond the *Bereitschaftspotential*: Consciousness of Self and Others as a Necessary Condition for Change." This chapter nicely draws together themes from earlier authors in this part of the volume, while looking ahead at practical implications considered in Part Four.

Spence reflects further on the significance of Libet's research. He agrees with previous authors in refusing to locate free will in the possibility of an immediate veto of the urge to act, and, with Frith, Thompson, and others, looks both to the longer-term and to the social context of action. His reflections are sharpened by raising questions about the significance of Libet's work for understanding patients with movement disorders. What does it mean, now, to ask whether an abnormal movement is or is not voluntary? Many patients with pathogenic movements exhibit the same *Bereitschaftspotential* beforehand as is found in normal movement. Some take this occurrence to mean that the movements are in fact voluntary, while others take its occurrence to indicate that normal actions are not voluntary. So when a killer raises a knife does it matter whether he exhibited a *Bereitschaftspotential* beforehand?

Spence concludes that these ambiguities show the importance of awareness of the consequences of our actions in determining responsibility, and it is not the awareness or lack of it in the milliseconds before acting. He asks: if we cannot, post-Libet, claim authorship of actions in the short term, over the milliseconds preceding them, how might we still maintain some form of responsibility, in moral, legal, and religious senses? He proposes that our moral accountability lies in whether we exercise "meta-responsibility" for our own future behavior, given that we know we cannot always take control of immediate responses. We live in long-term relations with our behaviors and propensities, and we become ourselves as we take charge of planning for them. There are simple cases such as deciding not to drink to excess because of knowledge of what one is likely to do when under the influence. In addition, an agent may choose over long time scales to rehearse certain behaviors in preference to others, in light of a cultivated understanding of and concern about the consequences of our actions for others. This ability to care *for* others depends on how we are cared for *by* others. In the right circumstances and in the right company, conscious awareness is potentially redemptive; it tells us about ourselves. The social world holds us in a kind of equilibrium. The character formation we receive when young puts us in position to choose actions which, in the long term, create our future behavior by forming appropriate brain circuits; it thereby allows us "to take care of our automatisms."

While Spence has introduced practical concerns (e.g., psychiatric diagnoses) related to the research reviewed in this volume, the first two chapters of **Part Four** turn specifically to application, particularly in the field of law. **David Hodgson** is a practicing judge, who regularly faces the question of the relevance of developments in neuroscience for determining sentencing guidelines. In chapter 14, "Criminal Responsibility, Free Will, and Neuroscience," Hodgson notes that reactions to neuroscientific findings range from *fear* that they will sound the death knell for notions of free will and responsibility, to those who *welcome* such a change because they see it as promoting a new approach to criminal behavior not distorted by primitive and inhumane ideas of retribution and vengeance.

This split raises the question of the purpose of punishment. There are two concepts here: a backward-looking focus on retribution and a forward-looking consequentialist position. The latter incorporates the goals of deterrence, restraint of the criminal from further crimes, placation of victims, and reassurance to the community that they are being protected from criminals. These two conceptions of punishment relate in various ways. One point of intersection regards the question of whether the defendant evidences not only a guilty act, but also a guilty mind; another is in determining what punishment is appropriate. "A defect of reason from disease of the mind" mitigates against guilt and thus can lead to lesser retributive punishment, but there are offenses of strict or absolute liability in which diminished capacity is not relevant because consequentialist considerations are sufficient to justify placing the onus on citizens to make sure the event in question does not occur.

Neuroscience now adds to the list of scientific developments that have led many in the past to call for the elimination of retributive punishment – from Laplace's total physical determinism to Freud's emphasis on unconscious drives. In the face of these arguments, Hodgson defends retribution as a guiding purpose of criminal law. His reasons include the following: (1) If the state only attends to the consequences of what it does to citizens this amounts to treating them as objects rather than responsible human beings. (2) Making punishment dependent on wrongdoing reassures the innocent that their compliance with the law will protect them from loss of liberty, and will deter them from taking justice into their own hands. Finally (3) proportionate retribution is consistent with the goals of consequentialist theories of punishment.

Hodgson concludes that the necessity of distinguishing between the guilty and the innocent requires that we maintain the policy of regarding people as free and responsible. Some make the argument that we could maintain the policy even if we believe that free will is an illusion. A second argument is that compatibilist free will is a sufficient basis for maintaining current legal practice. Hodgson argues that we in fact need libertarian free will in legal rationales, and defines it as the ability consciously to grasp and be guided by reasons.

Hodgson agrees with O'Connor and Spence in noting that the capacity to be guided by good reasons varies; we are greatly affected by who we are as we come into the world. He extends the metaphor of having been dealt a better or worse hand of cards by pointing out that we *are* the cards that circumstances have dealt. The capacity for conscious decision-making is the Joker in the hand that allows us, so long as the other cards are acceptable, to be responsible for our actions.

The previous chapters in this volume have shown the reasonableness of Hodgson's position, despite recent neuroscientific findings. In addition, Hodgson notes that neuroscience will continue to contribute in a positive way to the legal system, by increasingly helping to determine questions of responsibility, in identifying brain conditions that involve particular risks of criminal behavior and devising methods to minimize the risks, in devising programs for rehabilitation, and improving reliability in evaluation of evidence.

Chapter 15, "Law, Responsibility, and the Brain," by **Dean Mobbs, Hakwan Lau, Owen Jones**, and **Chris Frith**, contributes to the goal Hodgson sets for neuroscience of identifying brain conditions that contribute to risks for criminal behavior, and of assessing their implications for the legal system. Ever since the accident befalling the now famous Phineas Gage it has been known that brain damage can compromise one's ability to act in conformity to moral judgment. As previous chapters have argued, the prefrontal cortex, a latecomer in phylogenetic history, is essential for rationality and morality. Severe damage here can result in acquired sociopathy. The authors cite studies showing particular prefrontal regions associated with pro-social behavior: anterior cingulate cortex is associated with empathy; orbital PFC with regret; ventromedial PFC with ethical decisions; ventrolateral PFC with inhibition of behavior; and dorsolateral PFC with reasoning.

Mobbs and colleagues distinguish criminal behavior into two types. Affective aggression is impulsive, emotional, and involves autonomic arousal. Predatory aggression is premeditated, goal-directed, and emotionless. The value of this distinction has been demonstrated by research showing that impulsive murderers exhibited reduced activation in the bilateral PFC, while activity in limbic structures was enhanced. Conversely, predatory psychopaths had relatively normal prefrontal functioning, but increased right subcortical activity, which included the amygdala and hippocampus.

Mobbs and colleagues also present findings related to the causes of criminal behavior. Studies show that 25 percent of defendants are medically and legally incompetent to stand trial. Clinical diagnosis of antisocial personality disorder (APD), defined as lack of regard for others' feelings and failure to abide by societal rules, has been found to be ten times higher in the prison population than the rate in the general population. In addition, people with APD often have a history of childhood trauma and maltreatment.

This chapter adds to Hodgson's list of possible benefits of future neuroscientific studies: understanding how cognitive processes of trial participants such as judges and jurors affect outcomes; examining assumptions underlying evidentiary rules, including the limits of witness memories; learning how people determine "just" punishments and react to certain kinds of character evidence; and determining the extent of injury from accidents. However, the authors maintain, the primary role of neuroscience will be to improve the court's ability to identify those cases that fall within the category of "not guilty by reason of insanity." They illustrate the claim that neuro-imaging will be useful here with the example of a man who suddenly succumbed to pedophilia, and was found to have a large tumor in his right orbitofrontal cortex; and a fifteen-year-old who killed family and friends, and was then found to have cavities in his frontal lobe. They argue that the fact that PFC continues to develop up to the age of 25 should be taken into account in sentencing of offenders under that age. However, they conclude with cautions regarding the sorts of information that brain imaging *cannot* be expected to provide.

Hans Küng's chapter, "The Controversy over Brain Research," provides a fine overview of many of the conclusions reached in this volume. He argues that philosophers and theologians can no longer discuss human nature without taking the findings of neuroscience into account. In particular, they cannot merely postulate free will on theological grounds. However, the research by Libet and Wegner is not sufficient to show that in the normal case free will is an illusion and, in particular, it does not invalidate legal attributions of guilt.

With Mobbs and colleagues, Küng points to the limits of what neuroscience can tell us. It is never possible to read the feelings and thoughts of a person from brain images. Küng cites a manifesto by German neuroscientists Gerhard Roth and Wolf Singer, warning that while it is permissible to ask the big questions of neuroscience such as that of free will, it is unrealistic to think they will be answered soon. Küng notes in particular the lack of a widely accepted account of the relation between brain and consciousness.

Küng also provides an overview of reasons for rejecting neurobiological reductionism. He agrees with previous authors in pointing out the limited relevance of the small units of action in Libet-type experiments. With Spence he emphasizes the importance of human ability to set goals and pursue them over time, and with both Kelso and Spence, the importance of culture in supporting the cognitive achievements that contribute to our capacity for moral responsibility. With Newsome he points out that brain scientists themselves have to presuppose responsible authorship in themselves and their colleagues. With Spence he emphasizes that better understanding of our own automatisms can extend freedom, since we are able to take them into account in long-term planning, and this planning must involve care for the consequences of our actions for others, within a shared system of moral norms.

3 Analysis of the Volume

In this section I offer reflections on the achievement of the current volume. After Koch has set up the problems to be addressed, the book defends against over-interpretation of Libet's, Wegner's and others' research in four interrelated ways. First, it sets up a framework for rejecting reductionist accounts of human life in general, by considering emergence, downward causation, complex dynamical systems, and finally by applying system dynamics to brain and behavior. This is important for disputing the metaphysical thesis of universal determinism, and is particularly relevant in addressing the research reported here. Libet-style research involves what Warren Brown and I call Cartesian materialism, by which we mean the assumption that the real "I" is reducible to my consciousness or to any sort of event *inside* my head (Murphy & Brown 2007).[5] The attribution of agency to something inside the person – such as a brain event – is one instance of reductionism, in that it assumes that the parts unilaterally determine the behavior of the whole. In contrast, chapters 3 through 6 have shown that the brain in the body, considered as a complex dynamical system, should be expected to be affected by the actions of the person, especially the person's interactions with the social environment.

Second, this book examines the research itself in detail, relating it to other relevant cognitive-neuroscientific experimentation. Various authors note ambiguities in the research, and, more importantly, they call into question the overly hasty extrapolation from experiments involving quite trivial sorts of choices to grand conclusions about free will. This sets the stage, third, for consideration of the ways in which free will and responsibility pertain to the larger picture of human action, outside of the laboratory, in which we are able to recognize the degree of automaticity in our responses to stimuli. In light of long-term goals, of social expectations, and finally in light of ethical norms, we can become the authors of our own character. This is exactly the sort of conclusion that contributions in Part One should have led us to expect.

The fourth move of the volume as a whole is to turn the tables on the neuroscientific research, in the sense of using it to pursue the question of what brain regions and systems are involved in *enabling* responsible action: How can neuroscience help us to distinguish between responsible action and aberrant cases,

[5] Daniel Dennett coined this term, but uses it more narrowly to refer to scientists who believe that there must be some location in the brain (the Cartesian theater) where all neural/mental activity comes together.

and further, to understand how our remarkable neural systems in fact create the capacity for morally and legally responsible action?

In pursuing in our own work on questions similar to those addressed by this volume, Warren Brown and I have found moral philosopher Alasdair MacIntyre's account of the cognitive prerequisites for morally responsible action immensely helpful (MacIntyre 1999). Our summary of MacIntyre's account of the capacity for moral responsibility is *the ability to evaluate that which moves one to act in light of a concept of the good.* Note that his concern here is not to present a criterion by which particular actions can be judged as morally responsible, but rather to ask the philosophical question of what are the essential requirements for anyone's attaining the capacity to act in a fully mature, rational, responsible, and moral manner. Brown and I make one modification that takes into account the fact, noted by Hallett and Frith, that humans and other organisms are intrinsically and spontaneously active. A better formulation, then, is that one is morally responsible when one has the ability to evaluate, in light of a concept of the good, the factors that serve to shape and modify one's actions. Here is how MacIntyre ties together the capacities that comprise practical reasoning:

> as a practical reasoner I have to be able to imagine different possible futures *for me,* to imagine myself moving forward from the starting point of the present in different directions. For different or alternative futures present me with different and alternative sets of goods to be achieved, with different possible modes of flourishing. And it is important that I should be able to envisage both nearer and more distant futures and to attach probabilities, even if only in a rough and ready way, to the future results of acting in one way rather than another. For this both knowledge and imagination are necessary. (MacIntyre 1999, pp. 74–75)

Brown and I drew from this overview a list of more basic cognitive prerequisites:

1. A symbolic sense of self ("different possible futures *for me*").
2. A sense of the narrative unity of life ("to imagine myself moving forward from … the present"; "nearer and more distant futures").
3. The ability to run behavioral scenarios ("imagination") and predict the outcome ("knowledge"; "attach probabilities … to the future results").
4. The ability to evaluate predicted outcomes in light of goals.
5. The ability to evaluate the goals themselves ("alternative sets of goods … different possible modes of flourishing") in light of abstract concepts.
6. The ability to act in light of 1 through 5.

As I have pointed out in section 1, and Timothy O'Connor will confirm in his chapter, free-will language in philosophical debates is not well attuned to the realities of life. In particular, a stalemate has been created by rigidly categorizing concepts of free will as either libertarian or compatibilist. So Brown and I argued that free will be understood as having and using this capacity for morally responsible action. Our account does not make the (untestable) claim that for any particular act in the past, I could have done otherwise, but focuses instead on the question of whether I will be able to choose differently in similar situations in the future.

This MacIntyrean account of moral responsibility has been supported in various ways by work in this volume: by Thompson's emphasis on meta-awareness and the ability it gives us to inhibit impulses and desires to act, by Frith's distinction between bottom-up and top-down control of action, by Spence's emphasis on considering the consequences of our actions for others, and by O'Connor's distinctions between automatisms and urges to act on the one hand, and on the other, the ability to be aware of our desires and beliefs, along with our total motivational structure. The present volume, in a variety of ways, has shown the role of meta-awareness, meta-responsibility, top-down control – what Brown and I call self-transcendence – in freeing human behavior from both internal drives and external influences, and allowing for increasing flexibility and autonomy.

What MacIntyre's analysis shows as needing to be added to the emphases in this volume is a reflection on the role of symbolic language.[6] Symbolic language is necessary for a sense of self. M.R. Bennett and P.M.S. Hacker write that "[t]he idea of me" depends on the ability to use the words "I" and "me," and these words cannot be used correctly without acquisition of a system of words including second- and third-person pronouns (Bennett & Hacker 2003, p. 348).

Abstract symbolic language is also necessary for imagining long-term futures, making predictions, and for conceiving of abstract goals such as moral goodness. It is necessary for formulating the reasons that guide actions. MacIntyre emphasizes that complex syntactic abilities are required for evaluating actions. This sort of meta-level judgment requires language with the resources necessary for constructing sentences that contain as constituents a representation of the first-order judgment. That is, mature human rationality develops when children attain the ability to consider why they are doing what they are doing, and then to raise the question of whether there might be better reasons for acting differently (MacIntyre 1999, pp. 53–54). This requires the linguistic capacity to be able to say something like the following: "I wanted to smoke to impress my friends, but I decided that it was more important to take care of my health."

Bennett and Hacker also emphasize the role of language in the sort of meta-level awareness that enables character formation and moral responsibility. One who has developed the sophisticated linguistic powers

> to use proper names and pronouns, as well as psychological predicates and predicates of action, in both the first- and third-person cases and in the various tenses .... is a self-conscious creature, who has the ability to be transitively conscious of its own mental states and conditions, who can think and reflect on how things are with it, who can not only act but also become and be conscious of itself as so acting. And it will also have the ability to reflect on its own past, on its character traits and dispositions, on its preferences, motives and reasons for action. (Bennett & Hacker 2003, p. 334)

Given the extensive research that has been done on the neural correlates of language use, we see again that neuroscience does not so much threaten free will as

[6] Ellis does make this point briefly in chap. 3.

give us insight into the ways in which our complex neural equipment *enables* us, in Koch's terms, to "escape the iron law of cause and effect." While we do not claim to have solved *the* free-will problem, we do claim that it can help in seeing how there could be *space* for free will in human life. Further advances here will depend on developments in neuroscience, particularly on solving "the hard problem of consciousness," and we judge this solution to be still some distance away.

References

Baumeister, R.F., Sommer, K.L.: Consciousness, free choice, and automaticity. In: Wyer Jr., R.S. (ed.) Advances in social cognition, vol. 10. Erlbaum, Mahwah (1997)

Bennett, M.R., Hacker, P.M.S.: Philosophical foundations of neuroscience. Blackwell, Oxford (2003)

Berofsky, B.: Determinism. In: Audi, R. (ed.) The Cambridge dictionary of philosophy, pp. 199–200. Cambridge University Press, Cambridge (1995)

Butterfield, J.: Determinism. In: Craig, E. (ed.) Routledge encyclopedia of philosophy, vol. 3, pp. 33–39. Routledge, London (1998)

Campbell, D.T.: 'Downward causation' in hierarchically organised biological systems. In: Ayala, F.J., Dobzhansky, T. (eds.) Studies in the philosophy of biology, pp. 179–186. University of California Press, Berkeley and Los Angeles (1974)

Deacon, T.W.: Three levels of emergent phenomena. In: Murphy, N.C., Stoeger, W.R. (eds.) Evolution & emergence: Systems, organisms, persons, pp. 88–110. Oxford University Press, Oxford (2007)

Hempel, C.: Aspects of scientific explanation. Free Press, New York (1965)

Kornhuber, H.H., Deecke, L.: Hirnpotentialänderungen bei Willkürbewegungen und passiven Bewegungen des Menschen: Bereitschaftspotential und reafferente Potentiale. Pflügers Archiv Physiologie 284, 1–17 (1965)

Libet, B.: Do we have free will? In: Libet, B., Freeman, A., Sutherland, K. (eds.) The volitional brain: Towards a neuroscience of free will. Imprint Academic, Exeter (1999)

MacIntyre, A.C.: Dependent rational animals: Why human beings need the virtues. Open Court, Chicago (1999)

Murphy, N.C., Brown, W.S.: Did my neurons make me do it? Philosophical and neurobiological perspectives on moral responsibility and free will. Oxford University Press, Oxford (2007)

Nagel, E.: The structure of science. Harcourt, Brace, and World, New York (1961)

Pojman, L.: Freedom and determinism: A contemporary discussion. Zygon 22, 397–417 (1987)

Ryle, G.: The concept of mind. University of Chicago Press, Chicago (1949)

Sellars, R.W.: Principles of emergent realism: The philosophical essays of Roy Wood Sellars. In: Preston Warren, W. (ed.), Warren H. Green, Inc., St. Louis (1970)

Strawson, G.: Free will. In: Craig, E. (ed.) Routledge encyclopedia of philosophy, vol. 3, pp. 743–753. Routledge, London (1998)

Van Gulick, R.: Who's in charge here? And who's doing all the work? In: Heil, J., Mele, A. (eds.) Mental causation, pp. 233–256. Clarendon Press, Oxford (1995)

Wegner, D.M.: The illusion of conscious will. MIT Press, Cambridge (2002)

Part I: Physics, Emergence, and Complex Systems

2

Free Will, Physics, Biology, and the Brain

Christof Koch

Division of Biology and Division of Engineering and Applied Science
California Institute of Technology
Pasadena, CA 91125
koch.christof@gmail.com

Summary. This introduction reviews the traditionally conceived question of free will from the point of view of a physicist turned neurobiologist. I discuss the quantum mechanic evidence that has brought us to the view that the world, including our brains, is not completely determined by physics and that even very simple nervous systems are subject to deterministic chaos. However, it is unclear how consciousness or any other extra-physical agent could take advantage of this situation to effect a change in the world, except possibly by realizing one quantum possibility over another. While the brain is a highly nonlinear and stochastic system, it remains unclear to what extent individual quantum effects can affect its output behavior. Finally, I discuss several cognitive neuroscience experiments suggesting that in many instances, our brain decides prior to our conscious mind, and that we often ignorant of our brain's decisions.

Keywords: Determinism, indeterminism, free will, quantum indeterminacy, neurons, stochastic synaptic release, flies, roundworm, readiness potential, choice blindness.

> You see there is only one constant. One universal. It is the only real truth. Causality. Action, reaction. Cause and effect. *The Merovingian,* from *The Matrix Trilogy*

In a remote corner of the universe, on a small blue planet gravitating around a humdrum sun in the nonfashionable, outer districts of the Milky Way, arose organisms from the primordial mud and ooze through an epic struggle for survival that spanned the eons. Despite all evidence to the contrary, these bipedal creatures thought of themselves as extraordinarily privileged, as occupying a unique place in a cosmos of a trillion, trillion stars. Conceited as they where, they even believed that they, and only they, could escape the iron law of cause and effect that governs everything else by virtue of something they called free will. For this allowed them to do things without any material reason.

Can humans – and other animals as well – truly act freely? Can we do and say things that are not a direct consequence of our predispositions and circumstances? That is the topic of this book. Did you, the reader of these pages, choose to read this book of your own free will? To you, it felt like you voluntarily decided to browse through these pages in the face of competing interests – eating lunch, playing a video game, or running in the mountains. But is that the whole story? Were there not external causes that influenced you – a reading assignment for a mind-brain class, a friend mentioning its fluid style, and so on? You might argue that these causes were not sufficient, that something else had to intervene, your will. Yet the ancient doctrine of predestination and its modern version, determinism, holds that you could not have acted in any other way. You had no choice in the matter. You are a life-long indentured servant to an absolute tyrant. You never had the option of eating lunch. You were destined from the beginning of time to read these lines.

Richard Wagner's monumental *Der Ring des Nibelungen* is a twenty-one-hour mini-series centered on the conflict between fate and freedom. Its hero is Siegfried. Unrestrained by fear or by the mores of society, he kills the dragon, shatters the spear of Wotan and walks through the ring of fire to woo Brünhilde, thereby precipitating the destruction of the old world order of the Gods. Siegfried follows no laws but his inner desires and impulses. He is free but acts blindly, without understanding the consequences of his actions. It is left to the opera's heroine, Brünhilde, to freely and knowingly usher in the new age of Man by her self-sacrifice. This drama is set to some of the most glorious and moving music ever composed.

Whether or not biological organisms are free is no mere philosophical banter; it engages people in a way that few other scientific questions do. Free will touches upon our most cherished beliefs, institutions, and cultural practices. Ultimately it is about the level of control we can exercise over our life and how much responsibility we have for our actions.

The question of free will – what it means and whether it exists – is as old as philosophy itself, with an enormous literature (for a handbook on Western perspectives on free will see Kane 2002). Arcane and eristic arguments have been advanced for or against whatever position one might conceivably hold. Let us not be too distracted by these millennia of learned and disputatious philosophical debate, and focus on what physics, neuroscience, and psychology have to contribute to this aspect of the mind-body problem. Science has discovered matters that open up new ways of thinking about the ancient conundrum of free will.

1 Free Will Comes in Different Shades: Strong versus Pragmatic

Let me offer one intuitive definition of free will: *you are free if, under the same circumstance, you could have acted otherwise.* You voluntarily chose one but could have also chosen the other. This is the *libertarian* or *Cartesian* position, *Will* with a capital W. Think of the iconographic scene in *The Matrix,* where Neo must decide whether to swallow the blue pill Morpheus offered him – with its promise of blissful ignorance – or the red pill – a painful awakening into reality. Neo freely willing the latter meant that he could have taken, with equal ease, the blue pill, depriving us of one of the most compelling movies in recent memory.

Today, on December 31, 2007, at 12:33 p.m. in a small cafe in East Berlin where I am writing these lines, I ordered a glass of coke. If this was a truly free act, I could have ordered a cup of coffee instead. At question is not whether on the next day, sitting in the same cafe, I could have ordered coffee. For the next day, I might be less thirsty for sugary brown water, or it might be colder and the warmth of the coffee would appeal to me, or I might be tired and need the kick of the caffeine, or whatever other reasons could make me choose one over the other. No, the strong view on will demands that on December 31, 2007, at 12:33 p.m., with my brain in the identical state as the first time around – its hundred billion nerve cells firing or not firing, its trillions of synapses hot or not – I could have opted for the coffee instead of for the soft drink.

Of course, in the real world where I cannot travel back in time, I shall never know whether I could have done otherwise. Or as the ancient sage Heraclitus famously declared: *You can't step into the same river twice.*

The Cartesian view of will is the one most prevalent in the general American public. It is closely linked to the notion of a soul. Hovering above the brain like Casper the Ghost, the will freely decides this way or that, making the brain act out its wishes, like the driver who takes the car down this or that road.

For this sort of free will to exist, two different conditions must be meet. First, the universe must not be fully determined. That is, there must be genuine choice – a choice that is compatible with the laws of physics, but that is not fully dictated by physics. Otherwise the will has no traction, has nowhere to act. Second, free will, whatever it is, must be able to influence the world so that one of these possibilities is actually realized. That is, will must have some genuine causal powers.

Contrast this with a more nuanced conception of freedom, referred to as a *compatibilist* belief, advocated by Thomas Hobbes. It is the dominant view in legal and medical circles. You are free if you can follow your own desires and preferences and if you are not in the throes of some inner compulsion or addiction nor acting under the undue influence of other persons or powers.

Living in a Western-style liberal democracy, you can freely vote or articulate your opinion about the worthlessness of politicians and parties without fear of retribution. You are not free to shout "fire" in a crowded theater, but otherwise you have great latitude in what you can say. You are free to practice any religion, including none at all. It is for these political and religious freedoms that our immediate or more distant ancestors fought and died for.

Criminal law recognizes instances of diminished responsibility where the accused did not act freely. The husband who beats the lover of his wife to death in a blind rage when he catches them *in flagrante dilicto* is considered less guilty than were he to kill him weeks later in a premeditated manner. The law is more lenient when it can be proven that the culprit acted under some strong inner compulsion, say because he is diagnosed with schizophrenia and acted out what the voice in his head commands him to do. This is what the "not guilty by reason of insanity" defense is all about. Without such attenuating circumstances, the accused is assumed to be competent to stand trial. Contemporary society and the judicial system is built upon such a pragmatic, *psychological notion of freedom.*

We can try to dig deeper, trying to discern the underlying causes of such "free" actions. Your daily marathon is a gauntlet of choices – which shirt to wear, which radio station to listen to, which dish to order, when to come home, and so on. Yet your biases and habits effectively constrain these choices. My closet is full of colorful checkered shirts, I listen to National Public Radio, don't eat the flesh of mammals, and try to be home before midnight. So my freedom is restricted by the consistent choices I've made in the past.

The very riverbed that holds and channels your stream of consciousness is fashioned by the family and the culture you were raised in. Consider slavery. To us, children of the Enlightenment, owning other persons and disposing of them as property is an abhorrence. Yet for the ancient Greek philosophers – and for some of America's founding fathers – slavery was a natural state, a consequence of conquest of one people by another (Garnsey 1999). You did not freely decide that slavery is morally repugnant; rather you were born into a society that thought so. Yes, even as an ancient Greek you could have decided that slavery is not justifiable from an ethical point of view; but when nobody around you questions commonplace assumptions, it is very difficult for you to do so.

Freedom in this psychological sense leaves a residue of unease; absence of overt inner and outer coercion is necessary to feel free. But does it guarantee freedom in the strong sense? If all external factors that might conceivably influence you are accounted for – physical, genetic, neurobiological, environmental, and cultural ones – is there any room left to maneuver? Is there any freedom left to act this or that way? Isn't it likely that you are, unknowingly, an utter slave to these constraints and that your freedom is illusory? Has our conceptual spadework hit the underlying bedrock of determinism? Let's see what physics has to say about this matter.

2 Physics and Choice: The Clockwork Universe

A high point in humankind's ongoing process of understanding the cosmos occurred in 1686, the year Isaac Newton published his *Principia,* enunciated the law of universal gravitation and the three laws of motion. Newton's second law links the force brought upon an isolated system – say a planet orbiting a star or a billiard ball rolling on a velvet green table – to its changing velocity. It has profound consequences. For it implies that the positions and velocities of all the components making up an entity at any one particular moment in time, together with the forces between them, unalterably determine its fate, that is, its future location and speed.[1] Nothing else intervenes; nothing else is needed. The destiny of the system is sealed until the end of time. The domain of this law extends throughout the land – whether it is the force of gravity, electrical charges attracting or repelling each other, the shove and push of mechanical forces – the second law unifies them all. Given the forces and the precise location and the speed of all components of the system, the state of the system at any future point in time can be foretold.

This is the clockwork view of the universe. Knowing the mass, location, and velocities of the planets as they plow their orbits around the sun fully determines where they will be in a thousand, a million or a billion years from today, provided only that we properly account for all the forces acting on them. Or, put differently, the forces are sufficient causes for all future events; where the planets will be and how fast they will move is strictly necessitated by these forces and only by them.

The application of Newton's law requires that the system under study is a closed one, isolated from the rest of the universe. For otherwise something outside the system could reach inside, adjusting things. Consider billiard balls. They move along trajectories fully determined by collisions with other balls and the side of the table as well as by friction between the ball and the green cloth covering the table. Their trajectories are perfectly predictable using Newton's laws. But the player who places a cue ball on the table acts as an external agent, interfering with this predictability, supervening the simple laws of cause and effect that govern the motion of the balls on the table. The asteroid that crashed into Earth on a spring morning 65 million years ago, putting a fiery and permanent end to the Age of Dinosaurs, is an example of an outside, celestial agent perturbing the course of evolution on the planet. Indeed, the deeply religious Newton assumed that the Creator had to periodically intervene in the universe to prevent such catastrophes and keep the solar system on track – an interventionist God who does so without violating the laws of physics.

If the motions of the player are considered part of the billiard game or the trajectory of all asteroids as part of an investigation into the future of life on Earth, then both events would have been predictable and determinism rescued. So the reach and extent of the system under study must be extended until all possible factors that can conceivably influence its fate have been considered. The most grandiose expression of determinism avoids this complication by considering the entire universe as a whole, rather than any part of it in isolation. This conceptual leap finds its most eloquent proponent in the French mathematician Pierre Simon Marquis de Laplace (Laplace 1951). In 1814 he penned these lines:

[1] Technically, this is only true if the underlying differential equations have a unique solution for the time considered; that is, in the absence of any singularities. This cannot always be guaranteed, in particular for 1/r potentials (Siegel & Moser 1971).

> We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, and if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes.

The universe, once set in motion, runs inexorably on its course. It is a clockwork that slowly, ever so slowly, unwinds over billion of years, until it has run its course. To an all-knowing superior intellect – think of a supercomputer – the future is an open book. There is no freedom above and beyond that dictated by the laws of physics. All of our personal struggles to come to grips with our inner demons, to accept our deeds, both good and bad, is for naught. For the outcome was ordained when the universe was wound up at the beginning of time.

The first hint that this colossal machine was not quite as predictable as expected came in the closing decades of the nineteenth century from the mathematician Henri Poincaré in the context of studying the three-body problem (Strogatz 2000). But it took the digital computer in the second half of the twentieth century to reveal *deterministic chaos* for what it is – a full-blown setback for the notion that the future can be accurately forecast. It was the MIT meteorologist Edward Lorenz who discovered this in the context of solving three simple mathematical equations characterizing the motion of the atmosphere. The weather predicted by the computer program varied widely when he entered starting values that differed by less than a tenth of one percent: this is the hallmark of chaos – infinitesimal small changes, tiny perturbations in where the equations start off lead to radically different outcomes. Lorenz coined the term *Butterfly Effect* to denote such extreme sensitivity to initial conditions: the beating of a butterfly's wings creates barely perceptible ripples in the atmosphere that ultimately alter the path of a tornado elsewhere (Lorenz 1995).

Chaos is the reason why precise long-term weather prediction will never be in the cards. Meteorologists must record the local temperature, barometric pressure, humidity, solar radiation, wind speed, and so on quite accurately to assess future weather patterns. For the sake of this argument, let us assume that they must be measured to within a few percent of their true values in order to forecast coastal fog in the morning a couple of days hence. To forecast fog a week from now, these variables need to be estimated to within a fraction of a percent of their true value; if one wanted to know about fog in ten days time it would require a degree of accuracy unattainable in the real world due to all the uncertainties and fluctuations in the atmosphere.

The epitome of the Newtonian-Laplacian clockwork universe is celestial mechanics. Planets majestically ride gravity's geodesics, propelled by the initial rotation of the cloud that formed the solar system. It came as a mighty surprise when computer modeling demonstrated that Pluto has a chaotic orbit, with a divergence time of about ten to twenty million years (Sussman & Wisdom 1988). Put differently, astronomers cannot be certain whether Pluto will be on this side of the sun (relative to Earth's position) or the other side ten million years from now! No matter how small the residue of our measurement error, it will never vanish and therefore will always limit how far we can peer into the future.

If this uncertainty holds for the position of a planet-sized body in deep space, what does this portend for the predictability of a single synapse deeply embedded inside a brain, let alone the action of a nervous system of millions or billions of nerve cells, each one encrusted with thousands of synapses? Given the nonlinear and cooperative nature of such neural networks, their behavior is chaotic to a high degree.

But chaos in the mathematical sense of extreme sensitivity to initial conditions does not invalidate the law of cause and effect. It continues to reign supreme. We do not know where Pluto will be eons from now, but we are sure that its orbit is always completely in thrall to gravity and other physical forces and nothing else. In theory we could accurately forecast its position. What breaks down in chaos is not the chain of action and reaction but our practical ability to predict events. The universe is still a gigantic clockwork, even though we're not sure that the minute and hour hands will both simultaneously point to the top of the dial at midnight a week hence.

The same point can be made about biology. Any organelle, such as the nucleus of a cell or a synapse, is made out of a fantastically large number of molecules suspended in watery solution. These molecules incessantly jostle and move about in a way that can't be precisely captured; this is called *noise*. Physicists are unable to track individual molecules. To tame this noise, they borrow techniques from statistics and from probability theory, calculating the average kinetic energy of the molecules or the average time between synaptic release and so on.

A good example of such averaging is the theory of the random motion of small particles suspended in water or air. Known as *Brownian motion,* the glittering, tumbling movement of motes of dust in a beam of sun was already used by the Roman scholar Lucretius as proof for the existence of atoms. While it is impossible to forecast the motion of any one particle, one can predict quite precisely how quickly a large cloud of such particles, say a blot of ink dropped into water, disperses (Atkins 1984).

Once again, the randomness inherent in molecular motion and related processes is not because of any fundamental limit on precision or the breakdown and abandonment of determinism at the microscopic scale. No, it is for practical reasons that we will never be able to follow the ceaseless motion of a gazillion molecules. Practically speaking, Newton's laws can be profitably applied to a handful of billiard balls, missiles, or planets, but computers fail when asked to determine the fate of such vast number of molecular components. But under the laws of classical physics, there is no reason to deny that given the forces and the initial position and velocity of all molecules, their future state follows inexorably.

3 The Demise of the Clockwork Universe

This deterministic, not to say fatalistic, view of the universe changed decisively and radically with the birth of Quantum Mechanics in the 1920s. The terminal blow to the Newtonian-Laplacian dream – or nightmare, depending on your point of view – is the celebrated *Uncertainty Principle* formulated by Werner Heisenberg in 1927: an irreducible limitation on how precisely the position of a particular and its momentum can be measured. In its most common interpretation, Heisenberg's principle avers that the universe is built in such a way than any particle, say a photon of light or an electron, cannot at the same time have both a definite position and a definite momentum. If you know its speed very accurately, its position is correspondingly ill defined and vice versa. It is not any limitation of our instruments that we might be able to overcome with better technology. No – it is built into the very fabric of reality. Macroscopic, heavy objects like my red Mini convertible occupy a precise position in space while moving at a well-defined speed along the freeway. But microscopic things such as elementary particles or small atoms and molecules violate common sense: the more precisely you determine where they are, the more uncertain, the more fuzzy, is their speed, and vice versa.

Heisenberg's uncertainty principle is a permanent departure from classical physics, with repercussions that have not yet been fully worked out. It replaces dogmatic certainty with principled randomness. The ultimate reality is a mathematical abstraction called the wave function. It evolves in a deterministic manner, dictated by Schrödinger's law. From it, physicists can derive the probability of a family of events happening – say the probability that an electron occupies a particular orbit around the nucleus of a hydrogen ion. The probabilities themselves can be calculated accurately to a fantastic degree. But whether the electron will actually occupy this orbit is left up to chance – subject to these probabilities.

Consider an experiment that ends with the electron being here with a 90% chance and over there with a 10% probability. If we were to run the same experiment over and over, a total of one thousand times, on about nine hundred trials the electron would be here and on only one hundred trials would it be there. Yet this does not determine where on the next trial the electron will be. It is more likely to be here than there, but where it finally ends up is truly left up to chance. Albert Einstein could never reconcile himself to this random aspect of nature. But we know it to be a fact.

There is breathtaking evidence of this randomness if you know where to look. Galaxies are not spread evenly through the immensity of space. They cluster along elongated and thin strands, arranged in sheets and walls surrounding trackless and bottomless voids whose vast emptiness staggers the mind. It takes a ray of light millions of years to cross such an abyss! Our own Milky Way is part of the Virgo supercluster of galaxies with tens of trillions of stars. Superclusters are the largest structures in the known universe. According to the inflation theory of cosmology, superclusters were caused by stochastic quantum fluctuations when the universe was very young, far smaller than a head of a pin, an instant after the Big Bang. In the tight confides of the initial brew of mass-energy, things were a bit denser here and somewhat less so over there. The inflationary phase of the early cosmos, when this kernel of the universe exploded outwards to create space itself, amplified this quantum imprint to the stupendous and uneven distribution of galaxies we observe today (Turner 1999). Quantum uncertainty is written in the sky, too large to be seen with the unaided eye.

The universe has an irreducibly random character. If it is a clockwork, its cogs, springs, and levers are not Swiss-made; they do not follow a predetermined path. Physical indeterminism rules in the world of the very small as well as in the world of the very large.

But wait – I hear a serious objection. There is no question that the macroscopic world of our experience – that includes bathroom slippers and brains – is built upon the microscopic, quantum world. But this does not necessarily imply that these objects inherit all of the weird properties of quantum mechanics. These macroscopic systems are constituted by an unfathomable number of microscopic particles such that many of their properties – such as their uncertain positions and velocities – are washed out, and what remains is usually fully deterministic.

Take my convertible again. If I park it, it has zero velocity relative to the pavement. It does not have the infinite positional uncertainty seemingly demanded by Heisenberg. Assuming I did not forget where I last parked the car or that it was not towed or stolen, I will find it at the exact location where I left it. That is because the Mini is enormously heavy compared to an electron and so the fuzziness associated with the position is essentially zero for all intents and purposes. In the macroscopic world we live in, most things are deterministic. I push down on the car's speed pedal, and the engine growls reassuringly. So far, the car has accelerated every time I've hit the metal. And if one day the car stutters and jerks when I try to accelerate, I won't consult my quantum physicist friends but will talk to a car mechanic about quite classical effects, such as a clogged fuel line. So quantum indeterminacy does not appear to matter for the objects that humans and other animals interact with; operationally, they are governed by good old-fashioned Newtonian laws.

There is an alternative school of thought that dates back to the founding days of quantum theory (von Neumann 1932). It postulates an intimate link between quantum mechanics and human consciousness. One notion is that a conscious human observer – whether a monkey would also do has never been considered – is required for the probabilities that quantum mechanics deals with to *collapse* into one or another actual outcome (Wigner 1967). This is the infamous measurement process that has engendered an enormous literature. The British physicist Roger Penrose (Penrose 1994), the American anesthesiologist Stuart Hameroff (Hameroff & Penrose 1996), and others (Stapp 2003) speculate that the more otherworldly, weirder aspects of quantum mechanics, in particular nonlocality – the well-verified observation that certain quantum systems remain mysteriously entangled, no matter how far apart they are separated – are closely linked to consciousness. Entangled quantum systems, such as two coupled electrons or two coupled photons, can be highly correlated even though they are arbitrarily far away. Strands of Buddhism, a much older tradition, likewise argues that object and subject are inexorably linked and that consciousness is a fundamental feature of the physical universe (Wallace 2007).

Yet not all interpretations of quantum mechanics require a collapse of the wave function. Most prominently, the *Many Worlds interpretation* of quantum mechanics (Everett 1957) that is now enjoying a renaissance among physicists (Tegmark 2007) does not. If there is no collapse of the wave function, then the need to evoke the magic of consciousness is gone. Furthermore, the physics of the brain conspires against stable quantum entanglement. Two biophysical operations underpin much information processing in the brain: chemical transmission across the synaptic cleft, and the generation of the action potentials. Both operations involve thousands of ions and neurotransmitter molecules, coupled by diffusion or by the membrane potential that extends across tens of micrometers. Both processes would destroy coherent quantum states. Spiking neurons can only receive and send classical, rather than quantum, information; at each moment, a neuron either spikes – that is, generates, one of these binary pulses – or it does not. It is never in a superposition of spiking and nonspiking. Finally, what would be the advantage of quantum computations from a behavioral point of view? Most of the excitement in quantum computers flows from Shor's (1997) quantum algorithm for factoring large integers for data encryption. Yet the survival value to animals of factoring large numbers into their primes is probably low; so it is totally unclear what algorithmic or computational advantage would accrue to nervous systems that would exploit some of these feature of quantum mechanics (Koch & Hepp 2009). There is no evidence that any components of the nervous system – a warm and wet tissue strongly coupled to its environment – display quantum entanglement (Koch & Hepp 2006). What cannot be ruled out is that tiny quantum fluctuations deep in the brain are amplified by deterministic chaos and will ultimately lead to behavioral choices. This is the basis of Jordan's quantum amplifier hypothesis of free will (Jordan 1938). The release of a single synaptic vesicle may be dependent on some pre-synaptic quantum event. This might generate an action potential in the post-synaptic neuron that, in turn, triggers a cascade of active neurons that ultimately give rise to movement. Biological organisms – from bacteria to bugs to boys – may well act truly randomly, like the proverbial toss of the coin. In that case, the laws of cause and effect do not fully determine behavior. Physics would not, even in principle, predict whether I will choose the glass of coke or the cup of coffee. True choice would become possible.

Personally, I find determinism abhorrent. To believe that the entire evolution of the cosmos and all of its inhabitants is already inherent in the Big Bang evokes a feeling of helplessness in me. While indeterminism does not address the question on whether "I" can make a difference, whether I can start a chain of causation on my own, it insures that my environment and my behavior are, in general, not fully determined by the past. Instead, they are partially contingent on truly random events.

4 The Impoverished Freedom of the Mind to Realize One Quantum Event over Another

Of course, indeterminism is no substitute for free will. For surely my actions should be caused because I want them to happen for one or more reasons rather that they happen by chance. Trading in the certainty of determinism for randomness, for the toss of the coin, is not what Descartes had in mind. Thus, the libertarian conception of the mind requires that the mind controls the brain. The mind has to be able to take advantage of the situation and decide. How could that work?

The Austrian philosopher Karl Popper and the Australian neurophysiologist John Eccles are modern defenders of a soul (Popper & Eccles 1977). Popper is a famous philosopher of science and politics and Eccles a pioneer in the biophysical study of synaptic transmission, work for which he was awarded the Nobel Prize in 1963. So these are both reputable scholars.

According to Popper and Eccles, the mind inhabits its own world, that of subjective states. It is not the material world where stars, dogs, people, and brains reside. This world follows its own rules and regulations that are not the laws of physics. The conscious mind – made out of some sort of metaphysical ectoplasm – imposes its will onto the brain by affecting the way neurons communicate with each other in the part of the cerebral cortex concerned with the planning of movement and action (by adjusting the synaptic release probabilities). According to Beck and Eccles (1992), by promoting synaptic traffic between these neurons in one location and preventing it in another, the conscious mind imposes its will onto the material world. As the brain is exceedingly rich in synaptic connections, numbering perhaps a hundred trillion, the mind is going to be awful busy adjusting even a small fraction of these every time the brain executes a "voluntary" action, such as grasping a glass of water. For those raised with a belief in a strong will, the Popper-Eccles theory is appealing, as it seems to reconcile a religious point of view with a scientific stance.

But is this proposal reasonable on physical grounds? No! Not if the mind directly forces the brain, or some of its components, this way or that. For if the mind intervenes in the material world, it has to do work and this costs energy. And even the minute energy expenditures necessary to tweak synaptic transmission have to show up on nature's balance sheet. Physics does not allow any exceptions. The principle of energy conservation has been tested again and again and always comes out a winner.

That is the trouble with such a mind. If it is truly ephemeral, ineffable, like a ghost or a spirit, it cannot interact with our universe. It cannot be seen, heard, or felt. And it certainly could not make your brain do anything. For it to be able to influence matters, it must be more in the nature of a poltergeist, rumbling and tugging synapses. And to do that, it must expend energy. Nothing and nobody can intervene in the world without leaving a trace. And, to the best of our knowledge, there is no evidence for a spooky and unaccountable force that could do this.

The only freedom that such a mind could have is to realize one quantum-mechanical event rather than another one as dictated by Schrödinger's law. Say, for example, that at a particular point in time and at a particular synapse in cortex, a superposition of two quantum mechanical states occurs. There is a 10% chance that the synapse will switch – sending a chemical signal across the cleft separating two neurons – and a 90% chance that nothing happens. Put differently, if one could repeat the same experiment over and over, say one thousand times, in about one hundred trials the synapse would trigger, while in the remaining nine hundred cases no release would occur. But this does not tell us what would happen the next time around. All we can say is that it is much more likely than not that no synaptic event would occur.

Given our present interpretation of quantum mechanics, it cannot be ruled out that the conscious mind could exploit this idiosyncratic freedom. It is powerless to change these probabilities – that would cost energy – but it might be able to decide what happens on any one trial. The mind's action would always remain covert, *sub rosa*. For if we considered many trials, nothing out of the ordinary would occur; only what is expected from the physics of the situation. Conscious will would act in the world but only within the straightjacket of physics. Following Occam's razor, the notion of this free will appears redundant. It does not explain anything that cannot be explained using quantum indeterminacy.

So the maximal freedom afforded to the conscious mind is an impoverished choice among quantum possibilities. This presupposes that choosing one quantum event over another quantum event will make a difference to the brain and is not washed out by the thermal fluctuations that are such a hallmark of biology at the balmy temperatures where life exists. It also supposes that the conscious mind had the means to somehow select one outcome over another. We do not know whether this is even within the realm of the possible. But at least it cannot be ruled out.

That is not the end of the grave conceptual problems with a Popper-Eccles mind. We would need to know how the brain influences the mind, since surely the mind needs to know what the brain does? It would need to see with its eyes and feel its pain for it to be able to decide upon any course of action. This and other queries are unanswered and perhaps unanswerable.

5 Brains, Animals, and Randomness

A salient feature of nervous systems and their components are their noisy, random character. Individual voltage- and ligand-gated ionic channels – single proteins that are inserted into the neuronal membrane – enable neurons to communicate with each other via chemical synapses and generate and propagate all-or-nothing binary pulses, the action potentials that are the lingua franca of almost all nervous systems. The ionic currents flowing through such channels are microscopic, discrete, and stochastic. Figure 2.1 illustrates this for an acetylcholine-activated channel. Even though the membrane potential is held constant across the channel, the channel fluctuates between an open, high-conductance and a closed, low- (or zero-) conductance state (with some flickering in between). It is only if thousands of channels are closely packed together that the transition from microscopic, binary, and stochastic current flow to macroscopic, continuous, and deterministic current flow, as seen classically in the Hodgkin-Huxley squid axon, occurs (Strassberg & DeFelice 1993). Of course, small neuronal components, such as dendritic spines or distal dendritic sites, may only contain a few channels that thereby signal in a stochastic manner (Koch 1999). Given the large size of channel proteins – the acetylcholine-binding channel of figure 2.1 has a total molecular mass of 268,000 Daltons – it is generally believed that the stochastic character of ionic channels can be entirely explained by thermal fluctuations and does not rely on quantum indeterminacy (Hille 2001).

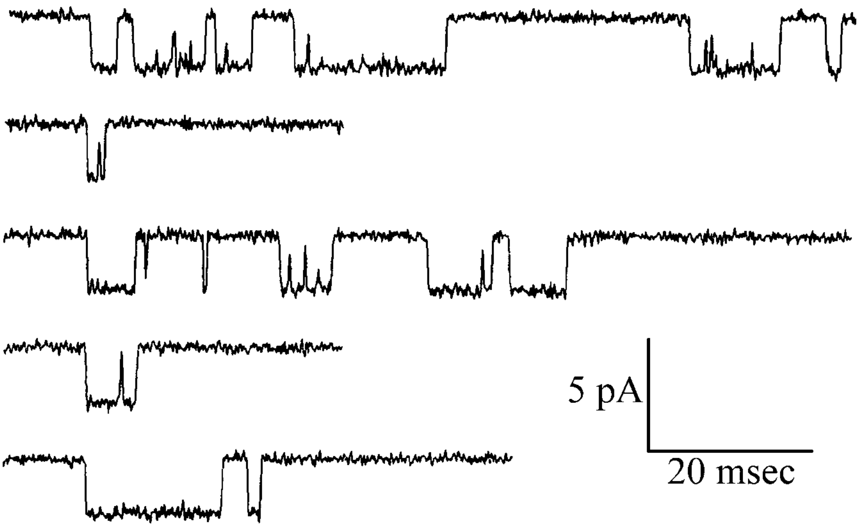

Fig. 2.1. Current Flowing through a Single Ionic Channel. Even though the voltage across this acetylcholine-activated channel, embedded into a muscle cell, is held constant, the channel fluctuates randomly between a closed and an open state. Given the channel's large size, it is believed that its stochastic behavior is entirely explainable by classical, thermal motion. The tiny fluctuations are due to instrument noise. From Koch 1999.

Randomness is also apparent at the level of action potentials. Say that a microelectrode, essentially a conductive wire, is placed close to a nerve cell in the brain of a monkey looking at a display of a randomly moving cloud of dots. Each time the display is turned on, the cell becomes excited and fires a set of all-or-none electrical pulses, "spikes" in neuro-lingo. These can be picked up by the microelectrode. Spikes are the principle means of rapid communication among nerve cells throughout the animal kingdom. If you look carefully, the precise pattern of spikes varies unpredictably from one trial to the next (fig. 2.2), while the average number of spikes remains reasonably constant.

Some of this variability is due to trembling eyes, the exact timing of the heart beat, breathing, and so on. The remaining unpredictability is thought to be accounted for by the incessant movement of the molecules, primarily water, making up the wet and warm brain – thermal motion that I mentioned above. This ceaseless motion cannot be predicted but is still subject to the laws of cause and effect. Biophysicists by and large believe that quantum mechanics has no essential role to play here. While nervous systems – like anything else – obey quantum mechanics, the collective effects of all these molecules frenetically moving about is to smear out any quantum indeterminacy. At the cellular level, neurons look to be firmly governed by classical physics.

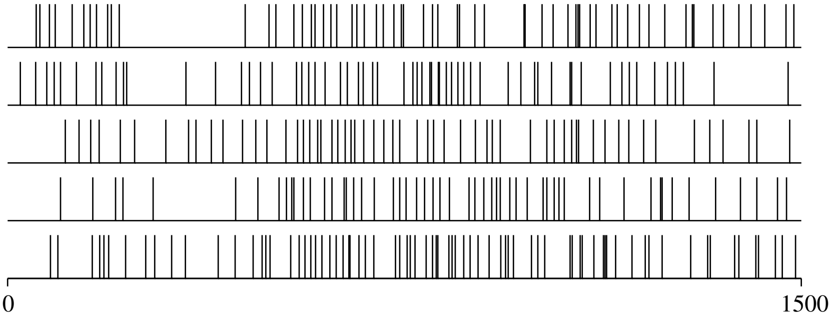

Fig. 2.2. Variability in the Response of a Cortical Neuron. A visual stimulus is repeatedly presented to the same neuron in the cortex of an awake macaque monkey. Each line corresponds to an action potential, with time (in msec) running from left to right. The stimulus was turned on at 0 msec. Notice the high degree of trial-to-trial variability in the detailed timing of the action potentials. This lack of reproducibility is one justification for a mean rate code in which only the average firing rate matters. From Koch 1999. Data from W. Newsome and K. Britten.

Randomness is also apparent at the behavioral level, where it manifests as *spontaneity.* Take a bevy of genetically identical fruit flies *Drosophila M.,* hatched at the same time, fed the same food, living in identical housing, subject to the same 12-hour-on, 12-hour-off light-dark cycle; such regimented control is far beyond anything even the most inhumane experimentalist could exert over identical twins. Yet the flies will still act capriciously. When confronted by a choice, most flies might turn one way, a minority will turn the other, while one or few will do something altogether different. This well-known propensity is enshrined in the adage: "Under carefully controlled experimental circumstances, an animal will behave as it damned well pleases."

This variability has been well studied in the fly flight simulator, in which the fly is tethered by a wire suspended from a torque meter. In this stationary flight, the animal can only turn left and right while its visual panorama is varied (Heisenberg & Wolf 1984; Maye et al. 2007). Variability shows up in unpredictable "body saccades," in which the animal makes a sudden turn, akin to an eye movement in humans. This can be thought of as a form of voluntary behavior. In the absence of any structured visual object, when the fly is surrounded by nothing but white walls, the animal stochastically executes these saccades following a fractal pattern. That is, the animal behaves neither completely randomly nor fully deterministically, but opts for something in between chance and necessity. It is well known that in complex environments where food or mates are distributed at unexpected locations, a pseudo-random search strategy is optimal (Viswanathan et al. 1999). Without concerning themselves with the question of the ultimate origin of this variability, Heisenberg and Wolf (1984) treat such spontaneous actions as voluntary behavior, akin to willed action in humans.

And what applies to organisms with a mere 100,000 neurons is also true for us, with vastly larger brains, containing on the order of 50,000,000,000 neurons. Nervous systems are indeterministic. Whether or not this indeterminism is grounded in quantum mechanics remains an open question. Your actions are not, and never will be, predictable. Even though the universe and everything within it obeys natural laws, the state of the future world is contingent in a way that, in general, cannot be computed from its current state.

Truman Capote's *In Cold Blood* is a chilling literary account of the senseless slaying of a farmer, his wife, and two children by two ex-convicts. The decision to murder the entire family appeared spontaneous, taken on the spot, without any compelling rationale, "just like that." It is easy to imagine that the criminals could have fled without committing this atrocity (for which they were later hanged). How many of life's critical choices are determined by such thought-less, possibly truly random, acts?

6 The Cognitive Science of Willful Intention

Let me describe a classical experiment by Libet and colleagues (Libet et al. 1983) that convinced many that free will must be an illusion.

The brain is ceaselessly active. One way to visualize this is to record the small fluctuations in the electrical potential on the scalp. Sometimes called brain waves, their amplitude is a minute fraction of a Volt. They can be measured by placing electrodes onto the skin of the head, a procedure known as electroencephalography (EEG). Every voluntary action, such as kicking the leg or turning the head, is accompanied by a slowly rising electrical potential that can best be recorded on the crown of the head. Called the "readiness potential," it precedes the actual onset of the motion by up to one second and was described in the mid 1960s (Kornhuber & Deecke 1965).

The readiness potential reflects neural activity in the motor planning and execution stages of the brain. Intuitively, therefore, the sequence of events must be as follows – first you consciously will to move your hand. Your brain translates that intention into electrical activity of neurons in motor cortex and elsewhere; this activity is then relayed to the motorneurons in your spinal cord that ultimately cause the muscles in your hand to contract. The mind decides and commands the brain to act, like the driver is in charge of her automobile, right?

Libet was not convinced that this was the actual sequence of events. He wanted to know what comes first: the conscious decision to move the hand or the onset of the brain's readiness potential? After millennia of learned debate, finally a question with an answer that can be obtained in a relatively straightforward manner.

Libet recruited subjects and asked them to spontaneously flex their wrist, whenever they felt like it, while their brainwaves were recorded with a simple EEG instrument. Participants looked at a screen where a bright light moved along a circular trajectory – like the pointer on a clock – as an aid for them to note when they first became aware that they wanted to move their wrist ("the light was at the 1:00 o'clock position when I decided to move my hand"). To confirm how well subjects could judge time, they had to indicate where the light was when they started to actually flex their hand. This time can be accurately, and objectively, established by measuring muscular activity with another electrode. The volunteers were, indeed, quite accurate in their judgments of onset of muscular motion. So it is likely that they were equally accurate in judging onset of the conscious decision to move the hand.

What became apparent was that the beginning of the readiness potential *preceded* the conscious decision to move by 0.3 and 0.5 sec. That is, the brain acted before the conscious mind did! This is a complete reversal of the deeply held intuition of mental causation – your brain and your body only act after your mind wills it. That is why this experiment was, and remains, controversial. But it has been refined in a number of ways over the intervening years, and its basic conclusion stands (Haggard & Elmer 1999; Haggard 2008). Recently, a fMRI variant of such an experiment was carried out in which subjects had to move either their left or their right hand. Hemodynamic activity in parietal and prefrontal cortex predicted which hand would be used up to 8 seconds prior to the actual onset of movement (Soon et al. 2008). The brain starts to act before the conscious mind decides. Somewhere in the brain's catacombs, possibly in the basal ganglia, a decision to move is made, say because some threshold has been spontaneously exceeded. Although it is your brain, your conscious mind does not know yet that a decision has already been taken. Signals are then sent to motor and premotor cortices, and those sectors of the brain come online, preparing to activate the relevant muscle groups in the necessary sequence. This furious activity of cortical nerve cells shows up outside the skull as the electrical readiness potential.

7 Agency or the Conscious Experience of Will

The experience of having consciously decided to move is generated much later, most likely by networks in the medial premotor and anterior cingulate cortices in the frontal lobe. Here, the appropriate neuronal activity triggers the feeling of willing an action, of being an autonomous agent who causes things to happen. Psychologists refer to this conscious experience as *intention.* When the action is actually executed, subjects experience the distinct conscious sensation of *agency* (Haggard 2008).

Take note here, for this is a radical idea – that the mind-brain nexus creates a specific conscious sensation for willing some behavior, a compelling experience of "I intended this action and made it happen." Like any other experience, the experience of being the cause, of feeling responsible for the action of one's body has subjective content, has qualia associated with it.

The critical point to remember in the context of the free-will debate is that the neuronal activity associated with the feeling of agency is only triggered after the actual decision has already been taken by some neural network. The conscious mind does not cause the action to come about. It is more in the nature of a marker for voluntary action, an afterthought.

The psychologist Wegner manipulates the sense of agency in unexpected ways (Wegner 2003). Depending on circumstances, his undergraduate subjects can be made to feel more or less responsible for actions that they undertook or wrongly ascribe a willful action to themselves that they manifestly were not responsible for.

For instance, Wegner had a volunteer dress in a black smock and white gloves and stand in front of a mirror. The subject's arms were hanging by her side. Directly behind her stood a confederate, dressed in a similar black smock and white gloves. His arms reached under the arms of the subject directly in front of him, so that when the subject looked into the mirror, she saw a pair of arms and gloved hands, much like her own. Both wore headphones through which Wegner could give instructions. For instance, the confederate would be told to clap his hands three times. When the subject could hear instructions previewing each movement, she reported an enhanced feeling of controlling the hands (even though they were not hers). This experience of agency did not occur when the instructions occurred after the movement had already taken place.

According to Wegner, the sense of agency is a psychological module that automatically and unconsciously assigns authorship to certain actions based on simple rules. If somebody tells me to snap my fingers and I look down and see what looks like my fingers snapping, it is not unreasonable to conclude that I was responsible for this action. Imagine you are walking alone through a forest and you hear a twig break. If this sound came just after you stepped on a branch, you are relieved as your agency module assumes that you are responsible for the sound and all is well. But if the sound occurred before you stepped on the branch, something or somebody might be following you and all of your senses will switch into high alert.

The experience of agency is a subjective sensation with an associated quale, no different in kind from the conscious experience of seeing red or tasting bitter almond. Like other percepts, it has a trigger – here an internal action rather than an event in the world. There are visual illusions in which one's visual percept does not correspond to what is really out there. Such illusions also occur with the sensation of agency: a movement may seem unwilled while another – see the above experiment – is experienced as willed even though somebody else caused it.

In well-practiced actions – rapid sensory-motor behaviors I refer to as zombie agents – willful experience may be reduced. This is certainly true for the involuntary reflexes – your pupil automatically constricts when a bright light is suddenly turned on or your hand shoots out to steady you on a slippery walkway. You forcefully exert your will, like an inner muscle, to overcome the fear of climbing past the exposed crux section. But once you are underway, your body manages quite well on its own, without you exerting any further will.

In automatism, the sense of agency may be missing all together. Well-studied examples are possession and trance in the context of religious ceremonies, post-hypnotic suggestions, Ouija board games, or divining, dowsing and similar pseudo-occult phenomena. Typically, participants will vehemently deny that they are responsible for actions that are provably theirs (Wegner 2003).

Mental diseases can lead to overt pathologies that likewise stunt the experience of will. The spectrum ranges from the lung-cancer patient who sneaks out of the hospital to smoke, to the drug addict who turns criminal to finance his habits, to the obsessive-compulsive who needs to wash her hands so often that they bleed, or to criminals with aggressive impulse-control disorders (Hollander & Berlin 2008). Intriguingly, most, if not all, obsessive-compulsive disorder patients realize that their behavior is pathological, is "crazy," a knowledge that causes them much distress. The same is true of some patients with frontal lesions who cannot control their impulsive behavior even though they know they are acting inappropriately (Berlin et al. 2004). Clearly, these patients have lost their freedom of choice.

Social psychology has learned that even if people firmly state, and believe, that they are undecided about some matter, for example, which presidential candidate they will be voting for in the upcoming election, their future behavior can be predicted quite well from their unconscious attitudes (Galdi et al. 2008). In some sense, they have already made up their mind, even though they have no attendant feeling of agency. When reading this literature, I am struck by how little insight humans have into how and why we act and decide.

Let me describe a final experiment, known as "Choice Blindness." More than one hundred volunteers were shown two photographs, each of a woman's face. After looking at both pictures for a few seconds, they had to choose the one that looked most attractive to them. Immediately after three such choices, subjects were shown again the face they had just chosen and were asked to explain their choice. They readily complied. On three other trials, the experimentalist, in a sleight of hand, exchanged the picture of the chosen woman with the opposite image. That is, immediately after deciding that woman A was more attractive, a double-card ploy was used to confront subjects with the picture of woman B and they had to explain why they chose her (the two women depicted on the photos were quite distinct). Remarkably, most of the time the subjects were fooled. Only in fewer than 25% of trials were participants aware that their original choice was not honored, that they had been fooled. Most of the time, they blithely ignored the discrepancy between their original conscious decision and what they were told they had decided. And even more remarkably, they proceeded to justify this choice even though it contradicted what they actually did a few seconds earlier: "She's radiant. I would rather have approached her at a bar than the other one. I like her earrings," even though the original choice looked solemn and had no earrings.

What choice blindness reveals is that people often have no idea why they choose the way they do. But their urge to explain their actions is such that this does not prevent them from making up a story on the spot, confabulating without knowing it.

8 Taking Stock of the Situation

Let me summarize. Classical physical determinism is out; the future is not fully determined by the current facts. Quantum mechanics teaches that randomness is inherent in the basic structure of the universe. There is always some probabilistic aspect to nature. Indeterminism implies that what you do is not fully determined by the past. The future is literally an open book; while the letters on the page you are reading right now are clearly visible, it is more and more difficult to be certain of the text in the following pages that foretell what happens next. They become progressively fuzzier and illegible.

The combination of quantum mechanics and deterministic chaos limits the accuracy and range of predictions that even the best informed neuroscientist of the future will be able to make about any one individual. Some behaviors will always appear stochastic, spontaneous, uncaused.

The strong, Cartesian version of free will, the belief that if you were placed in exactly the same circumstance, you could have acted otherwise, is difficult to defend given our current, possibly very limited, understanding of the brain-mind nexus. The trouble is to account for how the conscious mind, the refuge of the classical soul, could influence the brain without leaving telltale signs. Anything in the world happens for one or more reasons that are also part of the world; the universe is causally closed. All the mind could accomplish is to realize one of several quantum mechanical possibilities, without being able to do anything about the underlying probabilities. In particular, it could not make one outcome more likely and another one less likely. This is a meager freedom. Furthermore, the action of a truly free will could never be distinguished from a random choice.

Libet first compellingly demonstrated that the brain can make a simple decision well before the conscious mind does; his observation reveals the experience of willing an action to be secondary to the actual cause. The sense of agency, of feeling responsible for an act, shares with other, more sensory, forms of conscious experience phenomenal content, qualia. Psychological experiments and various psychiatric patients expose the reality of this aspect of our mind. And common to other conscious experiences, the actual workings of the sense of agency – why we choose the way we do – is opaque, hidden from conscious access. Like vision, the sense of agency often fails us in unexpected ways, creating illusions of will.

Common to all these subjective states is the fundamental mystery of how bioelectrical activity in a restricted part of the brain give rise to these ineffable experiences. In particular, what are the neuronal correlates of willful conscious experience?

A number of crucial scientific, moral, and practical questions remain. Where do our deliberations leave personal responsibility? I, for one, certainly believe that your brain must be held responsible for your actions. But I also think that the notion of responsibility for dysfunctional or for criminal behavior must be modified in light of the facts discussed here. How can families and society at large cultivate good habits that permit individuals to make wise decisions?

Another aspect of volition is the phenomenon of will power. While I can run relatively effortlessly for hours in the mountains, I cannot resist the temptation of a Mousse en chocolate. Is it a simple matter of not having enough will power to eat less? Or a question of proper motivation? Is will power something that can be trained, as many meditation and self-help gurus claim? Does it have a genetic basis? All of these are questions that demand answers.

I would like to end with a plea for humility. Humility because even though we are living in the age of science, we know so little. The cosmos is a strange place. Take the decade-old discovery that only four percent of the mass-energy of the universe is the sort of material out of which stars, planets, trees, you, and me are fashioned. One quarter is cold dark matter while the rest is something bizarre called dark energy. Cosmologists have no idea what exactly this is nor what laws it obeys. It is exceedingly strange stuff and cannot be seen. Is there some ephemeral connection between this spooky stuff and consciousness, as suggested by the novelist Philip Pullman in his trilogy "His Dark Materials"? Very likely not; but who is to say for certain. Our knowledge is only a fire lighting up the vast darkness around us, flickering in the wind. So let us be humble and be open to alternative, rational explanations.

References

Atkins, P.W.: The second law. Scientific American Books, New York (1984)

Beck, F., Eccles, J.C.: Quantum aspects of brain activity and the role of consciousness. Proceedings of the National Academy of Sciences of the United States of America 89, 11357–11361 (1992)

Berlin, H.A., Rolls, E.T., Kischka, U.: Impulsivity, time perception, emotion and reinforcement sensitivity in patients with orbitofrontal cortex lesions. Brain 127, 1108–1126 (2004)

Everett, H.: Relative state formulation of quantum mechanics. Reviews of Modern Physics 29, 454–462 (1957)

Galdi, S., Arcuri, L., Gawronski, B.: Automatic mental associations predict future choices of undecided decision-makers. Science 321, 1100–1102 (2008)

Garnsey, P.: Ideas of slavery from Aristotle to Augustine. Cambridge University Press, New York (1999)

Haggard, P.: Human volition: Towards a neuroscience of will. Nature Reviews Neuroscience 9, 934–946 (2008)

Haggard, P., Eimer, M.: On the relation between brain potentials and the awareness of voluntary movements. Experimental Brain Research 126, 128–133 (1999)

Hameroff, S., Penrose, R.: Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness. In: Hameroff, S., Kaszniak, A., Scott, A. (eds.) Toward a science of consciousness: The first Tucson discussions and debates. MIT Press, Cambridge (1996)

Heisenberg, M., Wolf, R.: Vision in Drosophila: Genetics of microbehavior. Springer, Berlin (1984)

Hille, B.: Ion channels of excitable membranes, 3rd edn. Sinauer Associates, Sunderland (2001)

Hollander, E., Berlin, H.A.: Neuropsychiatric aspects of aggression and impulse-control disorders. In: Yudofsky, S.C., Hales, R.E. (eds.) Neuropsychiatry and behavioral neurosciences, pp. 535–565. American Psychiatric Publishing, Washington DC (2008)

Johansson, P., Hall, L., Sikström, S., Olsson, A.: Failure to detect mismatches between intention and outcome in a simple decision task. Science 310, 116–118 (2005)

Jordan, P.: Die Verstärkertheorie der Organismen in ihrem gegenwärtigen Stand. Naturwissenschaften 26(33), 537–545 (1938)

Kane, R. (ed.): The Oxford handbook of free will. Oxford University Press, New York (2002)

Koch, C.: Computation and the Single Neuron. Nature 385, 207–211 (1997)

Koch, C.: Biophysics of computation: Information processing in single neurons. Oxford University Press, New York (1999)

Koch, C.: The quest for consciousness. Roberts Publishing, Denver (2004)

Koch, C., Hepp, K.: Quantum mechanics and higher brain functions: Lessons from quantum computation and neurobiology. Nature 440, 6161–6162 (2006)

Koch, C., Hepp, K.: The relation between quantum mechanics and higher brain functions: Lessons from quantum computation and neurobiology. In: Chiao, R.Y., Cohen, L.M., Leggett, A.J., Phillips, W.D., Harper Jr., C.L. (eds.) Amazing light: Visions for discovery: New light on physics, cosmology and consciousness, Cambridge University Press, New York (2009)

Kornhuber, H.H., Deecke, L.: Hirnpotentialänderungen bei Willkürbewegungen und passiven Bewegungen des Menschen: Bereitschaftspotential und reafferente Potentiale. Pflügers Archiv Physiologie 284, 1–17 (1965)

de Laplace, P.S.M.: A Philosophical Essay on Probabilities. Translated from the 6th French edn., Truscott, F.W., Emory, F.L. (eds.). Dover Publications, New York (1951)

Libet, B., Gleason, C.A., Wright, E.W., Pearl, D.K.: 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 (1983)

Lorenz, E.N.: The essence of chaos. University of Washington Press, Seattle (1995)

Maye, A., Hsieh, C.-H., Sugihara, G., Brembs, B.: Order in spontaneous behavior. PLOS One 5, e443 (2007)

Penrose, R.: Shadows of the mind. Oxford University Press, Oxford (1994)

Popper, K.R., Eccles, J.C.: The self and its brain. Springer, Heidelberg (1977)

Shor, P.W.: Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. SIAM. Journal on Computing 26(5), 1484–1509 (1997)

Siegel, C.L., Moser, J.K.: Lectures on celestial mechanics. Springer, New York (1971)

Soon, C.S., Brass, M., Heinze, H.-J., Hayner, J.D.: Unconscious determinants of free decisions in the human brain. Nature Neuroscience 11, 543–545 (2008)

Stapp, H.P.: Mind, matter and quantum mechanics. Springer, Berlin (2003)

Strassberg, A.F., DeFelice, L.J.: Limitations of the Hodgkin-Huxley formalism: Effects of single channel kinetics on transmembrane voltage dynamics. Neural Computation 5, 843–855 (1993)

Strogatz, S.H.: Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering. Da Capo Press (2000)

Sussman, G.J., Wisdom, J.: Numerical evidence that the motion of Pluto is chaotic. Science 241, 433–437 (1988)

Tegmark, M.: Many lives in many worlds. Nature 448, 23–24 (2007)

Turner, M.S.: Large-scale structure from quantum fluctuations in the early universe. Philosophical Transactions of the Royal Society of London: Series A, Mathematical, Physical and Engineering Sciences 357, 7–20 (1999)

Viswanathan, G.M., Buldyrev, S.V., Havlin, S., da Luz, M.G., Raposo, E.P., Stanley, H.E.: Optimizing the success of random searches. Nature 401, 911–914 (1999)

von Neumann, J.: Mathematische Grundlagen der Quantenmechanik. Springer, Berlin (1932); English translation: Mathematical foundations of quantum mechanics. Princeton University Press, Princeton (1955)

Wallace, A.B.: Contemplative science: Where Buddhism and neuroscience converge. Columbia University Press, New York (2007)

Wegner, D.M.: The illusion of conscious will. MIT Press, Cambridge (2003)

Wigner, E.P.: Symmetries and reflection: Scientific essays of Eugene P. Wigner. Indiana University Press, Bloomington (1967)

3

Human Freedom and “Emergence”

William T. Newsome

Department of Neurobiology
Stanford University School of Medicine
Fairchild Building, Room D200
Stanford, CA 94305
bill@monkeybiz.stanford.edu

Summary. Whether free will is a reality is an increasingly urgent problem, both from a scientific and a social point of view. An ability to make judgments and take actions that are “free” in some meaningful sense would seem a prerequisite for the process of scientific reasoning and for our ability to behave morally. How are we to reconcile the “autonomy” of a reasoning intellect with our scientific conviction that all behavior is mediated by mechanistic interactions between cells of the central nervous system? It seems that answers will ultimately lie in a deeper understanding of emergent phenomena in complex systems. This will help enrich our impoverished standard notions of causation in physical systems.

Keywords: free will, emergence, neuroscience, neural networks, casuality.

The question of human freedom is a vexing point of tension between humane and scientific worldviews. What are we to make of human freedom when, from a scientific point of view, all forms of behavior are increasingly seen as the causal products of cellular interactions within the central nervous system, which themselves are substantially influenced by the toss of genetic dice that occurred when each of us was conceived?

To frame the issue in an everyday context, can I really “choose” to have fish or chicken for dinner this evening, or do events already in motion restrict me to a predetermined course of action? And if our sense of choice is in fact illusory, can anyone reasonably be held responsible for his or her actions? A foundational assumption of our legal system is that individuals have a meaningful degree of freedom to choose between alternative behaviors. Our great religious traditions presume similarly: we can make loving, sacrificial choices in how we interact with others, or we can act in ways that are exploitative, or at worst overtly hateful and destructive. William Provine, biologist and historian of science has said, “There is no way that the evolutionary process as currently conceived can produce a being that is truly free to make moral choices” (Provine 1988).

Provine states the challenge in stark terms, and much hinges on our response to the challenge. Can we develop an understanding of “freedom” that is consistent with contemporary neuroscientific understanding of the brain and behavior?

The issue of human freedom is a tricky one. Some modern thinkers find refuge from strict determinism in quantum mechanics, which describes events probabilistically rather than deterministically. Quantum mechanics implies that we live in a fundamentally unpredictable world. Consider, for example, a classic quantum mechanical phenomenon: the absorption of photons by matter. Absorption of high energy photons by DNA in my skins cells can result in genetic damage and fatal cancer, irreversibly changing the course of my life and the lives of my family, friends, and colleagues. Yet the triggering events – photon absorptions – are fundamentally random and unpredictable, even in principle. To my mind, therefore, the model of a fully deterministic world can be set aside. Yet I am not yet convinced that quantum mechanics offers deep insight into human freedom. It is not clear to me that randomness provides an understanding of human freedom that is any more meaningful than that of strict determinism. In contrast to strictly deterministic and quantum mechanical views, our intuitive understanding of human freedom is that we have some meaningful degree of autonomy, or self-determination. While we are certainly influenced by random events (in the quantum mechanical sense) and by strictly determined events (in the Newtonian sense), we are at the complete mercy of neither.

Some of my scientific colleagues seem to feel that the notion of human freedom must be tolerated as a practical matter in order to maintain a functioning society, but that human freedom is likely to prove illusory in the final analysis. From this perspective, brains are extremely complex neurochemical machines, and their behavior will ultimately be understood in the same mechanical terms in which any other machine is understood. While notions of human freedom are convenient and probably even necessary to get along in everyday life, our subjective experience of freedom itself is no more than the result of machine-like activity within specific regions of the central nervous system.[1]

What this point of view fails to realize, however, is that the sense of human freedom, or autonomy, is just as important for scientific understanding as for everyday understanding of the world. Thoroughgoing determinism becomes entangled in profound logical difficulties in science no less than in everyday life. J.B.S. Haldane put the matter succinctly: “If my mental processes are determined wholly by the motions of the atoms in my brain, I have no reason to suppose that

[1] But as Charles Jennings has observed, throw a rock through the living room window of the most reductionistic neurophilosopher and you will probably find out just how quickly the dispassionate notion of behavioral determinism evaporates! (Editorial 1998).

my beliefs are true … and hence I have no reason for supposing my brain to be composed of atoms” (Haldane 2002, p. 209).

Haldane’s point is that the entire enterprise of science depends upon the assumption that scientists have freedom to evaluate evidence rationally and make reasoned judgments about the truthfulness of particular hypotheses and results. If, however, the scientist’s rational judgments, and her/his beliefs about the validity of the scientific method, simply reflect an inevitable outcome of the atomic, molecular, and cellular interactions within a particular physical system, how can we take seriously the notion that her/his conclusions about the world bear any relation to objective truth? (Ironically, the ardent determinist becomes an intellectual bedfellow of the ardent deconstructionist.) And if we cannot believe that the scientific approach leads to some approximation of truth, how can we take seriously the scientifically based assertion that mechanical determinism is the correct way to think about the world? The attempt to adopt a thoroughgoing determinism is like sawing off the branch that one is sitting on; the result is intellectual freefall. Like it or not, then, achieving a meaningful understanding of human freedom is profoundly important for science, for society, and for each individual person.

How are we to reconcile the “autonomy” of a reasoning intellect with our scientific conviction that all behavior is mediated by mechanistic interactions between cells of the central nervous system? Although I have no certain answer to this question, I suspect that answers will ultimately lie in a deeper understanding of emergent phenomena in complex systems. Emergence is a somewhat slippery concept and has been used in different ways by different authors.[2] By “emergence,” I mean that complex assemblies of simpler components can generate behaviors that are not predictable from knowledge of the components alone and are governed by logic and rules that are independent of (although constrained by) those that govern the components. Furthermore, the intrinsic logic that emerges at higher levels of the system exerts “downward control” over the low-level components. To foreshadow my ultimate argument, it is the phenomenon of downward control that endows a system with a behavioral autonomy, which in the case of biological organisms can be regarded as meaningful choice.

It is critical to be very clear on one point at the outset: the concept of “emergence” does not imply magic or mysticism. My discussion here will not invoke brain events that violate known physical principles. More than anything else, this reflects my biological intuition that the human brain, as a product of the natural evolution of the universe in general and life on earth in particular, operates in a manner consistent with (i.e., constrained by) known physical laws. It is certainly conceivable, and perhaps even likely, that that some aspects of human and animal consciousness will never be satisfactorily understood from the point of view of the

[2] There is a large literature, both formal and informal, on the theme of emergence in complex systems; for recent examples, see Clayton & Davies 2006; Clayton 2006. Also see the excellent review by Timothy O’Connor in the online *Stanford Encyclopedia of Philosophy* at http://plato.stanford.edu/entries/properties-emergent/.

reductive sciences (see, e.g., Nagel 1974), but one does not want to throw in the towel until forced!

Many authors have cited examples of emergent behavior in complex systems, a favorite example being the unicellular organism (for beautiful examples of emergence in the context of physics, see Laughlin 2005). The existence of unicellular organisms permits an enormous number of new phenomena that could not be predicted from knowledge of macromolecules alone and that operate on principles that go well beyond those that govern macromolecules: cellular motility, foraging for resources, competition with other organisms, and adaptation to environmental pressure by means of mutation, to name but a few. Each of these phenomena must be identified and described in-and-of-themselves, and their internal logical rules worked out, before rigorous links can be made to lower-level mechanisms. Competitive interactions between species, for example, are comprehended by observation at the behavioral level, not by inference from the molecular level. The behavior of the unicellular organism, in turn, exerts downward control over its constituent molecules. The motion of an organelle within the cell depends in one sense on pressure exerted from the cytoplasm as the organism moves. But in another, equally valid sense, the motion of the organelle depends upon the immediate behavioral goal of the organism. As far as we know, nothing about the life of unicellular organisms violates the laws of physics or the chemical laws that govern the behavior of macromolecules. The cell cannot behave in any way that is not permitted by the lower levels of organization of its constituent parts; the behavior of the cell is thus *constrained* but not *determined* by the lower levels.

Obviously, the crucial distinction here is between the words “constrained” and “determined.” This distinction comes into clear relief for me when considering the operation of the computer program that is running right now on my laptop computer. If I want to understand how Microsoft Word operates, I can tackle the problem at the mechanistic level of transistors, resistors, capacitors, and power supplies, or I can tackle the problem at the level of the software – the logical instructions that lie at the heart of the process of computing. It seems clear to me that the most incisive understanding of Microsoft Word lies at the higher level of organization of the software. One wants to understand the logical relationships that comprise computation: for-loops, if-statements, and the like. The logic of the computation exists independently of the physical system of electronics that make up the computer (the software can be transferred to another computer) and operates according to its own rules that cannot be predicted from knowledge of the hardware alone. The rules of computation logic, in turn, orchestrate (in a real, causal sense) the currents flowing through the myriad individual components that comprise the computer. Again, nothing magical or mystical is occurring here. The software is constrained by the hardware; the software cannot abrogate the laws of physics nor the principles that govern the behavior of electronic circuits. Nevertheless, the behavior of the computer as I type this manuscript is *determined* at a higher level of organization – the software – not by the laws of physics or the principles of electronic circuitry.

Although this computer example emphasizes the critical distinction between "constraint" and "determination," it is *not* an example of emergence because the software did not evolve from a natural process of self-assembly but was designed by human programmers. A better example of emergence in the computing world lies in the relatively new field of neural networks. In the neural net illustrated in figure 3.1A, multiple layers of "neuron-like" computing units are linked to each other in a hierarchical manner such that the behavior of each unit in a lower layer influences each unit in the next highest layer (arrows). The strength of the influence of any given lower-level unit upon units in the next higher level is governed by a set of "weights" that determines the effectiveness of the link between each pair of units. In the initial state of the network, the weights governing the many links are chosen randomly; some are positive, some are negative; some are strong, some are weak. An input is then provided to the lowest level of the network, and an output emerges at the highest level. In a backpropagation network (one of several types of neural nets), a software entity called a "teacher" then recognizes whether the actual output is similar to the desired output and adjusts all of the weights of the links between computing units accordingly. After many iterations of the input-output-adjustment cycle, the network "learns" to produce the correct output for a given input. (The backpropagation network is not particularly biological because it incorporates an independent "teacher" which orchestrates the manipulation of the connections between units. More recent neural nets, accomplish the same goal in biologically plausible ways.)

Neural networks can perform remarkable feats that are extremely difficult to accomplish by traditional computing methods that employ mathematically precise algorithms specified by a programmer. Some of the most impressive examples lie in the arenas of voice and pattern recognition and of robotics. Yet a remarkable intellectual quandary is often encountered in the neural network field: a network can be trained to solve a fiendishly difficult problem, and in the end, the human programmer who designed the network and invented the training rule may have little or no insight into *how* the problem has actually been solved! The programmer can show us the final pattern of weights between the individual computing units that somehow embodies the solution (figure 3.1B, for example), but we frequently remain embarrassingly ignorant concerning the algorithmic principle(s) the network has "discovered" in solving the problem.[3]

This example comes closer to the meaning of emergent order in complex systems. At a "low" level we know everything there is to know about the neural network and the digital computer on which it runs. We fully understand the physical principles underlying operation of the computer as well as the learning algorithm that enables the network to modify its connections as it interacts with the environment. Furthermore, at the end of the learning exercise the programmer has full knowledge of the learned connection weights, and s/he may transmit the "solution"

[3] A computationally savvy colleague of mine at Stanford refers to these networks, with a mixture of humor and derision, as "know-nothing networks" because at the end of the exercise, the scientist still may not *understand* the solution that has been achieved.

in the form of connection weights to anyone in the world who would like to implement it for their own purposes. This point cannot be emphasized too strongly: at a mechanistic level there is no causal gap in our understanding; we know *everything* that matters about the neural network – both its final state and precisely how it got there. Paradoxically, however, we are frequently unable to state, or write an equation for, the algorithmic principle that lies at the heart of the learned solution. Our situation resembles that of an electronics assembly technician who can solder components together to create a functioning, causally complete, electronic circuit, yet has little or no idea how the thing actually works at a high level.

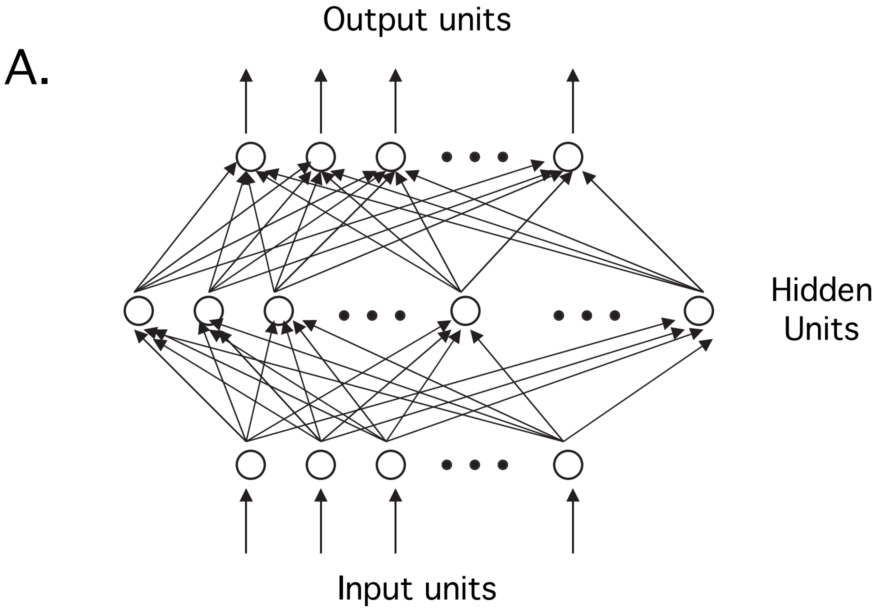

Fig. 3.1. Schematic diagram of a common multilayer neural network

A. Network architecture. Each circle represents a computing "unit." The units are arranged in three hierarchical layers: from bottom to top, "input" layer, "hidden" layer, and "output" layer. Signal flow is "feedforward" in the sense that a given layer exerts causal influences only on the next highest layer. Each unit in a given layer influences the activity of each unit in the next layer as illustrated by the arrows. The initial strengths, or "weights" of the connections between units are random. Some are positive (activity in the "sending" unit increases activity in the "receiving" unit) and others are negative. Some weights are strong (the sending unit has a large impact on the receiving unit), while others are weak. These weights are adjusted during the learning process according to the similarity of the actual outputs to the desired outputs. Diagram adapted from Rummelhart et al. 1986.

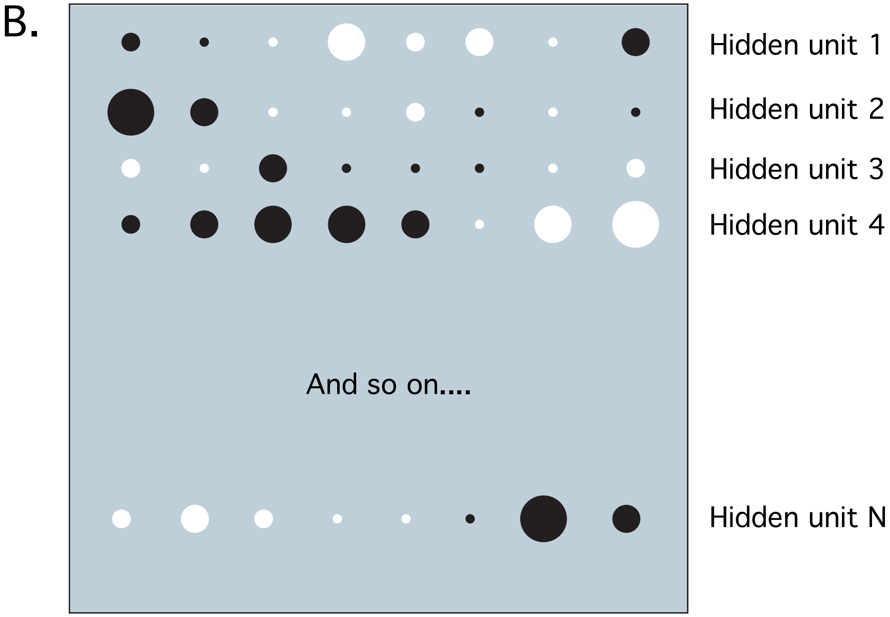

Fig. 3.1. (*continued*)
B. After the learning process, the final configuration of the network, which embodies the learned solution to the problem, is depicted as the final set of weights between the various units. If, for example, there are 8 input units and 20 hidden units in the network in A, the weights of the connections between each input unit and each hidden unit can be depicted as in B. White dots indicate positive weights and block dots depict negative weights. The size of the dot is proportional to the strength, or weight, or the input. The top row of dots depicts the weights from all 8 input units onto hidden unit one, and the second row depicts the weights from all 8 input units onto hidden unit two. The rows are iterated until the weights to all 20 hidden units are represented. A similar diagram (not shown) depicts the weights from each hidden unit onto each output unit. These "Hinton" diagrams (named after their originator, G.E. Hinton) fully describe the final state of the network and can be reproduced at will on any suitable digital computer.

This is a somewhat humiliating situation for a scientist to be in – understanding a system completely at a "low" level, while being quite ignorant of how it operates at a "high" level. Most of us feel intrinsically that we *must* understand the higher level of organization, which in the case of neural networks involves formal computational logic, to be intellectually satisfied with the result. One possible reaction to this dilemma is to deny that any "higher level" exists in the network. If we know the transfer function of each individual computing unit and the weights of all the connections, we can calculate the output for any given input and there is nothing else to know scientifically. For me, this is not a sustainable point of view. It brings to mind Thomas Nagel's observation: "To deny the reality or logical significance of what we cannot describe or understand is the crudest form of cognitive dissonance" (Nagel 1974).

How are we to reason about such a paradoxical state of affairs? In the case of the electronics technician, the answer is clear: the technician simply follows a design created by another intelligence – the circuit engineer. The circuit engineer is not imaginary or epiphenomenal, but rather is a critical locus of "downward" causal control in producing a functioning circuit. In the case of the neural network, end-users of the network exploit a design created during a learning interaction between the network and its environment. As is typical of systems that learn, the actual structure of the network changes as a result of new information acquired during the learning process, and the new (emergent) structure of the network embodies the learned solution to the problem. The higher-order interactions of a complex system formally resemble Darwinian selection mechanisms: hugely variable events in the word impact each organism through the selective filter of the organism's behavioral goals. The interaction of goals and selection during the learning process create nonreducible, high-level information in biological systems (see Ellis 2006a, 2006b). As with our circuit engineer, the intelligent solution embodied in the emergent structure of the network is not imaginary or epiphenomenal, but rather is a critical locus of downward causal control, implementing very practical solutions to complex problems.

Having wandered a bit from my original topic, let me now state exactly how my "toy" example of neural networks is – and is not – relevant to understanding human autonomy, which I take to be the essence of freedom. The most relevant lesson is this: A complex system endowed with the ability to learn possesses the autonomy to discover solutions (to problems) that cannot be captured satisfactorily by, or predicted in advance from, lower-level descriptions including the learning algorithm itself.[4] Information embedded at higher organizational levels is the most important locus of causal control of the system. A skeptic might argue that this toy example provides no understanding of autonomy (or freedom) whatsoever because every aspect of the network, including each step of the learning process, is causally determined. Given the same original set of weights between the computing units, the same learning algorithm, and the same set of inputs from the environment, the network would produce exactly the same solution by exactly the same series of steps each time it was run. My reply to this objection – which should be clear by now – is that a breach of causality is not a requirement for "autonomy"; a central point in my discussion of neural networks is that their autonomy is real even though their function is entirely causal. At the most fundamental level, I am arguing that our standard reductionistic notions of causation in physical systems are impoverished – not wrong, simply impoverished. Neuroscientists cannot satisfactorily understand cognitive phenomena such as attention simply in terms of causal interactions between molecules, just as the computer scientist cannot satisfactorily understand the operation of the neural network simply in terms of the interaction weights of the units. At certain levels of complexity, the primary drivers of system behavior are

[4] As a computational neuroscience colleague at MIT once said to me: "If we could figure out the solution in advance, we wouldn't have to throw a network at the problem."

the logical rules of operation intrinsic to higher levels of the system; no other level of explanation accurately captures the nature of the system.

A Caveat

I emphasize that the neural net heuristic is only that – a heuristic. It allows us to appreciate important points about complex systems, but it does *not* necessarily provide deep insight into the nature of human cognition, per se. Human brains, and those of other animals as well, are vastly more complex than the neural nets that we employ in our most advanced sciences, and new phenomena with their own intrinsic logic will certainly emerge with every added level of complexity within the nervous system.

Of particular importance are the abilities of humans to reason with symbols and to reason recursively about our own reasoning (see Deacon 1997). With these evolutionary accomplishments, the relationship between our highest-level behaviors and the underlying “wetware” (ion channels, membranes, single neurons) becomes even more indirect. The relationship exists, of course, which is a major reason why neuroscientists such as myself have jobs. But the relationship is more a matter of *constraint* than of *generation*. As always, the biophysics of the constituent wetware constrains the phenomena that are possible at higher levels, but the behavioral possibilities that are actually realized are determined by higher order interactions of an organism with its environment.

Concluding Remarks

Although I have no elegant solution to the problem of free will, I believe that understanding human freedom is the most important and most difficult long-term challenge facing the neuro-behavioral sciences. Our freedom is certainly restricted by our biology – more so than most of us would like to admit. The remarkable studies of identical twins raised apart, for example, emphasize the pervasive influence of our genetic composition on surprisingly varied aspects of behavior from basic temperament to small behavioral tics; we are not free to escape many aspects of our genetic heritage (see, e.g., Kendler 1993; McClearn et al. 1997). However, a meaningful capacity for self-determination (autonomy) – which I consider to be the core notion in our conception and experience of free will – is an irreplaceable foundation for taking seriously the notions of scientific truth as well as individual moral responsibility.

A satisfactory understanding of this capacity will ultimately lie in the concepts of emergence and downward causality within complex systems. Emergent behaviors of even simple learning systems are often surprising and deeply perplexing, yet they can get in touch with realities whose deeper foundations we struggle to discern long after we accept the validity of the behavior. Thus “emergence” becomes a pivotal concept for interpreting the reality of human life in all its complexity, from scientific endeavor to personal morality to religious understanding. Although emergence is a notoriously difficult phenomenon to study rigorously, few areas of study are likely to prove as intellectually and practically consequential in the long run.

Acknowledgments

This chapter is adapted from an earlier paper that was written for *Visions of Discovery: New Light on Physics, Cosmology, and Consciousness,* an edited volume honoring the ninetieth birthday of Charles Townes, Professor Emeritus of Physics at the University of California at Berkeley (in press, Cambridge University Press). I am greatly indebted to Professor George F.R. Ellis of the University of Capetown for extremely helpful contributions during preparation of this chapter.

References

Clayton, P.: Mind and emergence. Oxford University Press, Oxford (2006)

Clayton, P., Davies, P.C.W. (eds.): The re-emergence of emergence. Oxford University Press, Oxford (2006)

Deacon, T.: The symbolic species: The co-evolution of language and the human brain. Penguin, London (1997)

Editorial. Nature Neuroscience 1, 535–536 (1998)

Ellis, G.F.R.: On the nature of emergent reality. In: Clayton, P., Davies, P.C.W. (eds.) The re-emergence of emergence. Oxford University Press, Oxford (2006a)

Ellis, G.F.R.: Physics and the real world. Foundations of Physics, 1–36 (April 2006b), http://www.mth.uct.ac.za/~ellis/realworld.pdf

Haldane, J.B.S.: Possible worlds and other essays. Transaction Publishers, London (2002[1927])

Kendler, K.S.: Twin studies of psychiatric illness: Current status and future directions. Archives of General Psychiatry 50, 905–915 (1993)

Laughlin, R.B.: A different universe: Reinventing physics from the bottom down. Basic Books, New York (2005)

McClearn, G.E., Johansson, B., Berg, S., Pedersen, N.L., Ahern, F., Petrill, S.A., Plomin, R.: Substantial genetic influence on cognitive abilities in twins 80 or more years old. Science 276, 1560–1563 (1997)

Nagel, T.: What is it like to be a bat? Philosophical Review 83, 435–450 (1974)

O'Connor, T.: Emergent properties. Stanford Encyclopedia of Philosophy (2006), http://plato.stanford.edu/entries/properties-emergent/

Provine, W.: Evolution and the foundation of ethics. MBL Science 3, 25–29 (1988)

Rummelhart, D.E., McClelland, J.L., The PDP Research Group: Parallel distributed processing: Explorations in the microstructures of cognition, vol. 1. MIT Press, Cambridge (1986)

4

Top-Down Causation and the Human Brain

George F.R. Ellis

Mathematics Department
University of Cape Town
Rondebosch 7701, Cape Town
South Africa
George.Ellis@uct.ac.za

Summary. A reliable understanding of the nature of causation is the core feature of science. In this paper the concept of top-down causation in the hierarchy of structure and causation is examined in depth. Five different classes of top-down causation are identified and illustrated with real-world examples. They are (1) algorithmic top-down causation; (2) top-down causation via nonadaptive information control; (3) top-down causation via adaptive selection; (4) top-down causation via adaptive information control; and (5) intelligent top-down causation (i.e., the effect of the human mind on the physical world). Recognizing these forms of causation implies that other kinds of causes than physical and chemical interactions are effective in the real world. Because of the existence of random processes at the bottom, there is sufficient causal slack at the physical level to allow all these kinds of causation to occur without violation of physical causation. That they do indeed occur is indicated by many kinds of evidence. Each such kind of causation takes place in particular in the human brain, as is indicated by specific examples.

Key words: complex systems, hierarchy, causation.

1 Causation as the Core of Science

Physics is the basic science underlying physical reality, characterized by mathematical descriptions that allow predictions of physical behavior to astonishing accuracy. The key question is whether other forms of causation such as those investigated in biology, psychology, and the social sciences are genuinely effective, or are they rather all epiphenomena grounded in purely physical causation? The latter view is suggested by strong reductionist views based in the fact that all physical entities we see around us, including ourselves, are based in the same chemical elements composed from the same kinds of elementary particles, interacting with each other only through the four fundamental physical forces. How can there then be room for any other type of causation?

I will claim here that there are indeed other types of causation at work in the real world, described quite well by Aristotle's four types of causes. The overall framework for understanding these forms of causation and their interaction is the hierarchy of complexity (see table 4.1), ranging from particle physics and nuclear physics to astronomy and cosmology on the one hand, and to psychology and sociology on the other, with coarse-graining and consequent loss of detailed information relating each of the higher levels to lower levels. This structuring leads to the emergence of effective (phenomenological) laws at each of the higher levels, with apparent autonomy from the lower levels (Anderson 1972). It is this independence from the details of lower-level causation that allows phenomenological laws to be good effective theories of higher-level interactions (for they are levels of stable constitutive relationships); thus for example neurosurgeons do not have to understand particle physics or nuclear physics in order to ply their trade. Thus the context of the discussion is the modular hierarchical structures underlying complexity (Flood & Carson 1990; Simon 1992, chap. 7).

The key idea I will pursue is that as well as bottom-up causation, top-down causation takes place in these structures (Campbell 1974; Van Gulick 1995), due in particular to the crucial role of context in determining the outcomes of lower-level causation (Bishop & Atmanspacher 2006). I suggest there are at least five different types of top-down causation that can take place, depending on the context: namely, algorithmic top-down causation; top-down causation via nonadaptive information control; top-down causation via adaptive selection; top-down causation via adaptjive information control; and intelligent top-down causation. There could be others, but I claim that these can all be regarded as well established. In brief: *there are other forms of causation than those encompassed by physics and physical chemistry*. A full scientific view of the world must recognize this fact, or else it will ignore important aspects of causation in the real world, and so will give a causally incomplete view of things (Ellis 2005, 2006a, 2006b). This applies in particular in the human brain, and so is a key feature in the relation of the brain to the mind.

Table 4.1. The Hierarchy of Structure and Causation. This table gives a simplified representation of this hierarchy of levels of reality (as characterized by corresponding academic subjects) in living beings. Each lower level underlies what happens at each higher level, in terms of causation. For a more detailed description of this hierarchical structure, see http://www.mth.uct.ac.za/~ellis/cos0.html.

Level 8	Sociology/Economics/Politics
Level 7	Psychology
Level 6	Physiology
Level 5	Cell biology
Level 4	Biochemistry
Level 3	Chemistry
Level 2	Atomic physics
Level 1	Particle physics

2 Functional Context: Modular Hierarchical Structures

The context of the emergence of complexity is the hierarchical structure of matter and causal relations, characterized both by scale and by an appropriate classification and language of description for the entities that are recognized at each scale. It is a hierarchy of whole-part relations, which at the bottom levels can be seen as physical (one entity is physically a part of a larger one) but at the higher levels is a causal hierarchy (one entity provides the causal context for the other). For this hierarchical structure in the life sciences, illustrated in table 4.1, see Peacocke (1989), Campbell and Reece (2005). For the specific case of the brain, see Scott (1995).

Modularity

To enable true complexity to emerge, there will be numerous quasi-independent modules at each level of the hierarchy, interacting with each other in a network and enabling encapsulation, information hiding, abstraction, and inheritance (Booch 1994). This network structure is an irreducible higher-level characteristic. In addition to the properties of the units themselves, it is the set of relations between units, for example, large-scale topological relations as well as local causal motifs, that is crucial in building up complexity. These aspects cannot be reduced to lower-level variables.

Hierarchy

A hierarchical structure will be described by a corresponding hierarchy of variables appropriate to describing the different levels of the hierarchy. A *high-level variable* is a quantity that characterizes the state of the system in terms of a description using high-level concepts and language – it cannot be stated in terms of low-level variables. The higher levels of structure and causation cannot be reduced to lower-level terms, as the relevant concepts lie outside those that can be described in terms of lower-level concepts. As it is causal relations that count at the higher levels rather than physical nature or scale, high-level entities that occur in the life sciences hierarchy need not have a material nature (ethical values are an example, see below). They nevertheless have a clear place in the causal hierarchy, which can be thought of as bifurcating into a natural sciences and a life sciences branch with a single trunk (Murphy & Ellis 1995).

3 Bottom-Up Causation

Bottom-up causation is the ability of lower levels of reality to have a causal power over higher levels, in some cases uniquely determining what happens at the higher levels (fig. 4.1A). Examples are understanding of neuronal processes in terms of ion diffusion and the Hodgkin-Huxley equation at one level, and understanding it as a neural network built up from interacting neurons at a higher level.

The core of the strong reductionist view of science is that all can be explained by such bottom-up mechanisms based in the laws of physics, with no remainder.

4 Top-Down Causation

Top-down causation (Campbell 1974; Van Gulick 1995; Luisi 2002) is the ability of higher levels of reality to have a causal power over lower levels (fig. 4.1B). When dynamic effects take place, the outcome would be different if the higher-level context were different. Altering the high-level context alters lower-level actions; this is what identifies the effect as top-down causation. In such cases the high-level context variables are not describable in lower-level terms, and this is what identifies them as context variables.

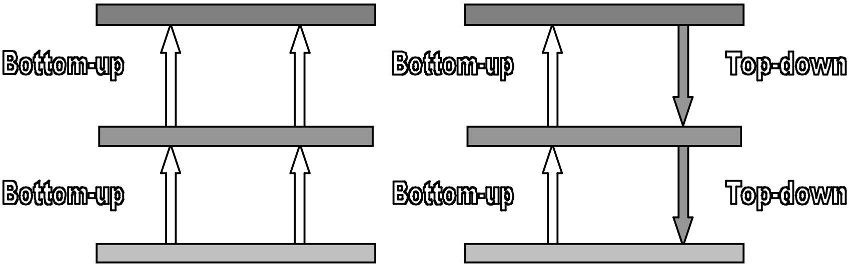

Fig. 4.1A. Bottom-up causation only **Fig. 4.1B.** Bottom-up and Top-down causation

How do you demonstrate top-down causation? You show that a change in high-level variables results in a demonstrable change in lower-level variables in a reliable way, after you have altered the high-level variable. It is the reliable nature of the change that characterizes it as causation and not just a random change; this is also what leads to predictability (the result is repeatable and thus testable). Thus you merely have to show that altering the high-level context alters the outcome in a way depending only on the top-level state, where the context variables are not describable in lower-level terms. Top-down causation as considered here means having causal power over lower levels, channeling causal effectiveness at those levels.

Top-down causation is ubiquitous in physics, chemistry, and biology, because the outcome of lower-level interactions is always determined by context. For example the wiring in a computer channels electrons from one specific component to another and thus enables logical computations to be performed. The kind of computation performed and resultant output, and hence the detailed switching of transistors at the micro level, depends both on the component connectivity and on the kind of program loaded into the computer (word processor, music, or graphics for example) – a high-level concept. These are constraints on the lower-level dynamics and so have causal power (Juarrero 1999).

Effective same-level action occurs when top-down causation combined with bottom-up causation leads to a resulting high-level outcome that depends only on the initial high-level state. In that case, the low-level dynamics commutes with coarse graining for all low-level states that correspond to each of the high-level states, and a coherent high-level dynamics emerges from the lower-level dynamics (Ellis 2006a) The resulting same-level action allows a phenomenological description of the higher-level action that is independent of the particular lower-level states that realize this action. This is the basis of the independence of higher-level descriptions from lower-level details and the reason that we can consider same-level causation at each level as ontologically real, expressed in terms of viable *effective theories* for the dynamics at that level (Hartmann 2001). When the lower-level dynamics does not commute with coarse graining in this way, no coherent higher-level dynamics emerges; for example, this occurs in chaotic dynamics.

While all top-down causation can be characterized as due to higher-level variables setting the context for lower-level action, there are clearly distinguishable ways this can happen. I suggest there exist at least *five different kinds of top-down causation*, themselves forming a hierarchy. They might not represent all the forms of top-down causation, but I believe these can all be regarded as well established. They can be acting simultaneously in the same physical system, different ones being effective at different scales. Note that we do not have to explain in detail how these various classes of causation work in order to determine both that they do indeed represent top-down causation, and that they are distinct from each other.

In the following I briefly discuss the nature of each of these five classes of top-down causation in turn, considering them in inverse order. For more details, see Ellis (2008).

4.1 Algorithmic Top-Down Causation

Algorithmic top-down causation occurs when high-level variables have causal power over lower-level dynamics through system structuring, so that *the outcome depends uniquely on the higher-level structural, boundary, and initial conditions*. The lower-level variables determine the outcome in an algorithmic way from the initial and boundary conditions as a consequence of the higher-level structural relations. The resulting high-level relations are then an inevitable consequence of the low-level interactions, given both the high-level context and the low-level dynamics (based in physics). It is the physical structuring and equations of state that determine the outcome resulting from particular boundary and initial conditions.

This is the kind of causation envisaged in the physicalist reductionist world-view, and occurs in all physical and natural systems as well as in biology. Algorithmic computational procedures in a digital computer that proceed on the basis of initial data only is an example; the algorithms (stored in high-level computer programs) determine the machine code that then determines the (low-level) switching of transistors (Tanenbaum 1990). This represents top-down causation, as different stored programs will employ different algorithms and so result in different transistor switching. Such machine-like processes controlled by algorithms may entail purpose (a desired outcome) without embodying goals (no feedback control system can be implemented in the algorithm, because in the stated context it cannot utilize updated information). Then they will run, tending to produce the desired outcome, but will be vulnerable to disturbances that they are unable to respond to. Thus in real world situations they will be unreliable. Examples are a stock control system that is not related to checks of the actual physical stock from time to time, and an aircraft autopilot that is not fed updated information on position and winds.

The way neuronal networks with given structure and weights process information in the brain, for example in the visual system, depends in an algorithmic way on the inputs to the system, and provides an important example (LeDoux 2002; Koch 2004). The output to the cortex indeed depends on these inputs, but is uniquely determined by them. It is this feature that underlies the reliability of sensory systems in animals, and that led Francis Crick to his famous aphorisim: "You, your joys and your sorrows, your memories and your ambitions, your sense of personal identity and free will, are in fact no more than the behavior of a vast assembly of nerve cells and their associated molecules" (Crick 1995).

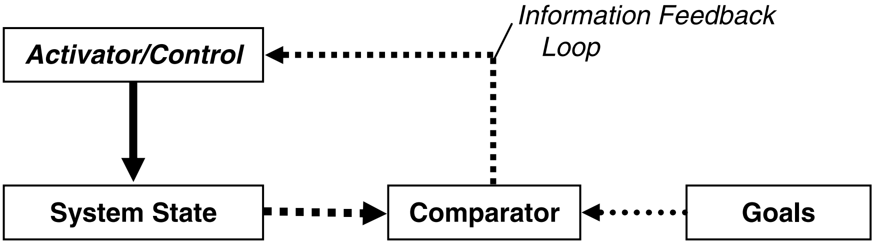

Fig. 4.2. The basic feedback control process. The goals tend to lead to a specific final state via a specific mode of physical action. The initial state of the system is then irrelevant to its final outcome, provided the system parameters are not exceeded.

4.2 Nonadaptive Information Control

In nonadaptive information control, higher-level entities influence lower-level entities so as to attain specific fixed goals through the existence of feedback control loops, whereby information on the difference between the system's actual state and desired state is used to lessen this discrepancy (S. Beer 1966; Simon 1992, chap. 1). A commonplace example is control of the temperature of water in a hot water cylinder by a thermostat controlling the water heater. Unlike the previous case, the outcome is not determined by the boundary or initial conditions; rather it is determined by the goals. The goal is attained through feedback control that functions by comparison of the current system state and the goal by a controller; information on the difference is fed back to the activator (see fig. 4.2). Thus feedback control systems depend essentially on information flows, plus an evaluation of that information relative to the chosen goals. The goals are embodied in the system structure and so do not change with time; there may however be some associated form of information storage and retrieval, and perhaps even implicit or explicit information processing. The control circuits are higher-level entities, as they are based in higher-level concepts (the high-level system state and the goal). The goals are intrinsic higher-level properties of the system considered, and they determine the outcome.

This is top-down causation because the goals are only expressible in higher-level terms, and are implemented by higher-level networks; these cannot be reduced to lower-level entities, precisely because it is the relations between the parts that make the network into a feedback control system. Taking the system apart destroys those relations. This is a core feature of physiology, and in particular brain function. For example, sodium and potassium levels in neurons are controlled by voltage-gated ion channels deployed in feedback control loops that return the membrane potential of an axon to its resting state, the resting potential. A *paramecium* has multiple systems for returning its membrane potential to the resting state (Greenspan 2007).

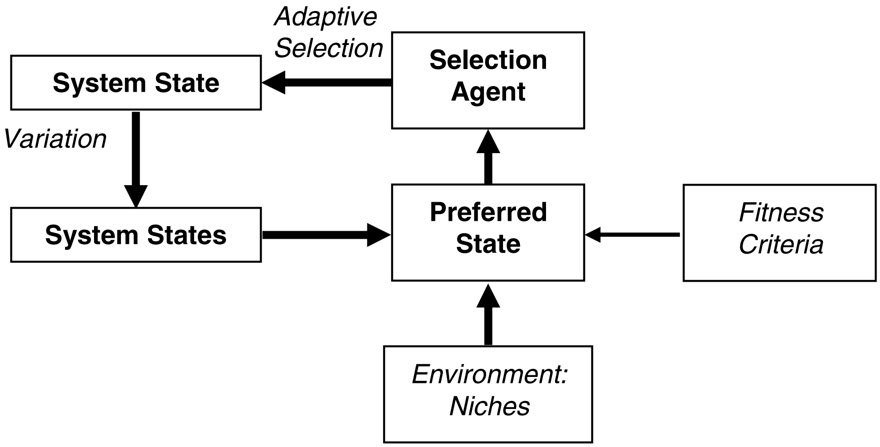

Fig. 4.3. Adaptive selection. The meta-goals embodied in the value system do not lead to a specific final state: rather they lead to any one of a class of states that tends to promote the meta-goals. Thus the final state is not uniquely determined by the meta-goals; random variation influences the outcome by leading to a suite of states from which an adaptive selection is made in the context of both the fitness criteria and the environment.

4.3 Adaptive Selection

Adaptive processes (Holland 1992) take place when many entities interact, for example, the cells in a body or the individuals in a population, and *variation takes place in the properties of these entities, followed by selection of preferred entities that are better suited to their environment or context* (fig. 4.3). Higher-level environments provide niches that are either favorable or unfavorable to particular kinds of lower-level entities; those variations that are better suited to the niche are preserved and the others decay away. Criteria of suitability in terms of fitting the niche can be thought of as fitness criteria guiding adaptive selection. On this basis a *selection agent* or *selector* (the active element of the system) accepts one of the states and rejects the rest; this selected state is then the current system state that forms the starting basis for the next round of selection, ultimately leading to the emergence and nature of biological form.

This can be thought of as a generalized feedback loop with a *meta-purpose* provided by fitness criteria classifying what kinds of outcomes are desirable and which not (these are higher-level purposes that are not directly attained as are the goals in a feedback control system, but still effectively guide what happens by selecting preferable outcomes). In some cases the fitness criteria may be implicit rather than explicit, being built in to the way the selection agent functions rather than being a separate function. Thus this is top-down causation from the context to the system. An equivalence class of lower-level variables will be favored by a particular niche structure in association with a specific fitness criteria. Unlike feedback control, this process does not attain preselected internal goals by a specific set of mechanisms or systems; rather it creates systems that favor the meta-goals embodied in the fitness criteria. This is an adaptive process rather than a control process. It is the way new information is generated that was not present before (Roederer 2005), and it enables emergence of complexity without dynamical attractors or specific goals guiding the process, but with an increase of complexity and embodied information, for the process searches the possible solution space in a way that is not preordained and adapts to the context. The outcome is usually not predictable either from the initial conditions or the meta-goals, because of the random element involved, although both clearly influence the outcome. This underlies all life, including cells and plants and animals.

For example, the training of artificial neural nets to perform a specific task (say, letter recognition) determines the interaction weights in the network (Bishop 1999). The niche is a particular set of letters to be recognized. The fitness criterion is correct pattern recognition, and the adaptive process is the training of the neural network. This is a form of top-down causation from the pattern to be recognized (a high-level concept, as it is defined in terms of the relation between the elements) to the low-level property of network weights. Different sets of weights can perform the same function, so the acceptable set of weights is an equivalence class. Decision-making is a property of the network rather than of any single cell (Greenspan 2007). The new weights are chosen in such a way as to probably provide better performance, so this is an example of predictive adaptation. Genetic algorithms (implemented on digital computers) are specifically designed to solve problems in an adaptive way (Mitchell 1998). A “fitness function” is defined over the genetic representation and measures the quality of the represented solution, thus providing the needed fitness criteria in this case.

In the brain, adaptive selection of the responses and selectivity of sensory neurons takes place in a dynamic manner so as to match changes in input stimuli (Gutnisky & Dragoi 2008). Neuromodulation allows patterns of neural activity to adapt to new conditions (Greenspan 2007). A key link of macro to micro conditions in the brain is a form of adaptive selection that has been called “neural Darwinism” by Gerald Edelman (1989), which refines neuronal connections on the basis of higher-level fitness criteria provided by a “value system” that guides brain plasticity in response to environmental interactions and that is made effective by neurotransmitters diffused to the cortex from the limbic system. This system-evaluating salience is nothing other than the various hard-wired primary emotions identified by Panksepp (Ellis & Toronchuk 2005). A particular case is *habituation*, that is, learning to ignore a stimulus that lacks meaning (M.E. Beer et al. 2007, pp. 763–67). In addition, the process of perception is a predictive adaptive process using Bayesian statistics to update the current perception on the basis of prediction errors (Frith 2007, pp. 125–27). This includes prediction of the intention of others, which is the basis of theories of other minds (Frith 2007, pp. 163–83).

To show that adaptive selection is operating, one has to show that variation occurs in some population of entities, followed by selection according to some identifiable fitness criterion, with this cycle occurring on a continuing basis. It is central to biological functioning as well as to Darwinian evolution; it is the prime way adapted complex structures can be built up in biology (Campbell & Reece 2005). One should note that it occurs on *all three biological timescales:* evolutionary, developmental, and functional. Thus it occurs in both phylogeny and ontogeny.

4.4 Adaptive Information Control

Adaptive information control takes place when there is *adaptive selection of goals in a feedback control system*, thus combining both feedback control and adaptive selection. The goals of the feedback control system are irreducible higher-level variables determining the outcome, but are not fixed, as in the case of nonadaptive feedback control; they can be adaptively changed in response to experience and information received. The overall process is guided by fitness criteria for selection of goals, and is a form of adaptive selection in that goal selection relates to future rather then present use of the feedback system. This allows great flexibility of response to different environments, indeed in conjunction with memory it enables learning and anticipation (Simon 1992, chap. 4) and underlies effective purposeful action as it enables the organism to adapt its behavior in response to the environment in the light of past experience, and hence to build up complex levels of behavior. It enables goals to be specific to individuals and to vary with time and experience.

To show that adaptive information control is operating, one has to show that information control is taking place, but now with the goals continually adapted to context and hence varying across individuals and with time. It underlies animal intelligence and action; indeed it separates the animal kingdom from plants. The classic example is associative learning in animals, such as Pavlovian conditioning: animal response to a stimulus such as a sound, which is taken as a sign of something else and causes physical reactions implemented by motor neurons. The training is causally effective by top-down action from the brain to cells in muscles. The fitness criterion is avoidance of negative stimuli. You demonstrate this top-down causation by changing the conditioning and finding that the response is different. This occurs in higher animals (dogs, for example) and in the snail *Aplysia* (M.E. Beer et al. 2007, pp. 763–71). In higher animals it is a form of predictive control, using a temporal difference algorithm to discover the best sequence of actions to perform in order to attain the goal (Frith 2007, pp. 95–97),

4.5 Intelligent Top-Down Causation

Intelligent top-down causation is the special case of feedback control with adaptive choice of goals, where *the selection of goals involves the use of symbolic representation to investigate the outcome of goal choices*. Here a symbolic system is a set of structured patterns, realized in time or space, that is arbitrarily chosen by an individual or group to represent objects, states, and relationships. It will generally involve hierarchical structuring and recursion, will use grammar and syntax as a vehicle for conveying semantic meaning, and has the potential to enable quantitative as well as qualitative investigation of outcomes.

This symbolic representation and choice of goals entails the causal efficacy of abstract entities such as action plans, the theory of the laser, and the value of money, represented symbolically. Thus the key feature of this higher-level of causation, distinguishing it from the general case of adaptive control systems, is its use of language (spoken or written) and abstract symbolism (Deacon 1997), extending to the quantitative and geometrical representations of mathematical models (Devlin 1996). These are all irreducible higher-level variables of an abstract nature: they form equivalence classes of representations, *inter alia* because they can be represented in different languages, and in spoken or written form. They enable information to be stored and retrieved, classified and selected as relevant or discarded, processed in the light of other information, and used to make qualitative and quantitative projections of outcomes and plan future actions in a rational way (Simon 1992, chap. 4), altering goals according to an intelligent understanding of past experiences and future expectations. Intentional action then enables one to implement the resulting plans, and so change the physical world. The outcome is thus the result of human agency (Frith 2007, p. 152). In doing so, one should recognize the causal power of images (Boulding 1961) and formal and informal causal models of the natural and social worlds (Frith 2007, p. 126) subject to predictive correction (Frith 2007, pp. 134–38), ranging from mental images of what might happen to elaborate quantitative models of physical entities and societies. These abstract entities (which are shared among many minds) play a large part in formulating our understandings and consequent actions, and hence are causally effective in the real world as they help us attain our goals.

An example is aircraft design: Plans for a jumbo jet aircraft result in billions of atoms being deployed to create the aircraft in accordance with those plans. This is a nontrivial example: It costs a great deal of money to employ experts in aerodynamics, structures, materials, fuels, lubrication, controls, and so forth to design and then to manufacture the aircraft in accordance with those plans. The plan itself is not equivalent to any single person's brain state: it is an abstract, hierarchically structured equivalence class of representations (spoken, drawn, in computers, in brains, etc.) that together comprise the design. It is clearly causally effective (the aircraft would not exist without it). Thus abstract plans and entities, such as social agreements, are causally effective. A second example is the value of money. Physically, money is just coins or pieces of paper with patterned marks on them. This does not explain its causal significance. The effectiveness of money, which can cause physical change in the world such as the construction of buildings, roads, bridges, and so on by top-down action of the mind to material objects, is based in social agreements that lead to the value of money (pricing systems) and exchange rates. These are abstract entities arising from social interaction over an extended period of time, and they are neither the same as individual brain states nor equivalent to an aggregate of current values of any lower-level variables (although they may be represented by and are causally effective through such states and variables).

Of course we do not fully understand how the mind is able to plan and make choices resulting in top-down action as discussed here. The fact that we do not know how it works does not affect the fact that we are certain it can happen and does happen (Simon 1992, chaps. 5 and 6). Indeed you could not be reading this chapter if it were not true: the marks on the paper that constitute the letters you are reading have the form they do because of top-down action from my mind to my hand. This chapter would not exist were this not possible. Thus this form of top-down causation has been demonstrated many millions of times.

Social roles are socially determined abstract entities that are causally effective in structuring society. They are a key aspect of the way individual behavior links with the social environment. Roles are developed by an adaptive process which is a combination of bottom-up and top-down interaction between society and the individuals who make up the society (Berger 1963). They are then inculcated into the individual by top-down social processes (Berger & Luckmann 1967; Cacioppo et al. 2002). Thereafter they become a core feature of individual psychology in relation to society(Longres 1990). Together with expectations guiding the choice of goals and actions, they are causally effective in a top-down way from the mind to the body. Roles embody social values, which, together with individual values relating to life purpose, guide the individual and communal choice of goals and the methods used to attain these goals. Thus the highest level adaptive goals are *values,* related to ethics, aesthetics, and meaning, which are all causally effective in a top-down way by determining the set of desirable lower-level goals (Murphy & Ellis 1995). The imperative to search for meaning is a key aspect of human nature (Frankl 1984), without which, for example, the entire edifice of science would not exist.

Thus *our understandings of meaning and purpose* are abstract entities that form the highest level in the hierarchy of causation in the mind and in organizations. The related ethical values are nonreducible higher-level variables; by determining the nature of acceptable lower-level goals, they are a set of abstract principles that are causally effective in the real physical world; indeed they crucially determine what happens. For example, wars will or will not be waged depending on ethical stances; large-scale physical devastation of the earth will result if thermonuclear war takes place. So the nature of ethical stances has crucial effects in the way human activity impacts on society and the world.

5 Freedom at the Bottom?

I have claimed here that top-down causation is causally effective, which means that even in principle, micro-level laws fail to fully determine outcomes of complex systems: causal closure is achieved only by appealing to downward causation. But this claim is clearly in trouble if the system is already causally closed at the micro level, as is supposed by most physicists. For higher levels to be causally efficacious over lower levels, there has to be some causal slack at the lower levels, otherwise the lower levels would be causally overdetermined. Where does the causal slack lie? Three key features are relevant.

First, in considering specific physical and biological systems, the slack lies partly in the *structuring of the system* so as to attain higher-level functions; for example, the specific connections in a computer (which could have been different) act as constraints on lower-level dynamics (Juarrero 1999), thus channeling how they function. This causal slack also lies partly in the *openness of the system*: new information can enter across the boundary and affect local outcomes. For example, cosmic rays may enter the solar system and alter the genetic heritage of individual humans; alteration in solar radiation can cause climate change on earth; telephone calls from afar convey vital information that changes how we act. This is top-down causation from the overall context to the system. Local systems are not isolated in either space or time, and their future evolution cannot be predicted from their internal properties alone. Thus, for example, weights of network connections shape the outcome of neural net and brain functioning, and these weights are developed through network training using external information. The "neurons" do not function independent of context.

The second key feature (Luisi 2002) is that *top-down causation changes the nature of the lower elements*. The situation does not merely consist of invariant lower-level elements obeying physical laws; rather, the nature of lower-level elements is being changed by context. Often the way this occurs ensures that the lower-level elements obey physical laws in a way that fulfils higher-level purposes. This is then an aspect of adaptive selection. For example, through the processes of developmental biology, cells differentiate to perform specific functions; this changes their nature relative to other cells in an adaptive way. Cells differentiate into neurons adapted to their location in the brain, into muscle cells adapted to their role in the heart, and so on. They each develop so as to fit into their allotted role in the body, creating the body and its biological form as they do so, and are then fine-tuned for their function. A particular case is the adaptive coding of sensory neurons (Gutnisky & Dragoi 2008). Another example is humans in society. Individual minds develop in the context of their interactions with other minds, and brain development cannot be understood outside this context (Donald 2001; Frith 2007, p. 187).

Individuals are shaped by society so that they fit into that society, for example, learning a specific language and a variety of societal roles and expectations (Berger 1963). This is top-down causation from the society to the individual, and indeed to the person's synaptic connections; one can say that the individual's brain is adapted to fit into the society in which he or she lives (Berger & Luckmann 1967; Cacioppo et al. 2002). Thus *the nature of micro-causation is changed by these top-down processes*, profoundly altering the mechanistic view of how things work.

Third, the required freedom lies in *micro-indeterminism* (random outcomes of microphysical effects), *combined with adaptive selection*: random outcomes at the micro level allow variation at the macro level, which then leads to selection at the micro level but based in macro-level properties and meaning. Statistical variation and/or quantum indeterminacy provides a repertoire of variant systems that are then subject to processes of Darwinian section, based on higher-level qualities of the overall system. For this to work, one needs amplifying mechanisms in order to attain macroscopic variation from quantum fluctuations. Some physical systems (such as photomultipliers and the human eye) amplify quantum effects to a macroscopic scale; some classically chaotic systems can amplify fluctuations in initial data that are of quantum origin; and some molecular biology processes (for example involving replication of mutated molecules) act as such amplifiers (Percival 1991). There is considerable evidence that these kinds of effects lead to indeterminacy in brain and behavior (Glimcher 2005). At a profound level the universe is indeterministic (Feynman 1992; Polkinghorne 2002), allowing the needed causal slack. By itself that does not lead to emergence of higher-level order; but it does allow this through the process of adaptive selection (Roederer 2005).

Whether these are sufficient to account for free will is not clear. But in any case the evidence for effective higher-level autonomy is very strong, as I discuss next, so there must be some way it is possible, even if we do not yet know what that way is.

My claim is that *there has to be adequate causal slack because it is needed in order to explain the detailed complexity that exists in the universe.* The claim made by bottom-up determinism is physical causal completeness: for any specific physical system, including human minds, physical laws alone give a unique outcome for each set of initial data. To see the improbability of this claim, one can contemplate what is required from this viewpoint when placed in its proper cosmic context. The implication is that the particles that existed at the time of decoupling of the Cosmic Background Radiation in the early universe (Silk 2001; Dodelson 2003) just happened to be placed so precisely as to make it inevitable that fourteen billion years later, human beings would exist and Crick and Watson would discover DNA, Townes would conceive of the laser, Witten would develop M-theory.

In my view, this is absurd. It is inconceivable that truly random quantum fluctuations in the inflationary era – the supposed source of later emergent structure (Dodelson 2003) – can have had implicitly coded in them the future inevitability of the Mona Lisa, Nelson's victory at Trafalgar, Einstein's 1905 theory of relativity. Such later creations of the mind are clearly not random; on the contrary, they exhibit high levels of order embodying sophisticated understandings of painting, military tactics, and physics, respectively, which cannot possibly have directly arisen from random initial data. This proposal simply does not account for the origin of such higher-level order. In any case it is not possible because of quantum uncertainty.

Quantum fluctuations can change the genetic inheritance of animals (Percival 1991) and so influence the course of evolutionary history on Earth. Indeed that is in effect what occurred when cosmic rays – whose emission processes are subject to quantum uncertainty – caused genetic damage in the distant past: "The near universality of specialized mechanisms for DNA repair, including repair of specifically radiation-induced damage, from prokaryotes to humans, … suggests that the earth has always been subject to damage/repair events above the rate of intrinsic replication errors," and that "radiation may have been the dominant generator of genetic diversity in the terrestrial past" (Scalo et al. 2003). Consequently the specific evolutionary outcomes of life on Earth (the existence of dinosaurs, giraffes, humans) cannot even in principle be uniquely determined by causal evolution from conditions in the early universe, or from detailed data at the start of life on Earth. Quantum uncertainty prevents this because it significantly affected the occurrence of radiation-induced mutations in this evolutionary history. The specific outcome that actually occurred was determined as it happened, when quantum emission of the relevant photons took place: the prior uncertainty in their trajectories was resolved by the historical occurrence of the emission event, resulting in a specific photon emission time and trajectory that was not determined beforehand, with consequent damage to a specific gene in a particular cell at a particular time and place that cannot be predicted even in principle.

Finally, we should recognize that the enterprise of science itself does not make sense if our minds cannot rationally choose between alternative theories on the basis of the available data; this would indeed be the situation if one takes seriously the bottom-up mechanistic view that the mind simply dances to the commands of its constituent electrons and protons, algorithmically following the imperatives of Maxwell's equations and quantum physics. A reasoning mind able to make rational choices is a prerequisite for the academic subject of physics to exist. The proposal that apparent rationality is illusory, being just the inevitable outcomes of microphysics, cannot account for the existence of physics as a rational enterprise. But this enterprise does indeed make sense; thus one can provisionally recognize the possibility that free will too is an active causal factor, not directly determined by the underlying physics. Indeed, I suggest a stronger statement: if your theory does not allow the existence of free will in a serious sense, then it is not a good enough theory – for you cannot engage in scientific activity without it!

6 Multiple Categories of Causation

Reductionist analysis "explains" the properties of the machine by analyzing its behavior in terms of the functioning of its component parts (the lower levels of structure). Systems thinking tries to understand the properties of the interconnected complex whole (Churchman 1968; Flood & Carson 1990) and "explains" the behavior or properties of an entity by determining its role or function within the higher levels of structure. For example, the question *Why is an aircraft flying?* can be answered in various ways:

- In *bottom-up terms*: It flies because air molecules impinge against the wing with slower moving molecules below creating a higher pressure as against that due to faster moving molecules above, leading to a pressure difference described by Bernoulli's law, this counteracts gravity, and so forth.
- In terms of *same-level explanation*: It flies because the pilot is flying it, after a major process of training and testing that developed the necessary skills, and she is doing so because the airline's timetable dictates that there will be a flight today at 16h35 from London to Berlin, as worked out by the airline executives on the basis of need and carrying capacity at this time of year.
- In terms of *top-down explanation*: It flies because it is designed to fly! This was done by a team of engineers working in a historical context of the development of metallurgy, combustion, lubrication, aeronautics, machine tools, computer aided design, and so on, all needed to make this possible, and in an economic context of a society with a transportation need and complex industrial organizations able to mobilize all the necessary resources for design and manufacture. A brick does not fly because it was not designed to fly.
- In terms of *ultimate explanation:* And why was it designed to fly? Because it will make a profit for the manufacturers and the airline company! Without the prospect of that profit, it would not exist.

These are all simultaneously true, nontrivial explanations; *the plane would not be flying if they were not all true at the same time.* The higher-level explanations involving goal choices rely on the existence of the lower-level explanations involving physical mechanisms in order that they can succeed, but are clearly of a quite different nature than the lower-level ones, and are certainly not reducible to them nor dependent on their specific nature. The bottom-up kind of explanation would not apply to a specific context if the higher-level explanations, the result of human intentions, had not created a situation that made it relevant.

This situation was captured by Aristotle through his proposal of four different kinds of causation. According to Falcon (2006), they are

- *the material cause*: "that out of which," for example, the bronze of a statue;
- *the formal cause*: "the form," "the account of what-it-is-to-be," for example, the shape of a statue;
- *the efficient cause*: "the primary source of the change or rest," for example, the artisan, the art of bronze-casting the statue, the man who gives advice;

- *the final cause*: "the end, that for the sake of which a thing is done," for example, health is the end of walking, losing weight, purging, drugs, and surgical tools.

The last is a *teleological explanation* – an explanation that makes a reference to *telos* or purpose. Additionally, circular causation is possible: things can be causes of one another – a relation of reciprocal influence.

These four kinds of causes correspond broadly to those identified above in the case of the flying aircraft. Indeed we can adapt Aristotle's categorization to the hierarchical context considered here by seeing the material cause as the lower-level (physical) cause, the efficient cause as the same-level (immediate) cause, the formal cause as the immediate higher (contextual) cause, and the final cause as the ultimate higher-level cause. The key point about causality in real-world contexts, then, is that simultaneous multiple causality (inter-level, as well as within each level) is always in operation in complex systems. For example, successful completion of a physics experiment such as observing particle production in a particle collider involves all the reinterpreted Aristotelian forms of causation. The material (physical) cause is the particle interactions that lead to the production of new particles. The efficient (immediate) cause is that the experimenter turns the accelerator and measuring equipment on at a particular time. The formal (contextual) cause is that the collider was designed and manufactured so that the collisions would take place and outcomes could be observed. The final cause might simply be that the experimenter wanted to understand the collision in the context of a theory of AdS/CFT duality, or it might be because she aspired to attaining a Nobel prize.

7 Conclusion

The fact that physics is not the only form of causation in the real world has been demonstrated above by numerous examples. Physics provides necessary conditions (but not the sufficient conditions) for what happens; it provides the possibility space for what happens, but does not determine the outcome. Top-down causation allows higher-level causes to be what they appear to be: real effective causes. Context is the key to physical outcomes: multiple causation is always at work. Random fluctuations along with quantum uncertainty provide the freedom at the bottom needed to allow this to happen. It enables the causal power of abstract entities – mathematics, theories, ethics, social constructs – and underlies the paradox of the experimenter in physical science: all scientific experiments are based on purposeful activity and free will, enabling decisions based in abstract analysis that lies beyond the explanatory scope of physical science.

This paper has focused on top-down causation because it is the mode of causation that is least considered at present. However it should be emphasized *that bottom-up, same-level, and top-down causation all occur at the same time, in concert, enabling the emergence of genuine complexity based in modular hierarchical systems.* The complex whole of physical organisms situated in the geographical and historical context of their environment and their evolutionary history arises from the interaction of these different modes of causation. Broadly speaking, same-level causation is where the action is; bottom-up causation enables it to happen; and top-down causation decides what happens. Furthermore, this is true for every level of the hierarchy (except the very top-most and very bottom-most; but we do not know what those levels are). A perceived reality of same-level causation at the cellular and molecular levels underlies Crick's dictum already quoted above (section 4.1). But nerve cells and molecules are made of electrons plus protons and neutrons, which are themselves made of quarks – so why not: "You, your joys and your sorrows, your memories and your ambitions, your sense of personal identity and free will, are in fact no more than the behavior of a vast assembly of quarks and electrons"? And these themselves are possibly vibrations of superstrings.

So why does Crick stop at the level he chooses? Undoubtedly because that is the level he best understands and is familiar with! Indeed scientists will perceive as fundamental the level they happen to work on and understand deeply in causal terms, so they usually assume that causality at that level is real. And that is a reasonable perception, if they are all real, as I take to be the case (a table is still a table even though it is made of atoms, for example; and the atoms are also real, as are the neutrons and protons). Crick's dictum either applies to all levels except the (unknown) bottommost one, or to none. If it applies to all levels, Crick's molecules are no more real than memories and ambitions; but he assumes the molecules are real, so his position is inconsistent. There is no reason to privilege molecules or cells in the hierarchy of structure. If we accept molecular reality, as I do, then we should also acknowledge the memories and ambitions as real too, for that is then the only consistent position.

References

Anderson, P.W.: More is different. Science 177, 377 (1972); Reprinted in Anderson, P.W.: A Career in Theoretical Physics. World Scientific, Singapore (1994)

Beer, M.E., Connors, B.W., Paradiso, M.A.: Neuroscience: Exploring the Brain. Lippincot, Williams and Wilkins, Philadelphia (2007)

Beer, S.: Decision and Control. Wiley, New York (1966)

Berger, P.: Invitation to sociology: A humanistic perspective. Doubleday, New York (1963)

Berger, P., Luckmann, T.: The social construction of reality: A treatise in the sociology of knowledge. Anchor, New York (1967)

Bishop, C.M.: Neural networks for pattern recognition. Oxford University Press, Oxford (1999)

Bishop, R., Atmanspacher, H.: Contextual emergence in the description of properties. Foundations of Physics 36, 1753–1777 (2006)

Booch, G.: Object oriented analysis and design with applications. Addison Wesley, New York (1994)

Boulding, K.E.: The image: Knowledge in life and society. University of Michigan Press, Ann Arbor (1961)

Cacioppo, J.T., Berntson, G.G., Adophs, R., Carter, C.S., Davidson, R.J., McClintock, M.K., Mcewan, B.S., Meaney, M.J., Schacter, D.L., Sternberg, E.M., Suomi, S.S., Taylor, S.E. (eds.): Foundations in social neuroscience. MIT Press, Cambridge (2002)

Campbell, D.T.: Downward causation. In: Ayala, F.J., Dobhzansky, T. (eds.) Studies in the philosophy of biology: Reduction and related problems, pp. 179–186. University of California Press, Berkeley (1974)

Campbell, N.A., Reece, J.B.: Biology, 7th edn. Pearson, Benjamin Cummings, San Francisco (2005)

Churchman, C.W.: The systems approach. Delacorte Press, New York (1968)

Crick, F.: The astonishing hypothesis The scientific search for the soul. Scribner, New York (1995)

Deacon, T.: The symbolic species: The co-evolution of language and the human brain. Penguin, London (1997)

Devlin, K.: Mathematics: The science of patterns. Henry Holt & Company, New York (1996)

Dodelson, S.: Modern cosmology. Academic Press, San Diego (2003)

Donald, M.: A mind so rare: The evolution of human consciousness. W.W. Norton, New York (2001)

Edelman, G.M.: Neural Darwinism: The theory of group neuronal selection. Oxford University Press, Oxford (1989)

Ellis, G.F.R.: Physics, complexity, and causality. Nature 435, 743 (2005), http://wwwnature.com/nature/journal/v435/n7043/edsumm/e050609-03.html

Ellis, G.F.R.: On the nature of emergent reality. In: Clayton, P., Davies, P.C.W. (eds.) The Re-emergence of emergence, pp. 79–110. Oxford University Press, Oxford (2006a)

Ellis, G.F.R.: Physics and the real world. Foundations of Physics 26(2), 227–236 (2006b)

Ellis, G.F.R.: On the nature of causation in complex systems. Transactions of the Royal Society of South Africa [Centenary Issue] 63, 69–84 (2008)

Ellis, G.F.R., Toronchuk, J.A.: Neural development: Affective and immune system influences. In: Ellis, R.D., Newton, N. (eds.) Consciousness and emotion, pp. 81–119. John Benjamins, Philadelphia (2005), http://www.mth.uct.ac.za/~ellis/AND.doc

Falcon, A. (2006), Aristotle on Causality. Stanford Encyclopedia of Philosophy, http://plato.stanford.edu/entries/aristotle-causality/

Feynman, R.: The character of physical law. Penguin, London (1992)

Flood, R.L., Carson, E.R.: Dealing with complexity: An introduction to the theory and application of systems science. Plenum, London (1990)

Frankl, V.: Man's Search for Meaning. Washington Square Press, New York (1984)

Frith, C.: Making up the mind: How the brain creates our mental world. Blackwell, Malden (2007)

Glimcher, P.W.: Indeterminacy in brain and behavior. Annual Review of Psychology 56, 25–56 (2005)

Greenspan, R.J.: An introduction to nervous systems. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (2007)

Gutnisky, D.A., Dragoi, V.: Adaptive coding of visual information in neural populations. Nature 452, 220–224 (2008)

Hartmann, S.: Effective Field Theories, Reductionism, and Scientific Explanation. Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 32, 267–304 (2001)

Holland, J.H.: Adaptation in natural and artificial systems. MIT Press, Cambridge (1992)

Juarrero, A.: Dynamics in action: Intentional behavior as a complex system. MIT Press, Cambridge (1999)

Koch, C.: The quest for consciousness: A neurobiological approach. Roberts and Company, Englewood (2004)

LeDoux, J.: Synaptic self. Viking, New York (2002)

Longres, J.F.: Human behavior in the social environment. F.E. Peacock, Itasca, IL (1990)

Luisi, P.L.: Emergence in chemistry: Chemistry as the embodiment of emergence. Foundations of Chemistry 4, 183–200 (2002)

Mitchell, M.: An introduction to genetic algorithms. Complex Adaptive Systems. MIT Press, Cambridge (1998)

Murphy, N., Ellis, G.F.R.: On the moral nature of the universe. Fortress, Minneapolis (1995)

Peacocke, A.R.: An introduction to the physical chemistry of biological organization. Oxford University Press, Oxford (1989)

Percival, I.: Schrödinger's quantum cat. Nature 351, 357 (1991)

Polkinghorne, J.: Quantum theory: A very short introduction. Oxford University Press, Oxford (2002)

Roederer, J.G.: Information and its role in nature. Springer, Berlin (2005)

Scalo, J.M., Wheeler, J.C., Williams, P.: Intermittent jolts of galactic UV radiation: Mutagenetic effects. astro-ph/0104209. In: Celnikier, L.M., Van Thanh, J.T. (eds.) Frontiers of life – XIIth Rencontres de Blois. Gioi Publishers, Hanoi (2003)

Scott, A.: Stairway to the mind. Springer, New York (1995)

Silk, J.: The big bang. Freeman, New York (2001)

Simon, H.A.: The Sciences of the artificial. MIT Press, Cambridge (1992)

Tanenbaum, A.S.: Structured computer organization. Prentice Hall, Englewood Cliffs (1990)

Van Gulick, R.: Who's in charge here? And who's doing all the work? In: Heil, J., Mele, A. (eds.) Mental Causation, pp. 233–256. Clarendon, Oxford (1995)

5

Top-Down Causation and Autonomy in Complex Systems

Alicia Juarrero

Emeritus, Prince George's Community College
Largo, MD 20774-2199
ajuarrero@pgcc.edu

Summary. Evolutionary evidence shows that complex dynamical systems become increasingly self-directed and decoupled from merely energetic forces over time. In this paper I analyze these transformations, concentrating on changes in the type of top-down causation that characterizes such self-organized and autopoietic processes. Specifically, I show that the top-down selection criteria of these systems makes some of them autonomous, and that because once evolution reaches humans the criteria according to which voluntary actions are selected are semantic and symbolic – and can be self-consciously chosen – human self-direction constitutes a form of strong autonomy that can arguably be considered "free will."

Keywords: autonomy, complexity, constraints, context-dependence, emergence, evolution, free will, multiple realizability, selection, self-determination, self-organization, top-down causation.

1 Introduction

Complex dynamical systems are neither completely rigid nor fully random; instead, they display a unique balance of integration, cohesion and robustness at the global level and, at the same time, differentiation and multiple realizability at the component level. In this chapter I first defend strong top-down causation in such systems. I base my defense on the cohesive properties supplied by the context-sensitive constraints that first create such complex systems and then hold them together. These constraints integrate previously independent parts into a unified whole that (a) incorporates the record of its history, (b) is embedded in its environment, and (c) possesses emergent properties. Such systems, which can remain the same *type* of phenomenon despite being composed of different *token* arrays, can change the very components that make them up and even alter their environment – and they do so on the basis of meaningful criteria defined at the global level. In the second half of the chapter I show how the autonomy and self-determination embodied in strong top-down causation is increasingly strengthened over the course of evolution. Iteration of the dynamics that constitute them progressively decouples complex systems from fundamental energetic forces by bringing those mechanisms responsible for top-down causation further and further inside the system. At each step of the iteration a corresponding change in the selective criteria on the basis of which that top-down causation is exercised also occurs. Such strong top-down causation and autonomy, I conclude, ultimately make room for a variety of free will worth wanting, without having to appeal to quantum indeterminacy or the Wheeler-Lloyd limit.

2 Dynamical Systems Theory

2.1 No Constraints

Closed, isolated systems that do not exchange matter and energy with their environment cannot decrease their entropy or become more ordered or complex. An agglomeration or conglomerate composed of particles independent of each other at equilibrium cannot differentiate into a complex organization with emergent properties, and particles related to each other only in terms of relative position at best produce agglomeration or conglomerates. Since the properties of particles do not change when they are merely elements of a conglomerate, any novel characteristics of aggregates near equilibrium – such as temperature and pressure – are merely nominally emergent features of the statistical average of the large number of particles (Bedau 2002; Fromm 2005b). No top-down causation is in evidence in this type of phenomenon and no significant emergence or decoupling from bottom-up energetic forces is therefore available to isolated independent particles at equilibrium.

2.2 Bottom-Up Integration

Now consider instead the convection flow of hexagonal Bénard cells, those physical *dissipative structures* that emerge when open systems are driven far from equilibrium as a result of exchanges of matter and energy with their environment. When a pan of water is heated uniformly from below, the combined constraints established by the continual influx of energy and the container walls take the system farther and farther from equilibrium until a phase change is precipitated. Suddenly, instead of dampening any fluctuations or perturbations, one of these becomes amplified and the system bifurcates into a new mode of organization. The phase change streamlines the system's organization and decreases its internal entropy by restricting, top-down, the degrees of freedom of the constituent particles; simultaneously, the overall system's phase space increases, thereby satisfying the requirements of the second law of thermodynamics.[1] The newly organized regime shows emergent macroscopic properties that de facto cannot be derived from the laws and theories pertaining to the microphysical level.

Even in physical dissipative structures there is top-down control. Once each water molecule is captured in the dynamics of a rolling hexagonal Bénard cell it is no longer related to the other molecules just externally; its behavior is contextually constrained by the global structure which it constitutes and into which it is caught up. That is, its behavior is what it is *in virtue of the individual water molecules' participation in a global structure*. Fromm's taxonomy of types of emergence based on different forms of feedback and cause-effect relationships identifies physical dissipative structures as examples of Type II emergence: "scale-crossing (top-down feedback)" indicates that these systems have crossed a barrier to a higher level of organization (Fromm 2005a).

3 Constraints

The type of causation involved in the above processes is best understood as the operation of *constraints* (Juarrero 1999). A pan of water at ambient temperature is a system in thermodynamic equilibrium; as such it has maximum Shannon entropy and information potential – but the uniform distribution or equiprobability of the water molecules means that as a whole the system can do no work. Likewise, a communications system at thermodynamic equilibrium can transmit no actual information.

Context-free constraints that take the system away from *equiprobability* and impose a gradient are necessary before work can actually be performed or information transmitted.[2] By taking the system away from equilibrium, context-free constraints reduce Shannon entropy thereby creating the potential for actual work or information transmission. Embodied as a prior probability distribution, *context-free constraints* turn amorphous Shannon entropy into real potential. But if only context-free constraints are available, message variety will be severely curtailed. At the limit, one message or phenotype would repeat (faithfully replicate) with probability 1 time and again, creating a bottleneck that impedes further increases in complexity and evolution (Moreno & Ruiz-Mirazo 2002).[3]

[1] Rod Swenson (1988) argues that complexification is nature's way of maximizing entropy production.

[2] Pistons are physical examples of context-free constraints, as are a particular language's prior probability of letter distribution, and modular, dedicated neurons or brain areas such as Wernicke's. In Bénard cells the gradient is created by the energy pumped into the system. Anything that acts as a template also essentially serves as a context-free constraint.

[3] The cellular automata simulations of bee foraging confirm this (Gambhir et al. 2004). Nevertheless, the value and importance of faithfully replicating *one* message should not be minimized. Even at this late date in cosmological and biological evolution, mitosis, cell division, and replication remain important components of biological reproduction and evolution.

To create order without stifling variety, *context-sensitive constraints* are necessary. These take the system away from *independence* by making the elements comprising the system interact in such a way that their behavior depends on one another's – and on what went before and what is occurring around them in the environment.[4] Once the probability that some event B will happen depends on and is altered by the presence of or interaction with some other object or event A, the two have become systematically and therefore internally related. When this happens a global structure AB defined by *conditional* probabilities has emerged. The coherence of this complex macrostate is reflected in the conditional probabilities that describe it. To phrase it otherwise, the integrated pattern of a Bénard cell just *is* the novel set of conditional probabilities that describes the range and behavior of microstate arrays. It is important not to reify the emergent. Constraints are not forces operating on isolated and independent systems. There is no need to invoke vitalist entelechies or other such *dei ex machinae.* Whenever components are coupled and interact, it is more accurate to speak of tendencies or propensities described in terms of conditional probabilities (Ulanowicz 2005). *The novel complex integration just is the changed probability distribution of the components' state space.* I call context-sensitive constraints that enable complexification *first-order context-sensitive constraints.* "Propensities, unlike forces, always arise out of a context, which invariably includes other propensities" (Ulanowicz 2005). This means that when components are dynamically coupled and coordinated, "a self-organizing network of components and its environment are in fact one system" (Rocha 2001, p. 97). The causal cascade of integrated propensities can be modeled using Granger causality (Seth 2005, 2006). Wheeler and Clark call it "causal spread" (Wheeler & Clark 1999).

3.1 Integration, Not Fusion

Paul Humphreys suggests that unlike aggregates, whose individual constituents retain their identities, emergence happens when microstates *fuse* (Humphreys 1997). As a result of a *fusion* operation, the components of unified wholes "no longer exist as separate entities and therefore do not have all their individual causal powers available for use at the global level." Because individual components "go out of existence" when fused, concerns about causal overdetermination and the causal closure of the physical are rendered moot and top-down causality becomes possible, Humphreys maintains. And yet, Humphreys' presupposition, that nature operates in classically mechanical fashion, is exposed when he cautions that fusion and multiple realizability are incompatible: that holding that (1) mental states have causal efficacy but only in virtue of being (identical to a set of multiply realized token) brain states, and (2) that mental properties can be variously instantiated in

[4] Context-sensitive constraints exist in metabolism, language, neurophysiology, and chemistry; they include syntactical rules, catalysts, neurotransmitters, and feedback processes. In language, the probability of the next letter in a word is conditioned by the one(s) that preceded it; catalysts increase the probability of chemical reactions; the presence of acetylcholine or dopamine changes the probability of a nearby neuron's firing.

components that do not “go out of existence,” reintroduces the threat of overdetermination.

The workings of complex systems tell us otherwise. Phase transitions, symmetry breaking, and other forms of dynamic transformations entrain components without thereby fusing them. Instead, the global patterns that emerge as a result of these qualitative changes are embodied in the conditional probability distribution of their components. The operation of *fusion*, a static notion that implies that once fused, there is no going back, is unlike the operation of *integration,* despite the thermodynamically irreversible nature of the latter.[5] Fusion, like context-free constraints, in the end closes off possibilities in a choke point that cuts off the open-endedness required for evolution. In contrast, context-sensitive constraints represent couplings that are Goldilocks-like – not too tight, not too loose – and that allow the same microstructure to participate in different global dynamics, both synchronically and diachronically. If the disruptive perturbation or fluctuation is strong enough, of course, the global structure will dissolve, but while the constraints hold, the complex dynamics remain integrated and coherent over time (Ulanowicz 2005) – and present emergent properties. The emergence brought about by dynamical integration is nothing but the effects of second-order context-sensitive constraints, embodied as a set of conditional probabilities that are invariant over time and that modulate and direct the behavior of individual – but now no longer independent – microphysical components in such a way that the global dynamics are maintained. In graph-theoretic terms, this unique balance between integration and differentiation can be measured in terms of a network’s *causal density* (the fraction of interactions among nodes in a network that are causally significant). Analysis shows that high causal density is consistent with a high dynamical balance between differentiation and integration, and therefore with high complexity (Seth 2005, 2006). The robustness characteristic of complex systems is thus due to dynamics that are globally coordinated while component details remain distinct; components do not fuse and yet the overall system displays a remarkable resilience and metastability despite radical differences in the arrangements of its component parts.

3.2 Emergent Properties

A key indicator of emergent novelty is that the higher-level variables are not reducible to the aggregation of lower-level ones. The causal relationships that the new codes specify about the higher level are for the most part sealed off from the energetic-type causes operating at the lower level. As a result, the former are to a great extent insensitive to details in the latter – that is, the higher level exists

[5] The irreversibility is provided not by unalterably fusing the micro-level components into a global structure, but by their historicity and context-dependence. These are embodied in the system’s internal dynamics, which “carry on their back” the conditions under which the systems were created and the trajectory they have undergone. Even snowflakes, for example, carry in their very structure information about both the atmospheric conditions that caused them and those they traversed before reaching the ground.

independently of the details of the lower level, which it nevertheless controls. Fromm (2005b) calls this Type IV emergence, where higher levels of complexity cannot be reduced *even in principle* to the direct effect of properties and laws of the elementary components. As evolution progresses, later emergents show increasing diachronic control, constraint, modulation, and regulation by the higher level over the lower. In the course of evolution, as I will describe, the higher level becomes more and more autonomous and self-directed as its capacity for constraint, modulation, and regulation is increasingly modularized and decoupled from energetic exchanges. And decoupling from buffeting by the external world is precisely the sort of trait one is searching for in any kind of free will worth wanting.

Even more importantly for our purposes, since levels are screened off from each other, new levels of dynamical organization involve the appearance of new capabilities at the uppermost level (Salthe 2001). The differentiation into a hierarchical system that is nevertheless dynamically integrated creates entirely new realms of meaning and possibilities. The overall system AB represents an enlarged phase space with more degrees of freedom than constituents A and B had separately: sentences can say things that words alone can't; words have significance that letters lack; amino acids folded into a protein structure possess properties that the amino acids on their own do not; neural patterns can carry meaning which individual neuron firings do not, and so forth. In contrast to context-free constraints, which limit message variety, the closure of context-sensitive constraints thus opens up possibilities and increases Shannon entropy by freeing up previously unavailable message variety.[6] Even in an artificial neural network integration can cause the appearance of *semantics*.[7] The top-down constraints that govern the complex system are selective – and causally effective – in terms of meaningful criteria determined at the higher level.

3.3 Minimal Functionality

Let us now return to the example of dissipative structures. Of particular significance for purposes of this paper is that particle A and particle B comprising a Bénard cell, for example, can no longer be defined in terms of their internal properties as (isolated, independent) elements. They have acquired a new identity in virtue of their participation in the global structure: they are now *components* of a differentiated, hierarchical, contextually embedded organization, the overall global system identifiable as Bénard cell AB. As such they are now *functional,* if only in the minimal sense that their behavior is "in the service" of sustaining the coherence of the cell as a whole. Their behavior, that is, is in virtue of their contribution to the whole. By definition, properties such as *functionality* that define a *relational* dynamics exist only after integration takes place. In addition to the weak type of

[6] It doesn't just activate previously existing potential; complexification creates heretofore nonexistent possibilities – and therefore new directions for the system.

[7] See Wheeler and Clark 1999 for additional examples of new properties created by the *causal spread* of context-sensitive constraints.

decoupling we see in Bénard cells – that between controller and controlled (Moreno & Ruiz-Mirazo 2002) – the properties defining the relational pattern of any global complex structure are not identical with those of its material basis. Bénard cells, for example, can also appear in a range of viscous fluids other than water; the identity of Bénard cells, therefore, is to be found in relational properties that signal an incipient decoupling of the global macrostate – the rolling hexagonal cell, in this case – from its specific material composition. Multiple realizability with top-down causation is a characteristic feature of complex dynamical systems, along with a partially arbitrary relationship between types and tokens that can embody symbolic, code-type representational properties (Wheeler & Clark 1999).

3.4 Top-Down Causation

The overall flavor of Humphreys' approach, however, remains sound: a token physicalism that nonetheless leaves room for a strong form of downward causation that is made possible when constituents are coupled and integrated – not fused – into a complex whole. Once the higher level dynamics self-organize, the state space of each of the components is indeed no longer what it was before: the components are now restricted to a smaller volume of their earlier state space, one that embodies the coherence of the constrained behavior of the particles (Brooks & Wiley 1988). We saw how as a function of the interdependence created by first-order context-sensitive constraints and the conditional probabilities[8] they represent, an integrated dynamical pattern, the AB system – the Bénard cell in this case – appeared. Once captured in the rolling hexagonal cell, each molecule's behavior depends on and is restricted top-down in virtue of its being taken up into the system's overall coherence and integration. Haken (1983) calls this phenomenon the "slaving" principle. Such part-whole and whole-part relationships are thus mereologically causally effective (Ellis 2007; Wheeler & Clark 1999). They do not operate as efficient causes, however; instead, by functioning as formal and final causes they do not violate (the efficient, energetic) causal closure at the particulate level. Unlike near-equilibrium, Markovian processes, context-dependence makes complex systems sensitive to initial conditions. Diachronically, emergent properties of complex dynamics will also be embodied in or realized by different microphysical configurations – this is still physicalism and not dualism. Nevertheless, even at the level of physical dissipative structures, differentiation into a hierarchical structure with top-down causally effective power appears. And it all happens without danger of overdetermination.

Philosophical discussions concerning the possibility of top-down causation by supervenient states inevitably raise the following concern: If mental states have causal efficacy in virtue of being physical states, and one assumes the causal closure of the physical, how can either overdetermination or redundancy be avoided? Hinton, Plaut, and Shallice's (1993) neural network described below illustrates that the answer concerning top-down causation can be found in the reverse formulation:

[8] These conditional probabilities are "in the function of" the overall AB system.

"Brain states can have causal efficacy *in virtue of being (embodying, being entrained into)* complex neural states with mental emergent properties." Put another way: brain states can have causal efficacy *qua* embodying mental properties. The integration without fusion of particles into coherent dynamical patterns that are embodied in a set of conditional probabilities can thus account for strong top-down *mental* causation without thereby risking overdetermination.[9]

Consider an admittedly artificial system such as Hinton and colleagues' (1993) word reading neural network. If lesioned after being trained without feedback loops, it makes errors characteristic of surface dyslexia: presented with the word *bed* the system's output might be *bad.* In contrast, if the neural network is trained with feedback loops – that is, with context-dependent constraints – and is then subsequently lesioned below the feedback loops, outputs characteristic of deep dyslexia appear instead: the same input might elicit the output *cot.* If presented with *band*, its output might be *orchestra.* Hinton and colleagues conclude that the only explanation for this remarkable phenomenon is that feedback caused the system to self-organize a semantic attractor, which subsequently constrained the output. Following Wheeler and Clark, we can also say that these self-organized states are representational and symbolic – and are causal *qua* representational and symbolic – insofar as *what matters is that the output of the system is dependent on the hidden variable configuration not because of the configuration's intrinsic physical properties but because of the information it carries.* In Hinton and colleagues' network, as in the case of natural complex systems, the physical (token-identified) configuration instantiating the same semantic attractor (type-identified) might be different on different runs of the same network, or in different networks. Likewise, the physical configuration embodying different semantic attractors (type-identified) might be the same (token-identified) among different networks. In the remarkable example above, the neural network's output – *cot* or *orchestra* – is produced top-down *by virtue of the emergent* semantic *properties of the global attractor,* a striking illustration of the way that in complex systems *top-down selection is carried out according to criteria of suitability determined at the higher level.* These top-down selection criteria define the direction of the system. In the word-reading example, the microphysical configuration of the network exercises its causal efficacy and produces a particular output *in virtue of being entrained into a higher-level dynamics that thereby embody (emergent) semantic features.*[10] As a result the output is thereby controlled and determined semantically.

The effects caused by the integration and context-sensitivity of complex systems can thus explain why the customary definition of *supervenience* – that there can be no differences in the mental without a corresponding difference in the physical – fails, as does the concept *realization,* which has of late replaced *supervenience* as the concept du jour among philosophers of mind with a

[9] Below I will discuss how various forms of self-organization that appear in the course of evolution also represent the progressive internalization of the system's regulatory mechanisms.

[10] That is, by virtue of their integration into a coherent dynamic pattern, a process resulting from context-dependent constraints. Recent findings (O'Connor 2004) showing that the biology of dyslexia varies with culture is consistent with this claim.

reductionist bent. Both fail to account for the possibility that two different complex systems with identical physical configurations might show different causal powers *depending on whether or not the components are entrained into a high-level complex dynamics and in virtue of the differences in initial conditions and historical trajectories of these complex dynamics.* Alternatively, the concept of *realization* fails to account for the different high-level properties instantiated in microphysical features depending on whether or not these are dynamically integrated into a global attractor at all. The equivalence class created through integration is established by the emergent content the global attractor embodies and is fixed by the counterfactuals which that content supports. If not entrained into a higher-level order parameter at all, the same (token-individuated) physical configuration or array will constitute only an aggregate and will not embody a high-level property. In the word-reading case, without feedback the output would not be produced *in virtue of* (or caused top-down by) a high-level feature at all.[11] Without such top-down constraints the output would be different – *bad*, not *cot,* say.[12] Carl Gillett argues that if some microphysical events have different causal powers depending on whether or not they are integrated into a complex system, philosophers and scientists must abandon the "Completeness of Physics" principle, the claim that "all microphysical events are determined by prior microphysical events and the laws of physics" (Gillett 2002) – whether or not these realize complex structures. Doing so, however, does not thereby require abandoning physicalism, the principle that "all individuals are constituted by, or identical to microphysical individuals, and all properties are realized by, or identical to, microphysical properties" (Gillett 2002).[13]

4 Autocatalysis

In this next section I describe the increasing autonomy that characterizes the evolution of complex dynamical systems from dissipative structures to biological and neural processes. Despite the incipient decoupling between their higher-level properties and their material basis, physical dissipative structures such as Bénard cells, hurricanes, and dust devils are only weakly emergent because the boundary constraints[14] that create and maintain them are exogenously imposed.[15] In the case

[11] This also shows why explaining complex systems necessarily involves reference to the trajectory of which the phenomenon investigated is the end point.

[12] This characteristic accounts for the higher level's support of counterfactuals.

[13] Since higher-level properties usually exist at different temporal and spatial scales from the microphysical properties that realize them, even the identity criterion may have to be rethought.

[14] Energy source and container boundary.

[15] In his defense of only a weak form of emergence, Bedau does not consider the *origin and construction* of ALife's glider streams and glider guns. Unlike these, biological systems not only manage and maintain the flow of energy through them to ensure continuation and preservation; their endogenous dynamics are responsible for their very origin and construction. Claims of stronger forms of emergence and autonomy are thus more plausible than Bedau maintains.

of the word-reading neural network, the feedback loops are imposed from without. Nevertheless, as we saw, the top-down regulation and modulation they exercise on their constituents are not trivial. By reversing the exergonic direction of classically thermodynamic processes and bringing a measure of control inside these systems– thereby retarding their entropic dissolution – the appearance of endergonic processes creates an integrity and self-direction that were previously absent. Nevertheless, the term *autopoiesis* or self-construction is customarily reserved for systems *that construct themselves as a result of their own endogenous dynamics.* Unlike physical dissipative structures, such self-organized systems create the very constraints that control the matter-energy flows that make the structure possible; in other words, the constraints giving rise to self-organization have themselves now been imported into the system's dynamics.[16] This capacity appears with the emergence of chemistry, and not before (Ruiz-Mirazo & Moreno 2000); only then does the recursive production of structures become possible. As a result of this organizational complexity, variety increases.

Autocatalytic cycles in which the product of the catalytic process is necessary for the activation of the process itself are paradigmatic examples of *autopoiesis*. The endogenous[17] dynamics of autocatalysis itself – reactions where the product of the process is necessary for the process itself – create the very constraints within which complexification occurs. Acting as a first-order context-sensitive constraint that drives the system farther from equilibrium, the dynamics' own runaway *positive feedback* takes the system to a critical threshold where, once again, a phase change occurs and the system is discontinuously driven to a new mode of organization, a novel order parameter with newly emergent properties. Once the positive feedback closure that characterizes these endogenous dynamics takes place, the bifurcation marks the transition into a new, hierarchically differentiated, system with emergent properties and whose microstates are constrained by the global structure top-down. That the context-sensitive constraints that make it possible are produced by the system's own dynamics marks a nontrivial distinction that signals the appearance of a measure of autonomy different from and more significant than that found in physical dissipative structures.[18] This newly emergent level of organization is another step in evolution's seemingly relentless creativity, one that also brings regulatory processes inside the system and thereby secures an additional measure of decoupling from spontaneous, that is, exergonic, tendencies.

[16] Note that "in" or "inside" in this context is used to mean that the control issues from the system's endogenous dynamics and in the top-down fashion described above.

[17] This is a somewhat misleading term given that these are open systems whose import and export of matter and energy imply feedback loops with the environment. Nonetheless, one can call the dynamics "endogenous" insofar as the context-sensitive constraints are self-produced.

[18] Phase changes in artificial neural networks as well as in autocatalysis are precipitated by the closure of circular causality, which appears to be the causal agent responsible for the integration – not fusion – of parts into dynamical wholes.

And it happens because the openness thereby achieved satisfies the second law of thermodynamics (Swenson 1988).

Biologist Francisco Varela identifies autonomy as

> mechanistic (dynamic) systems defined as a unity by their organization. We shall say that autonomous systems are organizationally closed. That is, their organization is characterized by processes such that (1) the processes are related as a network, so that they recursively depend on each other in the generation and realization of the processes themselves, and (2) they constitute the system as a unity recognizable in the space (domain) in which the processes exist. (Varela 1979, p. 55)

In contrast, Moreno and Ruiz-Mirazo (2002) reserve the classification *basic autonomy* for systems whose boundary constraints are endogenously produced. In my view, however, much more important than where one chooses to apply the term *autonomy* is the progressive "internalization of regulatory processes" that marks the evolution from the proto-autonomy of physical dissipative structures to the strong autonomy present in biological hereditary autonomous systems, and finally to that displayed in the exercise of human free will.

4.1 Selection Process

A. Selecting Materials to Import

As befits any complex system, the new multiply realizable macro-level system that emerges from autocatalytic closure is robust; it persists despite the deletion of individual components and perturbations entering from the environment. But even more significant is a novel feature: autocatalytic cycles can change the very type of components that make them up. As open systems that need to import matter and energy, autocatalytic cycles *can actively select* the molecules that participate in the overall cycle's continued coherence. Acting top-down as *second-order context-sensitive constraints*, the overall cycle adds, replaces or deletes component molecules in such a way that the far from equilibrium conditions necessary for its dynamical persistence are maintained. Thus chemical complexity both creates but also actively preserves itself *as itself,* again through a natural selection process whose fitness criterion is the persistence of the whole (Ulanowicz 2005) – and despite the radical alteration of components.[19] Because the criteria of selection are determined at the level of the emergent global structure, the system is that much more self-directed and autonomous. Ulanowicz describes how the selection process is such that autocatalytic cycle ABC might eventually transform into FDE (Ulanowicz 2005) while continuing to perform the same function. Because of the multiple realizability created by context-sensitivity, the range of token microstate arrays satisfying the newly emergent chemical *type* is in principle open-ended and indefinite. Autocatalysis thus represents the creation of unlimited ways of expressing the newly emergent dynamics. As phenomena that couple and thereby integrate internal processes and interactions with the environment, autocatalytic

[19] Persistence of the whole underwrites the equivalence class and its counterfactuals.

cycles have the potential to replicate and reproduce their dynamics in "an unlimited variety of equivalent systems, of ways of expressing that dynamics. These systems are not subject to any predetermined upper bound of organizational complexity, even if they are, indeed, [subject] to the energetic-material restrictions imposed by a finite environment and by the universal physico-chemical laws" (Ruiz-Mirazo et al. 2004, p. 331). Autopoiesis is thus a mechanism for creating unlimited *type variety – functional* type variety in this case. Once again, this newly self-organized system increases the production of Shannon entropy by freeing up heretofore nonexistent capabilities, thereby providing the emergent features and open-ended variety on which selection can operate.

B. Changing the Environment

To persist as complex structures, a constrained flow of energy that sustains them far from equilibrium is necessary. To provide this flow, chemical autocatalytic cycles can also actively alter even the conditions of their environment. And here too, the system itself does so in virtue of emergent criteria determined at the level of the global structure. Through selective transport processes, for example (Collier 1986), they alter the outside concentration levels. Unlike examples from the Game of Life[20] where the boundary conditions are established from without, biological systems not only create themselves endogenously, they also actively modify their environment, for example, the outside concentration levels, in order to ensure their continuous self-construction and persistence. In doing so, they realize an additional degree of autonomy that is both materially grounded and nontrivial. Whether one chooses to label this phenomenon the emergence of a "proto-self" (Ulanowicz 2005) or "agency" (Ruiz-Mirazo & Moreno 2000), the important thing to note is the increasing self-determination that progressively appears in the course of evolution.

4.2 Criteria of Suitability: Semiosis

That individual catalysts comprising the overall autocatalytic hypercycle are selected for inclusion and or discarded, and transport processes are altered, according to fitness criteria determined at the emergent, higher level cannot be overemphasized. In the case of autocatalysis, molecules are selected or discarded depending on whether they contribute to metabolic efficiency. We saw that even in those neural networks trained with feedback loops, too, output production is constrained by higher-level criteria – in Hinton and colleagues' example by the output's semantic appropriateness to the input, so to speak. In complex systems, that is, criteria of suitability of inclusion (Ellis 2007) – what counts as a "detail" as opposed to what counts as "essential" – are partitioned in terms of the goals of the newly organized

[20] Bedau bases his arguments for weak ontological emergence almost exclusively on the Game of Life (Bedau 2002).

system, and are formulated in terms of the goals of the new level, which makes a "significant (to it) interpretation of events at the lowermost level" (Salthe 2001). A normative and semiotic or representational process is at work in this formal cause-like process of top-down selection (Salthe 1998, 1999, 2001, 2008; Wheeler & Clark 1999). Instead of processes being determined by energetic considerations alone, selection based on criteria of suitability determined at the higher level defines a direction that is increasingly autonomous and decoupled from merely energetic considerations. Corresponding to this ontological transformation, the formal logic previously appropriate for homogeneous classes suddenly becomes unsuitable for heterogenous classes (Elsasser 1998). Yet another "barrier of relevance" has been crossed (Fromm 2005a, 2005b).

4.3 Summary Thus Far

More generally: whether exemplified in physical Bénard cells, chemical B-Z reactions, or later on in evolution, in biological *functions*, self-organization signals the creation of a hierarchy, a new ontological type constituted by a higher *relational* level that is *multiply instantiated* in the lower level. The appearance of functional units that are not token-token identical with their material constituents indicates the emergence of a new ontological *type* of existent defined by both integration and multiple realizability.[21] In a sudden burst of entropy, this novel phenomenon, whose defining criterion is the possibility of being instantiated by an indefinite *range* of token microstates, emerges. That the self-organized structure is determined top-down, shows lawful regularities, and supports counterfactuals regardless of its material composition is reason enough to consider it ontologically emergent – if only weakly so at the chemical level.

Whether one chooses to call it decoupling, information closure, proto-autonomy, or proto-agency, from the chemical stage onwards these complex endergonic phenomena embody an increasingly autonomous capacity for self-direction. It is an autonomy that is far from trivial in that the systems' own endogenous dynamics generate the constraints whereby the constraints themselves are re-generated and evolve. The dynamics themselves also select and delete components according to fitness criteria determined at (and meaningful for) the level of the coherent whole. The dynamics even actively bring about the environmental changes necessary to accomplish that goal.[22] Although only the first step towards a stronger form of decoupling such as exists between genotype and phenotype, autocatalytic reactions are thus *basically autonomous* (Ruiz-Mirazo & Moreno 2000).

[21] The ideas presented here are consistent with Koch and Tononi's (2008) arguments for an Integrated Information Theory of consciousness, according to which consciousness requires the availability of a large repertoire of states belonging to an integrated system. The resolution of that large and integrated repertoire into one particular state constitutes consciousness.

[22] One way it does so is by altering the outside concentration levels to allow selective exchange of molecules and energy with the environment.

4.4 Biological Function

Basic replication in the sense of growth and division is possible even at the level of chemistry, and, as we saw, autocatalytic cycles can generate new *types* of functional components and thereby expand variety (Moreno 2008). At a certain threshold, however, structural complexity becomes brittle and the potential of truly open-ended evolution requires the production of hereditary lineages that, on the one hand, can preserve any novelty constructed at a given stage while on the other simultaneously allowing continuous catalytic creation. Lila Gatlin claims that the emergence of this double mechanism was a major breakthrough. Biological systems discovered how to retain the faithful replication provided by context-free constraints while at the same time allowing context-sensitive constraints to expand their phase space. Agreeing with Gatlin, Moreno claims that the key mechanism that *strongly* decouples the biological from the chemical level and allows the functional open-endedness of *biological* organization is the evolutionary emergence of such *dynamical decoupling* (Moreno 2008).

Dynamical decoupling occurs when living systems produce two different *types* of *functional* components: first a new type – such as DNA – that serves as regulatory *record* of earlier functions and guarantees their faithful replication. The second is the persistence of the earlier dynamics that ensures the continual production of new types of catalysts. The latter continues the job of ongoing complexification while the former guarantees the preservation and heredity of important functions. The latter also continues to evolve towards greater metabolic efficiency but does not involve itself in the generation of the new *records* (Ruiz-Mirazo et al. 2004). Such *hereditary* autonomous systems thus depend on "two types of interdependent macromolecular components: some carrying out and coordinating directly the self-construction processes (catalysts); others storing and transmitting information which is relevant to carry out efficiently those processes in the course of subsequent generations" (ibid., p. 337). For purposes of this paper I wish to emphasize that with this new evolutionary breakthrough the hereditary/regulatory process itself is brought even further into the system's endogenous dynamics and modularized, thereby making it even more autonomous and self-directed.

Since these two different *types* of functional components cannot be directly linked on the basis of their intrinsic properties, their interaction [23] must be indirect – and

[23] All interactions "transfer (record) information about the state of the measured object into the state of the measuring device" (Rojdestvenksi & Cottam 2005, p. 116). In classical external measurements, the measurement becomes more exact as the measuring device becomes as complex as the measured. In those cases where the description – or record – of a system is embedded into a system itself – as for example, when a meta-language is embedded into a language, or a genetic apparatus is embedded in the metabolic system – an *internal* form of measurement is involved whose recursive embedding limits the measurement's possible accuracy. "Because the increase in complexity of the encoding results in a corresponding increase in complexity in the measuring device, the measured system becomes more complex as a result of the measurement itself" (ibid. 119). Evolution is thus a form of infinite recursive embedding: "Life evolves as a measurement of the environment, and becomes, through living organisms, embedded in this environment, which affects its further development by its own presence." And so forth.

integrated in a novel way. A *new type of semiosis* governing the criteria of selection thus becomes necessary: "It is only once hereditary autonomous systems start producing 'informational' components and mechanisms (i.e., once a *translation code* appears between two very different types of functional components in the system) that the 'genotype-phenotype' distinction becomes really significant" (ibid., p. 337) and for the first time in evolution a *code-type information*[24] semiosis appears.

Biological hereditary functions thus represent yet another novel way of integrating organizational structure. With the appearance of the genetic network, function becomes structure (Haken 1983): the product of previous *first-order context-sensitive* constraints becomes phylogenetically frozen into a structure and encapsulated as a higher-level *second-order context-sensitive* constraint. Because of the additional measure of decoupling accomplished by this novel level of integration between two different types of functional components, Ruiz-Mirazo and colleagues (2004) identify the appearance of this phenomenon with the emergence of what we can call the *strong autonomy* of biological systems. It is equally important to note that, once again, further decoupling, modularization, and internalization of regulatory processes bring with them even greater autonomy and self-direction.

5 Individuation

Before I close by directly addressing the topic of free will, I note one additional feature of complex systems with indirect implications to free will. One of the features of complex systems missed by philosophers working on the assumption that nature consists exclusively of simple aggregates near equilibrium is the historicity mentioned earlier in passing. Classical thermodynamics discovered an arrow of time – but there is no history in the closed structures of classical, near-equilibrium thermodynamics: Markovian processes are systems whose future is independent of their history; only their future, not their past, is packed into the present. History proper presupposes the integration and context-embeddedness provided by context-sensitive constraints. Open systems with progressively higher, coherently integrated levels of organization are not already there waiting to be unfurled; they embody instead uniquely individuated trajectories embodying irreversible discontinuities – both phylogenetic and ontogenetic – that emerge over time, while simultaneously remaining open to the future. As is true of all complex systems, during their lifetimes human beings too progressively evolve into uniquely individuated, multifaceted, and complex persons. The person's hierarchical dynamics become increasingly specified (Salthe 1998, 2001) and the person progressively becomes his or her own self. Behavior that issues from this highest

[24] By "information" Ruiz-Mirazo et al. (2004) mean a kind of causal connection in a system by which some (quasi) inert material patterns constrain, through a certain mechanism of "translation-interpretation," the metabolic dynamics of the system.

organizational or integrative level is not only increasingly autonomous and self-directed; it is increasingly, and uniquely, therefore, "one's own."

6 Free Will

As examples of dynamic decoupling, Moreno (2008) identifies the emergence of the neural organization, linguistic communication, and tools. I propose that additional dynamical decoupling of the regulatory process occurred with the appearance of the frontal cortex. I claimed earlier that the evolutionary open-endedness of biology appeared with the emergence of a double organizational structure that combines two different types of functional components – internalized quasi-inert *records* such as the genetic network that allow the storing and transmission of biological information, alongside the continuation of processes that increase the efficiency of metabolic tasks. I now postulate that a homologous transformation also occurred in the neural system with the appearance of a double organizational neurological structure that integrates new types of functional components with emergent properties – call them stable *mental records* (including conscious intentions and memories, and qualia) – that store, reproduce, and transmit meaning, and do so by constraining lower-level neurological processes that themselves self-organize behavior.[25] Since, once again, these two different types of functional components cannot be related directly through their intrinsic properties, this more recent evolutionary breakthrough, I suppose, also brought with it a new type of semiosis based on human symbolic language and communication with a higher-level translation code. In this manner an additional regulatory function was brought inside the system dynamics and modularized – and its subject freed even further from outside direction and control. In other words, because the criteria on the basis of which the top-down selection process is carried out are partitioned in terms of goals appropriate to the higher level, an even greater decoupling from energetic forces appeared with the emergence of the human mind – with self-consciousness, qualia, and the realm of the linguistically symbolic. If Rod Swenson is correct, this new evolutionary stage is just one more step in nature's relentless drive to maximize entropy production. But once it occurred we human beings became capable of intentional *actions*, that is, behavior that issues from and is determined by self-conscious symbolic constraints – by our uniquely individuated, symbolically organized character, in other words. Human beings thus embody the potential for an additional degree of autonomy – linguistically articulated and self-consciously chosen autonomy – that distances us even further from merely energetic exchanges (Donald 1991). This symbolic decoupling, I submit, warrants calling the bearers of such complex organization *maximally autonomous*, and behavior constrained symbolically in this manner, I maintain, is nothing less than an exercise in free will.

[25] Mental representation should be interpreted along the lines of Wheeler and Clark (1999). Mirror neurons might an early cortical analog of the "records" that permitted open-ended evolution.

7 Further Research: Boundaries

Dynamical closure always generates a boundary between the new emergent and the background. In the case of autopoietic structures the boundary is self-created by the very dynamics of the system. It can take the form of either a physical permeable boundary[26] between the new system and its environment, as is the case of a cell's membrane,[27] or as a dynamic phase separation between the emergent structure[28] and the environment, or between the structure and its components. Phase separation is clearly demarcated wherever crisp differences in time scales exist between the higher-level emergent dynamics and processes at the lower level. The dynamics of the higher informational regulatory system – the genetic and neural systems – often operate at a slower and longer time scale than that of their constituents. Salthe (2001) notes that higher-level laws and principles apply to dynamics that are slower and longer than those of the lower level; that these slower and longer dynamics nonetheless constrain the lower level constitutes what Haken calls "slaving." But the higher level is not always slower or faster than the lower: the regulatory genetic system operates faster than the metabolic system; the regulatory neural system works faster than the metabolic system.

Research on the role of time in brain processing is still in its early stages (Carey 2008). However, in light of the controversy occasioned by Libet's work, further research concerning the role of phase differences in intentional behavior is warranted, especially concerning phase differences between the regulatory informational system and the lower-level functional networks under the former's control. Since regulatory records operate in a time-independent mode and exert top-down causal influence on lower-level metabolic processes operating in a dynamic and time-dependent mode, arguments such as Libet's attempts to refute the possibility of free will (Libet 2004) based on temporal relations between neural events and experience call for closer scrutiny. Libet's experiments clearly disallow mechanistic interpretations of the concept of free will that conceive of voluntary intentions as conscious representations distinctly separate and preceding the actions they forcefully bring about as efficient causes. In light of the fact that functional, informational, symbolic, and representational processes operate as formal – not efficient – causes operating as second-order context-sensitive constraints that temporally span both the onset and terminus of behaviors under their control, the

[26] The boundary of a dissipative structure cannot entirely seal off the system from energetic and material exchanges.

[27] Accustomed to thinking in terms of rigid edges and borderlines that mark off crisp logical categories from undesirable fuzzy concepts, we often fail to recognize that a cell membrane is not a wall but an *active site* without which, for example, hearing would be impossible (Cilliers 2001, 2004). Boundaries as active sites need further research, not least because they complicate questions of complex system identity, a topic related to that of free will.

[28] The term "structure" should not imply reification. We are describing not a static thing but a "structure of process," a network of interconnected dynamical events (Earley 1981).

question of when an action is voluntary and when it is not must be rethought.[29] The approach described here, which is consistent with Andy Clark's (1997) and Scott Kelso's (1995), provides a better framework for understanding how volitions can meaningfully guide actions that take time to complete; how conscious intentions can operate as standing or structural causes of long-term actions such as "running for public office"; how integrated neural process can be both representational and serve as the *mental* causes of actions, etc. The important role of context sensitivity, however, also suggests that philosophical investigation could more fruitfully be redirected towards questions such as What first-order contextual conditions contribute to the development of *akrasia*, or to its opposite, robust character formation? When, dynamically speaking, is psychological intervention likely to make an impact? Under what conditions do human personalities mature into adaptive and resilient adults – or into rigid and therefore brittle individuals? And so forth. This is in contrast to rehashing over and over again the old question, "Do we have free will or not?" and assuming the answer will be formulated in terms of the kind of causality suitable for billiard balls, but certainly not complex dynamical systems.

Finally, determining that there exists a variety of free will that can serve as the ground of moral responsibility and explanation but which is inextricably linked to context and environmental embeddedness also brings with it the recognition that we are all responsible for contributing to the psychological, social, and economic conditions that both enable such autonomy and facilitate its nourishment.

References

Auletta, G., Ellis, G.F.R., Jaeger, L.: Top-down causation by information control: From a philosophical problem to a scientific research programme. Journal of the Royal Society Interface 5(27), 1159–1172 (2008)

Bedau, M.: Downward causation and autonomy in weak emergence. Principia revista internacional de epistemologia 6, 5–50 (2002); Reprinted in Bedau and Humphreys 2008

Bedau, M., Humphreys, P. (eds.): Emergence: Contemporary readings in the philosophy of science. MIT Press, Cambridge (2008)

Brooks, D., Wiley, E.: Evolution as entropy, 2nd edn. University of Chicago Press, Chicago (1988)

Carey, B.: Anticipating the future to 'see' the present. New York Times, June 10 (2008)

Cilliers, P.: Boundaries, hierarchies and networks in complex systems. International Journal of Innovation Management 5(2), 135–147 (2001)

Cilliers, P.: Knowledge, limits and boundaries. Futures 37, 605–613 (2004)

Clark, A.: Being there: Putting brain, body and world together again. MIT Press, Cambridge (1997)

Collier, J.: Entropy in evolution. Biology and Philosophy 1, 5–24 (1986)

Crutchfield, J.: Is anything ever new? Considering emergence. In: Cowan, G.A., Pines, D., Meltzer, D. (eds.) Complexity: Metaphors, Models, and Reality. Westview Press, Cambridge (1999); Reprinted in Bedau and Humphreys 2008

[29] See Juarrero 1999 for an attempt to reinterpret intentions as dynamic attractors and action as behavior constrained by such attractors.

Davies, P.C.W.: Emergent biological principles and the computational properties of the universe. Complexity 10(2), 11–15 (2004)

Donald, M.: Origins of the modern mind: Three stages in the evolution of culture and cognition. Harvard University Press, Cambridge (1991)

Earley, J.: Self-organization and agency: In chemistry and process philosophy. Process Studies 11, 242–258 (1981)

Ellis, G.F.R.: On the nature of causation in complex systems. Royal Society of South Africa (2007), http://www.sabinet.co.za/abstracts/royalsa/royalsa_v63_n1_a6.xm

Elsasser, W.M.: Reflections on a theory of organisms: Holism in biology. Johns Hopkins University Press, Baltimore (1998)

Fromm, J.: Ten questions about emergence (2005a) arXiv:nlin/0509049v1 [nlin.AO]

Fromm, J.: Types and forms of emergence (2005b) arXiv:nlin/0506028v1 [nlin.AO]

Gambhir, M., Guerin, S., Kauffman, S., Kunkle, D.: Steps toward a possible theory of organization. In: Proceedings, International Conference on Complex Systems (ICCS), Boston, MA (2004)

Gillett, C.: The varieties of emergence: Their purposes, obligations and importance. Grazer Philosophische Studien 65, 95–121 (2002)

Guerin, S., Kunkle, D.: Emergence of constraint in self-organizing systems. Nonlinear Dynamics, Psychology and Life Sciences 8(2), 131–146 (2004)

Haken, H.: Synergetics: An introduction. Springer, Berlin (1983)

Hinton, G., Plaut, D.C., Shallice, T.: Simulating brain damage. Scientific American 269, 76–82 (1993)

Humphreys, P.: How properties emerge. Philosophy of Science 64, 1–17 (1997); Reprinted in Bedau and Humphreys 2008

Juarrero, A.: Dynamics in action: Intentional behavior as a complex system. MIT Press, Cambridge (1999)

Kelso, J.A.S.: Dynamic patterns: The self-organization of brain and behavior. MIT Press, Cambridge (1995)

Koch, C., Tononi, G.: Integrated information theory of consciousness (2008), http://www.spectrum.ieee.org/jun08/6278/3

Libet, B.: Mind time: The temporal factor in consciousness. Harvard University Press, Cambridge (2004)

Liljenström, H., Svedin, U. (eds.): Micro, meso, macro: Addressing complex systems couplings. World Scientific, Hackensack (2005)

Moreno, A.: Organization and evolution: The principle of dynamical decoupling. Paper presented at Complejidad 2008, Havana, Cuba (unpublished) (2008)

Moreno, A., Ruiz-Mirazo, K.: Key issues regarding the origin, nature and evolution of complexity in nature: Information as a central concept to understand biological organization. Emergence 4(1&2), 63–76 (2002), http://www.informaworld.com/smpp/title~content=t787353389~db=all~tab=issueslist~branches=4-v4

O'Connor, A.: Biology of dyslexia varies with culture, study finds. New York Times, September 7 (2004), http://www.nytimes.com/2004/09/07/science/07read.html?_r=1&scp=1&sq=Biology%20of%20Dyslexia%20varies%20with%20Culture&st=cse&oref=slogin

Rocha, L.M.: Evolution with material symbol systems. BioSystems 60, 95–121 (2001)

Rojdestvenksi, I., Cottam, M.G.: Time rescaling and generalized entropy in relation to the internal measurement concept. In: Liljenström, H., Svedin, U. (eds.) Micro, meso, macro: Addressing complex systems couplings, World Scientific, Hackensack (2005)

Ruiz-Mirazo, K., Moreno, A.: Autonomy and emergence: How systems become agents through the generation of functional constraints. In: Farre, G.L., Oksala, T. (eds.) Emergence, complexity, hierarchy, organization (Selected and edited papers from the ECHO III Conference), pp. 273–282 (1998); Acta Polytechnica Scandinavica Ma91. The Finnish Academy of Technology, Espoo-Helsinki

Ruiz-Mirazo, K., Moreno Bergareche, A.: Searching for the roots of autonomy: The natural and artificial paradigms revisited. Communication and Cognition – Artificial Intelligence (CC-AI): The Journal for the Integrated Study of Artificial Intelligence, Cognitive Science and Applied Epistemology, Special Issue on Autonomy 17, 209–228 (2000)

Ruiz Mirazo, K., Moreno, A.: The maintenance and open-ended growth of complexity in nature: Information as a decoupling mechanism in the origins of life. In: Capra, F., Sotolongo, P., Juarrero, A., van Uden, J. (eds.) Rethinking complexity: Perspectives from north and south, pp. 55–72. ISCE Publisher (2006)

Ruiz-Mirazo, K., Peretó, J., Moreno, A.: A universal definition of life: Autonomy and open-ended evolution. Origin of Life and Evolution of Biospheres 34(3), 323–346 (2004)

Salthe, S.N.: Semiosis as development. In: Albus, J., Meystel, A. (eds.) Proceedings of the 1998 IEEE ISIC/CRA/ISAS, On Intelligent Systems, pp. 730–735. IEEE Press, Gaithersberg (1998)

Salthe, S.N.: A semiotic attempt to corral creativity via generativity. Semiotica 127, 481–496 (1999)

Salthe, S.N.: Summary of the principles of hierarchy theory (2001), http://www.nbi.dk/~natphil/salthe/Summary_of_the_Principles_o.pdf

Salthe, S.N.: The system of interpretance: Naturalizing meaning as finality. Biosemiotics (2008), http://dx.doi.org/10.1007/s12304-008-9023-3

Seth, A.K.: Causal connectivity of evolved neural networks during behavior. Network: Computation in Neural Systems 16, 35–54 (2005)

Seth, A.K.: Causal networks in neural systems: From water mazes to consciousness. In: Aleksander, I., et al. (eds.) Proceedings, 2006 Meeting on Brain Inspired Cognitive Systems, BICS 2006 (2006)

Seth, A.K., Edelman, G.M.: Neural Computation, 19, 910–33 (2007)

Swenson, R.: Emergence and the principle of maximum entropy production: Multi level system theory, evolution, and nonequilibrium thermodynamics. In: Proceedings of the 32nd Annual Meeting of the ISGSR, pp. 32–52 (1988); Reprinted, http://www.entropylaw.com/thermoevolution12.html

Ulanowicz, R.: A revolution in the middle kingdom. In: Liljenström, H., Svedin, U. (eds.) Micro, meso, macro: Addressing complex systems couplings. World Scientific, Hackensack (2005)

Varela, F.: Principles of biological autonomy. North-Holland, New York (1979)

Wheeler, M., Clark, A.: Genic representation: Reconciling content and causal complexity. British Journal for the Philosophy of Science 50, 103–135 (1999)

6

Toward a Complementary Neuroscience: Metastable Coordination Dynamics of the Brain

J.A. Scott Kelso and Emmanuelle Tognoli

The Human Brain and Behavior Laboratory
Center for Complex Systems and Brain Sciences
Florida Atlantic University
777 Glades Road
Boca Raton, FL 33431
kelso@ccs.fau.edu,
tognoli@ccs.fau.edu

Summary. Metastability has been proposed as a new principle of behavioral and brain function and may point the way to a truly complementary neuroscience. From elementary coordination dynamics we show explicitly that metastability is a result of a symmetry-breaking caused by the subtle interplay of two forces: the tendency of the components to couple together and the tendency of the components to express their intrinsic independent behavior. The metastable regime reconciles the well-known tendencies of specialized brain regions to express their autonomy (segregation) and the tendencies for those regions to work together as a synergy (integration). Integration ~ segregation is just one of the complementary pairs (denoted by the tilde [~] symbol) to emerge from the science of coordination dynamics. We discuss metastability in the brain by describing the favorable conditions existing for its emergence and by deriving some predictions for its empirical characterization in neurophysiological recordings.

Keywords: brain, metastability, the complementary nature, coordination dynamics, consciousness.

1 Prolegomenon

This chapter starts with some *considerata* for science in general and cognitive computational neuroscience in particular. It then focuses on a specific, empirically grounded model of behavioral and brain function that emanates from the theoretical framework of coordination dynamics. This model contains a number of attractive properties, one of which, metastability, has been acclaimed as a new principle of brain function. The term metastability is on the rise; it is well known in physics and has been embraced by a number of well-known neuroscientists. As we explain, it is not the word itself that matters, but rather what the word means for understanding brain and cognitive function. In coordination dynamics, metastability is not a concept or an idea, but a fact that arises as a result of the observed self-organizing nature of both brain and behavior. Specifically, metastability is a result of broken symmetry in the relative phase equation that expresses the coordination between nonlinearly coupled (nonlinear) oscillators. The latter design is motivated by empirical evidence showing that the structural units of the brain which support sensory, motor, and cognitive processes typically express themselves as oscillations with well-defined spectral properties. According to coordination dynamics, nonlinear coupling among heterogeneous components is necessary to generate the broad range of brain behaviors observed, including pattern formation, multistability, switching (sans “switches”), hysteresis, and metastability. Metastable coordination dynamics reconciles the well-known tendencies of specialized brain regions to express their autonomy, with the tendencies for those regions to work together as a synergy. We discuss metastability in the brain by describing the favorable conditions existing for its emergence and by deriving some predictions for its empirical characterization in neurophysiological recordings. A brief dialogue follows that clarifies and reconciles the present approach with that of W. Freeman. Finally, we explore briefly some of the implications of metastable coordination dynamics for perception and thinking.

2 Toward a Complementary Science

Up until the time of Bohr, Heisenberg, and Pauli, scientists debated over whether light, sound and atomic scale processes were more basically particle-like or wave-like in character. Philosophy spoke of thesis and antithesis, of dialectic tension, of self and not self, of the qualitative and the quantitative, the objective and the subjective, as if they were either/or divisions. This tendency to dichotomize, to divide the world into opposing categories, appears to be a “built in” property of human beings, arising very early in development and independent of cultural background (Talbot 2006). It is, of course, central to the hypothetico-deductive method of modern science, which has made tremendous progress by testing alternative hypotheses, moving forward when it rejects alternatives. Or so it seems.

For Bohr, Pauli, and Heisenberg, three giants of twentieth-century science and chief architects of the most successful theory of all time, it became abundantly clear that sharp dichotomies and contrarieties must be replaced with far more subtle and sophisticated complementarities. For all of nature, human nature (and presumably human brains) included. Probably Pauli expressed it best:

> To us the only acceptable point of view appears to be one that recognizes both sides of reality – the quantitative and the qualitative, the physical and the psychical – as compatible with each other. It would be most satisfactory of all if physics and psyche could be seen as complementary aspects of the same reality. (Pauli 1994, p. 260).

The remarkable developments of quantum mechanics demonstrating the essential complementarity of both light and matter should have ushered in not just a novel epistemology but a generalized complementary science. However, they did not. Thinking in terms of contraries and the either/or comes naturally to the human mind. Much harder to grasp is the notion that contraries are complementary, *contraria sunt complementa* as Bohr's famous coat of arms says. One step in this direction might be if complementary aspects and their dynamics were found not just at the level of the subatomic processes dealt with by quantum mechanics, but at the level of human brains and human behavior dealt with by coordination dynamics.

3 Toward a Complementary Brain Science

How might a complementary stance impact on understanding the brain? The history of brain research over the last few centuries is no stranger to dichotomy: it contains two conflicting theories of how the human brain works (see Finger 1994 for an excellent treatment). One theory stresses that the brain consists of a vast collection of distinct regions each localizable in the cerebral cortex and each capable of performing a unique function. The other school of thought looks upon the brain not as a collection of specialized centers, but as a highly integrated organ. In this view, no single function can be the sole domain of any unique part of the cortex. Obeying the old dictum, the holistic brain is greater than and different from the sum of its parts. Like debates on nature versus nurture, learning versus innateness, reductionism versus holism, these two conflicting views of how the brain works have shed more heat than light. Yet surprisingly, the two either/or contrasts still survive. In modern parlance, researchers ask if the brain is "segregated" into its parts or "integrated" as a whole, if information is represented in a modular, category-specific way or in a distributed fashion in which many distinct areas of the brain are engaged, each one representing many different kinds of information.

In the last twenty years or so some new ideas about brain organization have emerged that may provide deeper insight into the human mind, both individual and collective. One step in this direction is by Sungchui Ji (1995). In promoting his "complementarist" epistemology and ontology, Ji draws on the biology of the human brain, namely, the complementary nature of its hemispheric specializations. For Ji, the left and right hemispheres have relatively distinct psychological functions, and "ultimate reality," as perceived and communicated by the human brain, is a complementary union of opposites (Ji 1995). This is a picture painted with a very broad brush. On a much finer-grained scale, Stephen Grossberg (2000) in a paper entitled "The Complementary Brain" has drawn attention to the complementary nature of brain processes. For example, the visual system is divided by virtue of its sensitivity to different aspects of the world, form and motion information being carried by ventral and dorsal cortical pathways. For Grossberg, working memory order is complementary to working memory rate, color processing is complementary to luminance processing, and so forth. Grossberg believes that the brain is organized this way in order to process complementary types of information in the environment. For him, a goal of future research is to study more directly how complementary aspects of the physical world are translated into complementary brain designs for coping with this world.

If the brain, like the physical world, is indeed organized around principles of complementarity, why then do we persist in partitioning it into contraries? What is it that fragments the world and life itself? Is it the way nature is? Or is it us, the way we are? (see how pernicious the either/or is!). Of course, this age-old question goes back thousands of years and appears again and again in the history of human thought, right up to the present (Frayn 2006; Kelso & Engstrøm 2006). Outside quantum mechanics, however, no satisfactory answer from science has emerged. Motivated by new empirical and theoretical developments in coordination dynamics, the science of coordination, Kelso and Engstrøm (2006) have offered an answer, namely, that the reason the mind fragments the world into dichotomies (and more important how opposing tendencies are reconciled) is deeply connected to the way the human brain works, in particular its *metastable coordination dynamics* (e.g., Bressler & Kelso 2001; Jirsa & Kelso 2004; Kelso 1991, 1992, 1995; Tschacher & Dauwalder 2003; Perez Velazquez 2005). Let us summarize some of the general aspects of coordination dynamics, before focusing in on its core mathematical form.

4 Coordination Dynamics of the Brain: Multistability, Phase Transitions, and Metastability

From being on the periphery of the neurosciences for fifty years and more, brain dynamics is steadily inching toward center stage. There are at least four reasons for this. One is that techniques at many levels of description now afford an examination of both structure and function in real time: from gene expression to individual neurons to cellular assemblies, and on to behavior, structures, and their interrelation. The second is that slowly and surely the concepts, methods, and tools of self-organizing dynamical systems are taking hold. It is twenty years since a review article in *Science* laid out the reasons why (Schöner & Kelso 1988). The third is that dynamics is a language for connecting events from the genetic to the mental (Kelso 1995). Dynamics is and must be filled with content, each level possessing its own descriptions and quasi-autonomy (everything is linked, from a particle of dust to a star). The fourth is that empirical evidence indicates that dynamics appear to be profoundly linked to a broad range of disorders ranging from Parkinson's disease to autism and schizophrenia.

The theory of coordination dynamics is based on a good deal of empirical evidence about how brains are coordinated in space and time. One key set of results is that neurons in different parts of the brain oscillate at different frequencies (for excellent reviews see Başar 2004; Buzsáki 2006). These oscillations are coupled or "bound" together into a coherent network when people attend to a stimulus, perceive, think, and act (Eckhorn et al. 1988; Gray et al. 1989; Munk et al. 1996; Bressler 1996; Steinmetz et al. 2000; Mima et al 2000; Fries et al. 2001; Varela et al. 2001; Brown & Marsden 2001). This is a dynamic, self-assembling process, parts of the brain engaging and disengaging in time, as in a good old country square dance. Such a coordinative mechanism may allow different perceptual features of an object, different aspects of a moving scene, separate remembered parts of a significant experience, even different ideas that arise in a conversation to be bound together into a coherent entity.

Extending notions in which the "informational code" lies in the transient coupling of functional units, with physiological significance given to specific phase-lags realized between coordinating elements (König et al. 1996), we propose that phase relationships carry information, with multiple attractors (attracting tendencies) setting alternatives for complementary aspects to emerge in consciousness (Kelso 1994). In the simplest case, oscillations in different brain regions can lock "in-phase," brain activities rising and falling together, or "anti-phase," one oscillatory brain activity reaching its peak as another hits its trough and vice versa. In-phase and anti-phase are just two out of many possible multistable phase states that can exist between multiple, different, specialized brain areas depending on their respective intrinsic properties, broken symmetry, and complex mutual influence.

Not only does the brain possess many different phase relations within and among its many diverse and interconnected parts, but it can also switch flexibly from one phase relation to another (in principle within the same coalition of functional units), causing abrupt changes in perception, attention, memory, and action. These switchings are literally "phase transitions" in the brain, abrupt shifts in brain states caused by external and internal influences such as the varying concentration of neuromodulators and neurotransmitter substances in cell bodies and synapses, places where one neuron connects to another.

Coordination dynamics affords the brain the capacity to lock into one of many available stable coordinative states or phase relations. The brain can also become unstable and switch to some completely different coordinative state. Instability, in this view, is a selection mechanism picking out the most suitable brain state for the circumstances at hand. Locking in and switching capabilities can be adaptive and useful, or maladaptive and harmful. They could apply as easily to the schizophrenic or obsessive-compulsive, as they could to the surgeon honing her skills.

A third kind of brain dynamic, called metastability, is becoming recognized as perhaps the most important of all for understanding ourselves. In this regime there are no longer any stable, phase, and frequency-synchronized brain states; the individual regions of the brain are no longer fully "locked in" or interdependent. Nor, ironically enough, are they completely independent. According to a recent review:

> Metastability is an entirely new conception of brain functioning where the individual parts of the brain exhibit tendencies to function autonomously *at the same time* [emphasis ours] as they exhibit tendencies for coordinated activity (Kelso, 1991; 1992; 1995; Bressler & Kelso, 2001; see also Bressler, 2003). (Fingelkurts & Fingelkurts 2004)

As the Fingelkurtses remark, metastability is an entirely new conception of brain organization, not merely a blend of the old. Individualist tendencies for the diverse regions of the brain to express themselves coexist with coordinative tendencies to couple and cooperate as a whole. In the metastable brain, local and global processes coexist as a complementary pair, not as conflicting theories. Metastability, by reducing the strong hierarchical coupling between the parts of a complex system while allowing them to retain their individuality, leads to a looser, more secure, more flexible form of function that can promote the creation of new information. No dictator tells the parts what to do. Too much autonomy of the component parts means no chance of coordinating them together. On the other hand, too much interdependence and the system gets stuck, global flexibility is lost.

Metastability introduces four advantageous characteristics that neurocognitive models are invited to consider. First, metastability accommodates heterogeneous elements (e.g. brain areas having disparate intrinsic dynamics; brain areas whose activity is associated with the movement of body parts or events in the environment). Second, metastability does not require a disengagement mechanism as when the system is in an attractor and has to switch to another state. This can be costly in terms of time, energy, and information processing. In the metastable regime, neither stochastic noise nor parameter changes are necessary for the system to explore its patternings. Third, metastability allows the nervous system to flexibly browse through a set of possibilities (tendencies of the system) rather than adopting a single "point of view." Fourth, the metastable brain favors no extremes. Nor is it a "balance" of opposing alternatives. For example, it makes no sense to say the brain is 60% segregated and 40% integrated. Rather, metastability is an expression of the full complexity of the brain.

A number of neuroscientists have embraced metastability as playing a role in various neurocognitive functions, including consciousness (e.g. Perez Velazquez 2005; Varela et al. 2001; Bressler & Tognoli 2006; Edelman 2004, 2006; Edelman & Tononi 2000; Freeman & Holmes 2005; Friston 1997; Koch 2004; Sporns 2004). As we explain below, it is not the word itself that matters, but what the word means for *understanding*. In coordination dynamics, metastability is not a concept or an idea, but a consequence of the observed self-organizing and pattern-forming nature of brain, cognition, and behavior (Kelso 1995; Schöner & Kelso 1988; Haken 1996; Kelso et al. 1992). Specifically, metastability is a result of the broken symmetry of a system of (nonlinearly) coupled (nonlinear) oscillators called the *extended HKB model* (Kelso et al. 1990): *HKB* stands for Haken, Kelso, and Bunz (Haken et al. 1985) and represents a core (idealized) dynamical description of coordinated brain and behavioral activity (see e.g. Jirsa et al. 1998). Importantly, it is the symmetry-breaking property of the extended HKB model (Kelso et al. 1990) that has led to metastability and the new insights it affords.

5 The Extended HKB Model

Etymologically, *metastability* comes from the Latin *meta* (beyond) and *stabilis* (able to stand). In coordination dynamics, metastability corresponds to a regime near a saddle-node or tangent bifurcation in which stable coordination states no longer exist (e.g., in-phase synchronization where the relative phase between oscillating components lingers at zero), but attraction remains to where those fixed points used to be ("remnants of attractor~repellors"). This gives rise to a dynamical flow consisting of phase trapping and phase scattering. Metastability is thus the simultaneous realization of two competing tendencies: the tendency of the components to couple together and the tendency for the components to express their intrinsic independent behavior. Metastability was first identified in a classical model of coordination dynamics called the extended HKB (Kelso et al. 1990), and later seen as a potential way by which the brain could operate (Bressler & Kelso 2001; Kelso 1991, 1992, 1995; Fingelkurts & Fingelkurts 2004; Bressler & Tognoli 2006; Friston 1997; Perez Velazquez & Wennberg 2004; Werner 2007).

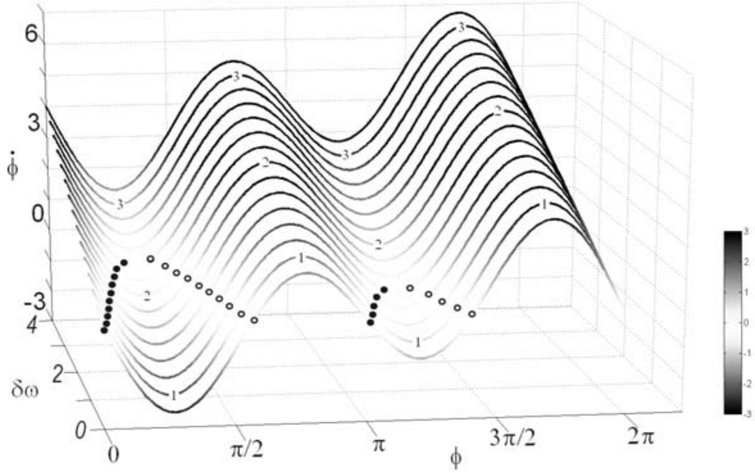

Fig. 6.1. Surface formed with the flows of the coordination variable φ (in radians) for increasing values of $\delta\omega$ between 0 and 4. For this example, the coupling is fixed: a=1 and b=1. When $\dot\varphi$ reaches zero (flow line becoming white), the system ceases to change and fixed point behavior is observed. Stable and unstable fixed points at the intersection of the flow lines with the isoplane $\dot\varphi$=0 are represented as filled and open circles respectively. To illustrate the different regimes of the system, three representative lines labeled 1 to 3 fix $\delta\omega$ at increasing values. Following the flow line 1 from left to right, two stable fixed points (filled circles) and two unstable fixed points (open circles) exist. This flow belongs to the multistable (here bistable) regime. Following line 2 from left to right, one pair of stable and unstable fixed points is met on the left, but notice the complete disappearance of fixed point behavior on the right side of the figure. That is, a qualitative change (bifurcation; phase transition) has occurred. The flow now belongs to the monostable regime. Following line 3 from left to right, no stable or unstable fixed points exist, yet coordination has not disappeared. This flow corresponds to the metastable regime, which is a subtle blend of coupling and intrinsic differences between the components.

The equation governing the coordination dynamics of the extended HKB model describes changes of the relative phase over time ($\dot\phi$) as:

$$\dot\phi = \delta\omega - a\sin\phi - 2b\sin(2\phi) + \sqrt{Q}\xi t \qquad (1)$$

where φ is the relative phase between two interacting components, a and b are parameters setting the strength of attracting regions in the system's dynamical landscape, $\sqrt{Q}\xi t$ is a noise term, and $\delta\omega$ is a symmetry breaking term arising from each component having its own intrinsic behavior. The introduction of this symmetry breaking term $\delta\omega$ (equation 1) changes the entire dynamics (layout of the fixed points, bifurcation structure) of the original HKB system. It is the subtle interplay between the coupling term (k=b/a) in equation 1 and the symmetry breaking term, $\delta\omega$, that gives rise to metastability.

The flow of the coordination dynamics across a range of values of $\delta\omega$ is presented in figure 6.1 for a fixed value of the coupling parameter, k = b/a=1 where a=1 and b=1). Stable fixed points (attractors) are presented as filled circles and unstable fixed points (repellors) as open circles. Note these fixed points refer to the coordination variable or order parameter: the relative phase (see Section 7 for further discussion of the order parameter concept). A fixed point of the coordination variable φ represents a steady phase- and frequency relationship between the oscillatory components or *phase-locking*. The surface shown in figure 6.1 defines three regions under the influence of the symmetry-breaking term $\delta\omega$. In the first region present in the lower part of the surface, the system is multistable. Following the representative line labeled 1 in figure 6.1 from left to right, two stable fixed points (filled circles) are met which are the alternatives for the system to settle in. Which one, depends on the initial conditions and the size of the basin of attraction. In an intermediate region, following the line labeled 2 from left to right, one observes that the weakest attractor near anti-phase (right side) disappears after it collides with its associated repellor somewhere near $\delta\omega$=1.3, but the strongest attractor (left side) is still present as well as its repellor partner. Finally in the third region in the upper part of the surface, the regime becomes metastable. Following the line labeled 3 from left to right, no fixed points exist anymore (this part of the surface no longer intersects the isoplane $\dot{\varphi}$=0 where the fixed points are located).

What does coordination behavior look like in the metastable regime? Although all the fixed points have vanished, a key aspect is that there are still some traces of coordination, "ghosts" or "remnants" of where the fixed points once were. These create a unique dynamics alternating between two types of periods which may be called dwell time and escape time. Escape times are observed when the trajectory of the coordination variable, relative phase, drifts or diverges from the horizontal. Dwell times are observed when the trajectory converges and holds (to varying degrees) around the horizontal. In figure 6.2c we show two locations for the dwell times: one that lingers a long time before escaping (e.g. figure 6.2c, annotation 1) slightly above the more stable in-phase pattern near 0 rad (modulo 2π), and the other that lingers only briefly (e.g. figure 6.2c, annotation 2) slightly above π (modulo 2π). The dwell time is reminiscent of the transient inflexions observed near the disappeared attractor-repellor pairs in the monostable regime (figure 6.2b, annotation 3). These inflexions recur over and over again as long as the system is maintained in the metastable regime, that is, as long as it does not undergo a phase transition.

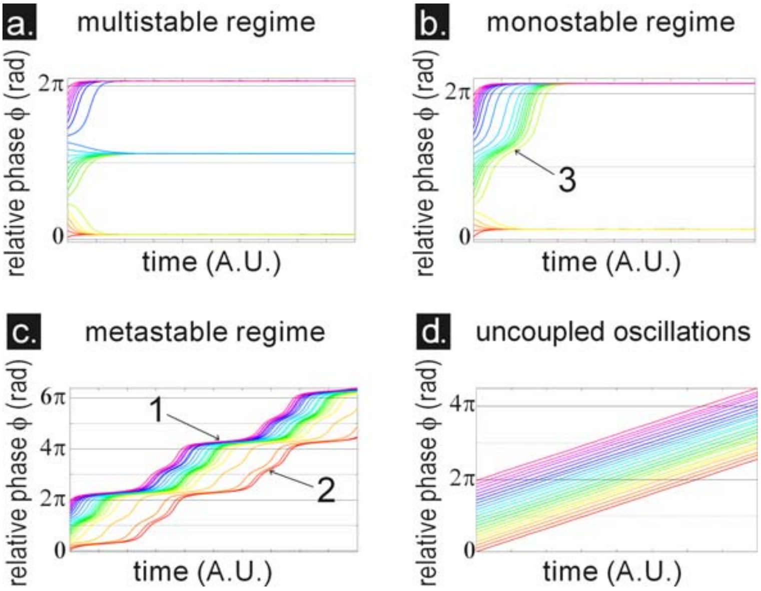

Fig. 6.2. Examples of trajectories of the coordination variable, relative phase φ arising from a range of initial conditions sampled between 0 and 2π radians, in the multistable (a), monostable (b) and metastable regimes (c) of the extended-HKB model. Trajectories in the multistable regime converge either to an attractor located slightly above 0 rad. modulo 2π or to another attractor located slightly above π rad. modulo 2π. In the monostable regime (a), trajectories converge to an attractor located slightly above 0 rad. modulo 2π. In the trajectories of relative phase for the metastable regime (c. unwrapped to convey continuity), there is no longer any persisting convergence to the fixed points, but rather a succession of periods of rapid drift (escape time) interspersed with periods inflecting toward, but not remaining on the horizontal (dwell time). Note dwells nearby 0 rad. modulo 2π in the metastable regime (e.g. dwell time at about 4π rad. annotated 1 in fig. 6.2c) and nearby π rad. modulo 2π (dwell time at about 3π rad. annotated 2 in c.) are reminiscent of the transient obtained for certain initial conditions in the monostable regime (b. annotation 3). For reference, the relative phase of uncoupled oscillators is displayed in (d.).

Despite the complete absence of phase-locked attractors, the behavior of the elements in the metastable regime is not totally independent. Rather, the dependence between the elements takes the form of dwellings (phase gathering) nearby the remnants of the fixed points and is expressed by concentrations in the histogram of the relative phase (see Kelso 1995, chap. 4). Can the brain make use of such a principle? In contrast to or as a complement of theories of large-scale organization through linear phase-coupling (Eckhorn et al. 1988; Gray et al. 1989; Bressler 1996; Varela et al. 2001), our thesis is that the ability of the system to coordinate or compute without attractors opens a large set of possibilities. The classical view of phase-locked coordination prescribes that each recruited element looses its intrinsic behavior and obeys the dictates of the assembly. When such situations arise, from the functional point of view, individual areas cease to exert an influence for the duration of the synchronized state, and the pertinent spatial level of description of the unitary activity becomes the synchronous assembly itself. However, phylogenesis promoted specialized activity of local populations of neurons (Bressler & Tognoli 2006; Ebbesson 1984; Deacon 1990; Jacobs & Jordan 1992; Chklovskii et al. 2002). In theories proposing large-scale integration through phase synchronization, the expression of local activity can only exist when the area is not enslaved into an assembly, whereas in the metastable regime, the tendency for individual activity is more continually preserved (see also Friston 1997).

As exemplified explicitly in the extended HKB model, a delicate balance between integration (coordination between individual areas) and segregation (expression of individual behavior) is achieved in the metastable regime (Kelso 1992, 1995). Excessive segregation does not allow the proper manifestation of cognition as seen for instance in autism and schizophrenia (Andreasen et al. 1999; Tononi & Edelman 2000; Brock et al. 2002; Niznikiewicz et al. 2003; Welsh et al. 2005; Liang et al. 2006). On the other hand, excessive integration does not appear to be adaptive either. Resting states measured during cognitive idling are characterized by widespread oscillations across large cortical territories (Berger 1929; Chatrian et al. 1959; Chase & Harper 1971; Kuhlman 1978; Hughes & Crunelli 2005) that appear to block or inhibit the proper expression of a local area's activity. Furthermore, propagation of synchronous activity leads to epileptic seizures (Schiff & Plum 2000; Glass 2001; Kostopoulos 2001; Blumenfeld & Taylor 2003; Dominguez et al. 2005) and is ultimately characterized by a total loss of cognition and consciousness once a certain mass of neurons is recruited. In a critical range between complete integration and complete segregation the most favorable situation for cognition is deemed to occur (Werner 2007; Atlan 1979; Chialvo 2004). Studies of interareal connectivity both at the anatomical and functional level (Friston 1997; Tononi et al. 1994; Tononi et al. 1998; see also Sporns 2004) support this view by showing that measures of complexity reach a maximum when the balance between segregative and integrative forces is achieved. Note, however, that such measures are based upon stationarity assumptions whereas metastability in coordination dynamics is a "stationary transient." That is, the holding and releasing of the relative phase over time appears to be of a transient nature, but is actually quite stationary.

Another interesting feature related to the absence of attractors is the ability of the system to exhibit more than one coordination tendency in the time course of its life. This property is reminiscent of the multistable regime with attractors, with the difference that no transition is required to switch from one state to the other. Evidence of "multistability" and spontaneous switching in perception and action abounds both at behavioral and brain levels (e.g., Kelso et al. 1992; Almonte et al. 2005; Başar-Eroglu et al. 1996; Hock et al. 1993; Keil et al. 1999; Kelso 1981, 1984; Kelso et al. 1991; Tuller et al. 1994). Aside from the multistable regime with attractors undergoing phase transition, the metastable regime is also suitable to explain those experimental results. The tendencies of the metastable regime toward the remnants of the fixed points readily implement spontaneous reversals of percepts and behaviors described in these studies (Kelso et al. 1995). From the perspective of coordination dynamics, the time the system dwells in each remnant depends on a subtle blend of the asymmetry of the components (longer dwelling for smaller asymmetry) and the strength of the coupling (longer dwelling for larger values of a or b). Such a mechanism provides a powerful means to instantiate alternating thoughts/percepts and their probability in both biological systems and their artificial models (e.g. alternating percepts of vase or faces in ambiguous Rubin figures, or alternative choices in the solving of a chess game).

Both a multistable regime with attractors and a metastable regime with attracting tendencies allow so-called perceptual and behavioral "multistability." Which attractor is reached in the multistable regime primarily depends on initial conditions. Once the system is settled into an attractor, a certain amount of noise or a perturbation is required to achieve a switching to another attractor. Or, if control parameters such as attention are modified, a bifurcation or phase transition may occur, meaning an attractor looses stability as the regime changes from multistable to monostable or vice versa (see Ditzinger & Haken 1989, 1990 for excellent examples of such modeling). In the metastable regime, successive visits to the remnants of the fixed points are intrinsic to the time course of the system, and do not require any external source of input. This is an important difference between multistability and metastability, and likely translates into an advantage in speed which is known to be an important constraint in neurocognitive systems (Thorpe et al. 1996) and a crucial aspect of their performance (Jensen 1993).

6 Metastability in the Brain

What is the anatomical basis in the brain for metastable coordination dynamics? As noted earlier, the fundamental requirements for metastability are the existence of coupled components each exhibiting spontaneous oscillatory behavior and the presence of broken symmetry. There are several spatial scales at which the collective behavior of the brain spontaneously expresses periodic oscillations (Chen et al. 2003a, 2003b; Freeman 2000; Wright et al. 2001; Buzsáki & Draguhn 2004) and represents the combined contribution of groups of neurons, the joint action of which is adequate to establish transfer of information (Braitenberg & Schuz 1991; Douglas & Martin 1991; Buxhoeveden & Casanova 2002). The oscillatory activity of the brain may be captured directly via invasive neurophysiological methods such as LFP and iEEG, or indirectly from EEG scalp recordings (commonly at the price of estimating local oscillations by bandpass filtering of the signals). The coupling between local groups of neurons is supported by long-range functional connectivity (Varela et al. 2001; Bressler 1995; Sporns & Kötter 2004). Broken symmetry has several grounds to spring from, including the incommensurate characteristic frequencies of local populations of neurons (Freeman 2001) and their heterogeneous connectivity (Jirsa & Kelso 2000).

If the conditions required for metastable coordination in the brain are easily met, it remains to establish that the brain actually shows instances of operating in this regime. This empirical characterization encounters some difficulties. Before any attempt to find signatures of metastability, a first question is to identify from the continuous stream of brain activity some segments corresponding to individual regimes. In other words, it consists in finding the transitions between regimes, a task undertaken by only a few (Lehmann et al. 1998; Kaplan & Shishkin 1999). Provided adequate measurement/estimation of local oscillations in the presence of noise and spatial smearing is possible, insights can be gained by identifying episodes of phase-locking (these forming states) and ascribing their interim periods as transitions (e.g. Varela et al. 2001). In the absence of ad hoc segmentation of the EEG, it remains as a possibility to use behavioral indices as cues to when brain transitions occur (e.g. Kelso et al. 1992; Keil et al. 1999). The case of metastability is evidently more difficult since the regime is stationary but not stable. Initial attempts have targeted the more recognizable dwell time as a quasi phase-locked event (Fingelkurts & Fingelkurts 2004). To gain understanding on the mechanism, it seems necessary to elaborate strategies that comprise the coordination pattern of the metastable regime in its entirety (dwell and escape time as an inseparable whole) and to establish criteria for the differentiation of state transitions and dwell ~ escape regimes.

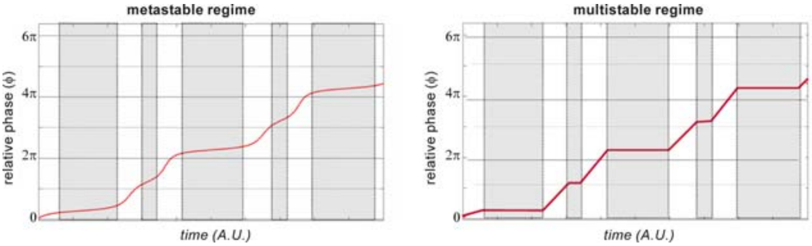

Fig. 6.3. Comparison of relative phase trajectories in the metastable and multistable regime for a temporal window of arbitrary size. Coordination (multistable regime) and tendency to coordinate (metastable regime) are shown in grey boxes. In the multistable regime (right), a succession of states (stable relative phase near 0 and pi radians) is interweaved with transitions. Horizontal segments are lost in the metastable regime (left) which only shows tendencies for synchronization toward in-phase and anti-phase. In a situation in which coordination is estimated from a broadband signal in the presence of noise, distinguishing between the two regimes may difficult. The transitions on the right however are induced by parametric change; the flow on the left is not.

To identify and understand potential mechanisms, it is of prime importance to be able to distinguish between the different regimes. For instance, a transition between the metastable and the monostable regime could be a way the brain instantiates a process of decision among a set of possibilities. This amounts to the creation of information (Kelso 2002). Figure 6.3 shows the isomorphism of simulated systems belonging to both regimes in their relative phases' trajectory. In this window of arbitrary size, a succession of states is shown in the multistable regime (right) separated by transitions. It differs from the metastable regime (left) by the presence of horizontal segments (stable relative phase) during the states and by sharp inflections of the relative phase at the onset and offset of transitions. The corresponding histograms of the relative phase cumulated over this period of time are similar as well. The ability to distinguish the multistable regime from the metastable regime in a nonsegmented EEG depends critically on the precision of the estimation of the components' frequency and phase. Unfavorable circumstances are met since the EEG is a noisy, broadband signal (Pritchard 1992), and because each component's frequency shifts when coupled in a metastable regime.

Other criteria might be sought for to distinguish between those regimes. State-transition regimes have been conceptually formulated and empirically verified by a line of studies initiated by Eckhorn et al. (1988), Gray et al. (1989). The theory of "transient cell assemblies" has gathered numerous empirical findings at the microscale (Engel et al. 1991; Castelo-Branco et al. 2000), mesoscale (Bressler 1995; Eckhorn & Obermüller 1993; Bressler et al. 1993), and macroscale (Başar-Eroglu et al. 1996; Tallon-Baudry et al. 2001; Müller et al. 1996, Rodriguez et al. 1999). This set of studies relies on linear pairwise phase synchronization methods applied both to Single- and Multi-Unit Activity and Field Potentials. Whereas many studies have focused on the "state" part of the state transition, an interesting feature is seen in the study by Rodriguez et al. (1999) of coherent oscillations elicited by Mooney faces. Two periods of intense synchronization at 250 msec and 700 msec are separated by a period of intense desynchronization that the authors described as phase scattering. They suggest that phase scattering is a mechanism by which the brain realizes the transition from a coherent assembly to another assembly – both belonging to stable regimes. Such a mechanism is unnecessary in the succession of tendencies that are characteristic of metastable coordination dynamics.

In summary, the brain by virtue of its properties forms a suitable ground for metastability to take place. The characterization of metastable onsets, however, is a matter which will certainly require some methodological developments outside the linear approach of transient phase synchronization. In the meantime, indices of metastability are found in the distribution of dwell times near phase-locked states.

7 Clarifying Nonlinear Brain Dynamics: The Freeman-Kelso Dialogue[1]

Recently, the eminent neurophysiologist Walter Freeman published an article entitled "Metastability, instability and state transitions in neocortex" (Freeman & Holmes 2005) that led to a discussion with the authors which we think may be useful for clarificatory purposes and to expand awareness of nonlinear brain dynamics. Here we highlight some of the main issues – FAQ about metastable neurodynamics, if you like – in part as a tribute to Freeman and his pioneering work and its relation to coordination dynamics.

First the concept itself. Freeman draws heavily from the solid state physics literature, where he notes that the concept of metastability has been in use for over

[1] With the blessing of Walter Freeman.

thirty years. Although this is correct and many useful analogies have been made between brains and other kinds of cooperative phenomena in nature (e.g., Kelso 1995; Haken 1983, 1996; Haken et al. 1985; Kelso & Haken 1995) notice here that metastability arises because of a specific symmetry-breaking in the coordination dynamics. That is, intrinsic differences in oscillatory frequency between the components are sufficiently large that they do their own thing, while still retaining a tendency to coordinate together. Thus, the relative phase between the components drifts over time, but is occasionally trapped near remnants of the (idealized) coordinated states, for example, near 0 and π radians (cf. fig. 6.2). As a consequence of broken symmetry in its coordination dynamics, both brain and behavior are able to exhibit a far more variable, plastic, and fluid form of coordination in which tendencies for the components to function in an independent, segregated fashion coexist with tendencies for the system to behave in an integrated, coordinative fashion.

Second, Freeman inquires about the order parameter in coordination dynamics. Freeman himself pursues spatial patterns of amplitude which he understands as manifestations of new attractors that form through learning. It is these amplitude patterns of aperiodic carrier waves derived from high density EEG recordings that constitute his order parameter. Freeman regards these as evidence for cortical dynamics accessing nonconvergent attractors for brief time periods by state transitions that recur at rates in the theta range. Although we originally also used the physicist’s terminology of order parameters (e.g. Kelso et al. 1992; Haken et al. 1985) we now prefer the term “collective variable” or “coordination variable” as a quantity that characterizes the cooperative behavior of a system with very many microscopic degrees of freedom. Indeed, our approach is called coordination dynamics because it deals fundamentally with informationally meaningful quantities (Kelso 1994). Coordination in living things is not (or not only) matter in motion. The spatiotemporal ordering observed is functional and task-specific. Because in our studies the variable that changes qualitatively under parametric change is the relative phase, relative phase is one of the key order parameters of biological coordination. Relative phase is the outcome of a nonlinear interaction among nonlinear oscillatory components, yet in turn reciprocally conditions or “orders” the behavior of those components. In a system in which potentially many things can be measured, the variable that changes qualitatively is the one that captures the spatiotemporal ordering among the components. This strategy of identifying order parameters or coordination variables in experimental data goes back to Kelso (1981, 1984) and the early theoretical modeling by Haken, Kelso, and Bunz (1985). Recent empirical and theoretical research contacts Freeman’s work in that it shows that phase transitions can also arise through the amplitudes of oscillation (Assisi et al. 2005). Both routes are possible depending on the situation, for example, amplitude drops across the transition, the relative phase changes abruptly. Again, this is all under parametric control. In coordination dynamics, you do not “know” you have a coordination variable or order parameter and control parameters unless the former change qualitatively at transitions, and the latter – when systematically varied – lead the system through transitions. Order parameters and control parameters in the framework of coordination dynamics are thus co-implicative and complementary. An issue in the olfactory system concerns what the control parameters are, for instance, that might lead the system from a steady state to a limit cycle or to chaotic oscillations and itineracy. Both Freeman's and our approach appeal to Haken's (1983) synergetics and so-called circular or reciprocal causality: whether one adopts the term "order parameters" or "coordination variables," both arise from the cooperation of millions of neurons and in turn are influenced by the very field of activity they create (see also Jirsa et al. 1998; Jirsa & Haken 1997) for an elaboration and application of neural field theory).

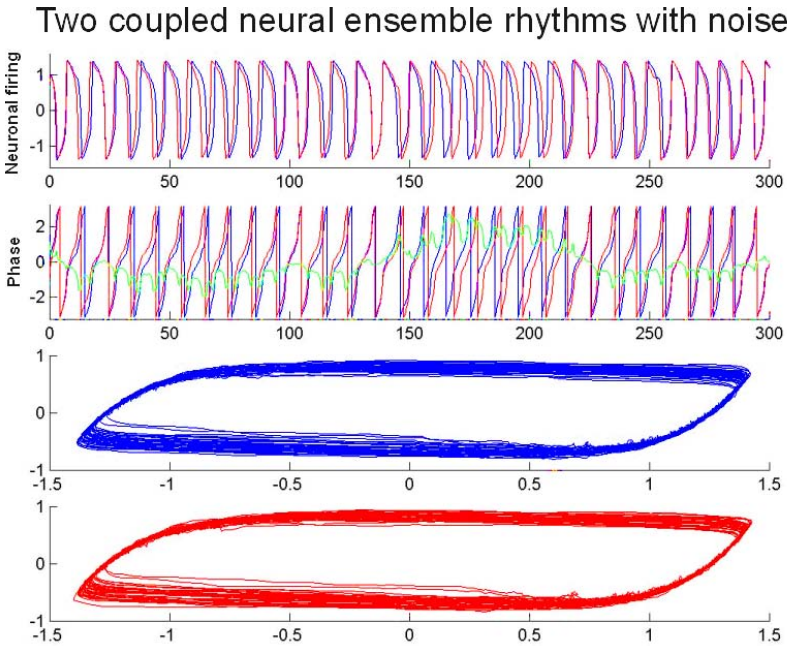

Fig. 6.4. A simulation of two coupled neural ensembles composed of an array of Fitzhugh-Nagumo oscillators (courtesy Viktor Jirsa). See text for description.

$$\frac{dX}{dt} = F(X) = \frac{\partial}{\partial t} S(X - Y) + noise$$
$$\frac{dY}{dt} = F(Y) = \frac{\partial}{\partial t} S(Y - X) + noise \qquad (2)$$

Both Freeman's approach and coordination dynamics appeal to nonlinear coupling among neural oscillators as the basis for varying degrees of global integration. Both make significant use of basin attractor dynamics to interpret experimental data. In coordination dynamics, the latter constitutes a step of idealization that is necessary in order to understand what it means to break the symmetry of the coordination dynamics (Kelso et al. 1990; Kelso 2002; Kelso & Jeka 1992). Both approaches nevertheless invoke symmetry breaking, coordination dynamics from the loss of attractors of the relative phase dynamics, and Freeman in the emergence of spatial patterns of amplitude and phase from EEG recordings by convergence to a selected *a posteriori* attractor in a landscape of *a priori* attractors. There are obvious parallels between the two bodies of work; both are testament to the richness of detail and power of nonlinear theory. Freeman envisages transient but definite access to a succession of basins of attraction. Metastable coordination dynamics, on the other hand, has a very precise meaning: it is not about states but a subtle blend of integrative and segregative *tendencies*. Notably, these integrative tendencies to phase gather and segregative tendencies to phase wrap can be shown at the level of coupled neural ensembles. Figure 6.4 illustrates a set of coupled neural ensembles each composed of one hundred Fitzhugh-Nagumo oscillators connected by a sigmoidal function, which is the usual consequence of the summation of threshold properties at cell bodies (equation 2). A small amount of random noise has been added only for illustrative effect. Looking from top to bottom, the neuronal firing activity of each ensemble (X,Y) is shown, followed by the individual oscillatory phases, their relative phase and respective phase plane trajectories indicating limit cycle properties, along with a simple mathematical representation.

The intent here is only to establish proof of concept. It is quite clear that the relative phase between the neural groups dwells near phi = 0, wanders and then returns, indicating nicely the transient metastable tendencies to integrate and segregate. As Fingelkurts and Fingelkurts note:

> One may note that the metastabilty principle extends the Haken synergetics rules.… Metastabilty extends them to situations where there are neither stable nor unstable states, only coexisting tendencies (see Kelso, 2002). (Fingelkurts & Fingelkurts 2004)

8 A Short Afterthought

It has not escaped our attention that the metastable coordination dynamics of brain and behavior invites extension to the processes of thought and thinking. A case can be made that multistable perception and ambiguity offer test fields for understanding the self-organization of the brain. The perceptual system, when faced with a constant stimulus, may fluctuate between two or more interpretations. Through ambiguities, as Richard Gregory (2000) remarks, we can investigate how the brain makes up its mind. One may speculate that when a naïve perceiver views a figure such as the hidden Dalmatian (fig. 6.5), a series of mental phases and their associated brain states takes place. In a first stage, the naïve observer may attempt to group the blackened areas in various ways. There will be multistability and phase transitions. Eventually, he/she will correctly organize the picture and a monostable state will be reached in which the Dalmatian's picture is salient. Finally, the observer may think of the artist's work and consider simultaneously the fragmentation that allows the Dalmatian to disintegrate into the scene as well as the organization that hints its presence. A metastable dynamic will arise in which both the figure and its hiding texture will simultaneously be present in the mind. As Heisenberg noted fifty years ago:

> We realize that the situation of complementarity is not confined to the atomic world alone; we meet it when we reflect about a decision and the motives for our decision or when we have a choice between enjoying music and analyzing its structure. (Heisenberg 1958, p. 179)

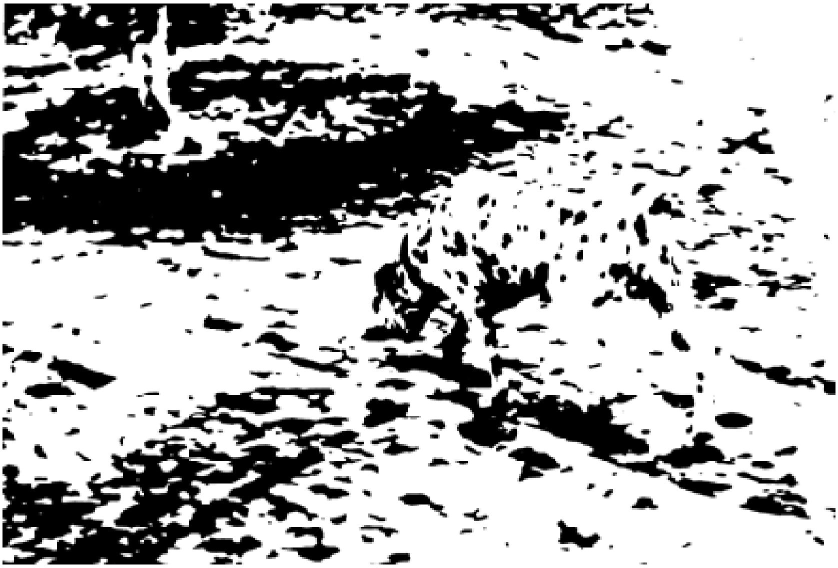

Fig. 6.5. The Hidden Dalmatian
(Richard Gregory, "The Intelligent Eye," McGraw Hill, 1979. Photographer: Ron James.)

Of course, Heisenberg, Bohr, and Pauli's philosophy came from considerations in quantum mechanics. Here both the philosophy of the complementary nature and a complementary neuroscience are rooted in the metastable coordination dynamics of the brain.

Acknowledgments

Previously published in R. Kozma and L. Perlovsky (Eds.). *Neurodynamics of higher-level cognition and consciousness.* Heidelberg: Springer, 2007.

This work was supported by NIMH Grant MH42900, NINDS Grant NS48229 and ONR Grant N00014-05-1-0117.

References

Almonte, F., Jirsa, V.K., Large, E.W., Tuller, B.: Integration and segregation in auditory streaming. Physica D 212, 137–159 (2005)

Andreasen, N.C., Nopoulos, P., O'Leary, D.S., Miller, D.D., Wassink, T., Flaum, L.: Defining the phenotype of schizophrenia: Cognitive dysmetria and its neural mechanisms. Biological Psychiatry 46, 908–920 (1999)

Assisi, C.G., Jirsa, V.K., Kelso, J.A.S.: Synchrony and clustering in heterogeneous networks with global coupling and parameter dispersion. Physical Review Letters 94, 18106 (2005)

Atlan, H.: Entre le cristal et la fumée. Paris, Seuil (1979)

Başar, E.: Memory and brain dynamics: Oscillations integrating attention, perception, learning, and memory. Conceptual advances in brain research, vol. 7. CRC Press, Boca Raton (2004)

Başar-Eroglu, C., Struber, D., Kruse, P., Başar, E., Stadler, M.: Frontal Gamma-band enhancement during multistable visual perception. International Journal of Psychophysiology 24, 113–125 (1996)

Berger, H.: Über das Elektroenkephalogramm des Menschen. Archiv für Psychiatrie und Nervenkrankheiten 87, 527–570 (1929)

Blumenfeld, H., Taylor, J.: Why do seizures cause loss of consciousness? The Neuroscientist 9(5), 301–310 (2003)

Braitenberg, V., Schuz, A.: Anatomy of the cortex. Springer, Berlin (1991)

Bressler, S.L.: Large-scale cortical networks and cognition. Brain Research Reviews 20, 288–304 (1995)

Bressler, S.L.: Interareal synchronization in the visual cortex. Behavioral and Brain Research 76, 37–49 (1996)

Bressler, S.L., Kelso, J.A.S.: Cortical coordination dynamics and cognition. Trends in Cognitive Sciences 5, 26–36 (2001)

Bressler, S.L., Tognoli, E.: Operational principles of neurocognitive networks. International Journal of Psychophysiology 60, 139–148 (2006)

Bressler, S.L., Coppola, R., Nakamura, R.: Episodic multiregional cortical coherence at multiple frequencies during visual task performance. Nature 366, 153–156 (1993)

Brock, J., Brown, C.C., Boucher, J., Rippon, G.: The temporal binding deficit hypothesis of autism. Development and Psychopathology 14, 209–224 (2002)

Brown, P., Marsden, J.F.: Cortical network resonance and motor activity in humans. Neuroscientist 7, 518–527 (2001)

Buxhoeveden, D.P., Casanova, M.F.: The minicolumnar hypothesis in neurosciences. Brain 125(5), 935–951 (2002)

Buzsáki, G.: Rhythms of the brain. Oxford University Press, Oxford (2006)

Buzsáki, G., Draguhn, A.: Neuronal oscillations in cortical networks. Science 304(5679), 1926–1929 (2004)

Castelo-Branco, M., Goebel, R., Neuenschwander, S.: Neural synchrony correlates with surface segregation rules. Nature 405, 685–689 (2000)

Chase, M.H., Harper, R.M.: Somatomotor and visceromotor correlates of operantly conditioned 12–14 cycles per second sensorimotor cortical activity. Electroencephalography and Clinical Neurophysiology 31, 85–92 (1971)

Chatrian, G.E., Petersen, M.C., Lazarte, J.A.: The blocking of the central wicket rhythm and some central changes related to movement. Electroencephalography and Clinical Neurophysiology 11, 497–510 (1959)

Chen, Y., Ding, M., Kelso, J.A.S.: Long range dependence in human sensorimotor coordination. In: Rangarajan, G., Ding, M. (eds.) Processes with long-range correlations, pp. 309–323. Springer, Berlin (2003a)

Chen, Y., Ding, M., Kelso, J.A.S.: Task-related power and coherence changes in neuromagnetic activity during visuomotor coordination. Experimental Brain Research 148, 105–116 (2003b)

Chialvo, D.R.: Critical brain networks. Physica A 340(4), 756–765 (2004)

Chklovskii, D.B., Schikorski, T., Stevens, C.F.: Wiring optimization in cortical circuits. Neuron 34(3), 341–347 (2002)

Deacon, T.W.: Rethinking mammalian brain evolution. American Zoologist 30, 629–705 (1990)

Ditzinger, T., Haken, H.: Oscillations in the perception of ambiguous patterns. Biological Cybernetics 61, 279–287 (1989)

Ditzinger, T., Haken, H.: The impact of fluctuations on the recognition of ambiguous patterns. Biological Cybernetics 63, 453–456 (1990)

Dominguez, L.G., Wennberg, R., Gaetz, W., Cheyne, D., Snead III, O.C., Perez Velazquez, J.L.: Enhanced synchrony in epileptiform activity? Local versus distant synchronization in generalized seizures. Journal of Neuroscience 25(35), 8077–8084 (2005)

Douglas, R.J., Martin, K.A.: A functional microcircuit for cat visual cortex. Journal of Physiology 440, 735–769 (1991)

Ebbesson, S.O.E.: Evolution and ontogeny of neural circuits. Behavioral and Brain Science 7, 321–366 (1984)

Eckhorn, R., Obermüller, A.: Single neurons are differently involved in stimulus-specific oscillations in cat visual cortex. Experimental Brain Research 95(1), 177–182 (1993)

Eckhorn, R., Bauer, R., Jordan, W., Borsch, M., Kruse, W., Munk, M., Reitboeck, H.J.: Coherent oscillations: A mechanism of feature linking in the visual cortex? Multiple electrode correlation analyses in the cat. Biological Cybernetics 60(2), 121–130 (1988)

Edelman, G.M.: Naturalizing consciousness: A theoretical framework. Proceedings of the National Academy of Science USA 100(9), 520–524 (2004)

Edelman, G.: Second nature: Brain science and human knowledge. Yale University Press, New Haven (2006)

Edelman, G., Tononi, G.: A universe of consciousness. Basic Books, New York (2000)

Engel, A.K., König, P., Singer, W.: Direct physiological evidence for scene segmentation by temporal coding. Proceedings of the National Academy of Science USA 88, 9136–9140 (1991)

Fingelkurts, A.A., Fingelkurts, A.A.: Making complexity simpler: Multivariability and metastability in the brain. International Journal of Neuroscience 114(7), 843–862 (2004)

Finger, S.: Origins of neuroscience. Oxford, New York (1994)

Frayn, M.: The human touch: Our part in the creation of the universe. Faber & Faber, London (2006)

Freeman, W.J.: Neurodynamics: An exploration in mesoscopic brain dynamics. Springer, Berlin (2000)

Freeman, W.J.: Making sense of brain waves: The most baffling frontier in neuroscience. In: Parelus, P., Principe, J., Rajasekaran, S. (eds.) Biocomputing, pp. 33–55. Kluwer, New York (2001)

Freeman, W.J., Holmes, M.D.: Metastability, instability, and state transition in neocortex. Neural Networks 18(5–6), 497–504 (2005)

Fries, P., Reynolds, J.H., Rorie, A.E., Desimone, R.: Modulation of oscillatory neuronal synchronization by selective visual attention. Science 291, 1560–1563 (2001)

Friston, K.J.: Transients, metastability, and neuronal dynamics. NeuroImage 5, 164–171 (1997)

Glass, L.: Synchronization and rhythmic processes in physiology. Nature 410, 277–284 (2001)

Gray, C.M., König, P., Engel, A.K., Singer, W.: Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. Nature 338(6213), 334–337 (1989)

Gregory, R.L.: Editorial. Perception 29, 1139–1142 (2000)

Grossberg, S.: The complementary brain: A unifying view of brain specialization and modularity. Trends in Cognitive Sciences 4, 233–246 (2000)

Haken, H.: Synergetics: An introduction. series in synergetics. Springer, Berlin (1983)

Haken, H.: Principles of brain functioning. Springer, Berlin (1996)

Haken, H., Kelso, J.A.S., Bunz, H.: A theoretical model of phase transitions in human hand movements. Biological Cybernetics 51, 347–356 (1985)

Heisenberg, W.: Physics and philosophy: The revolution in modern physics. Harper & Row, New York (1958)

Hock, H.S., Kelso, J.A.S., Schöner, G.: Bistability and hysteresis in the organization of apparent motion patterns. Journal of Experimental Psychology: Human Perception Performance 19, 63–80 (1993)

Hughes, S.W., Crunelli, V.: Thalamic mechanisms of EEG Alpha rhythms and their pathological implications. The Neuroscientist 11, 357–372 (2005)

Jacobs, R.A., Jordan, M.I.: Computational consequences of a bias towards short connections. Journal of Cognitive Neuroscience 4, 323–336 (1992)

Jensen, A.R.: Why is reaction time correlated with psychometric g? Current Directions in Psychological Science 2, 53–56 (1993)

Ji, S.: Complementarism: A biology-based philosophical framework to integrate western science and eastern tao. In: Proceedings of the 16th International Congress of Psychotherapy, pp. 518–548 (1995)

Jirsa, V.K., Haken, H.: A derivation of a macroscopic field theory of the brain from the quasi-microscopic neural dynamics. Physica D 99, 503–526 (1997)

Jirsa, V.K., Kelso, J.A.S.: Spatiotemporal pattern formation in continuous systems with heterogeneous connection topologies. Physical Review E 62(6), 8462–8465 (2000)

Jirsa, V.K., Kelso, J.A.S. (eds.): Coordination dynamics: Issues and trends. Springer, Berlin (2004)

Jirsa, V.K., Fuchs, A., Kelso, J.A.S.: Connecting cortical and behavioral dynamics: Bimanual coordination. Neural Computation 10, 2019–2045 (1998)

Kaplan, A.Y., Shishkin, S.L.: Nonparametric method for the segmentation of the EEG. Computer Methods and Programs in Biomedicine 60(2), 93–106 (1999)

Keil, A., Muller, M.M., Ray, W.J., Gruber, T., Elbert, T.: Human Gamma band activity and perception of a gestalt. Journal of Neuroscience 19, 7152–7161 (1999)

Kelso, J.A.S.: On the oscillatory basis of movement. Bulletin of the Psychonomic Society 18, 63 (1981)

Kelso, J.A.S.: Phase transitions and critical behavior in human bimanual coordination. American Journal of Physiology 246(6 Pt. 2), R1000–R1004 (1984)

Kelso, J.A.S.: Behavioral and neural pattern generation: The concept of neurobehavioral dynamical system (NBDS). In: Koepchen, H.P., Huopaniemi, T. (eds.) Cardiorespiratory and motor coordination, pp. 224–234. Springer, Berlin (1991)

Kelso, J.A.S.: Coordination dynamics of human brain and behavior. Springer Proceedings in Physics 69, 223–234 (1992)

Kelso, J.A.S.: The informational character of self-organized coordination dynamics. Human Movement Science 13, 393–413 (1994)

Kelso, J.A.S.: Dynamic patterns: The self-organization of brain and behavior. MIT Press, Cambridge (1995)

Kelso, J.A.S.: The complementary nature of coordination dynamics: Self-organiza¬tion and the origins of agency. Journal of Nonlinear Phenomena in Complex Systems 5, 364–371 (2002)

Kelso, J.A.S., Engstrøm, D.: The complementary nature. MIT Press, Cambridge (2006)

Kelso, J.A.S., Haken, H.: New laws to be expected in the organism: Synergetics of brain and behavior. In: Murphy, M., O'Neill, L. (eds.) What Is life? The next 50 years. Cambridge University Press, Cambridge (1995)

Kelso, J.A.S., Jeka, J.J.: Symmetry breaking dynamics of human multilimb co-ordination. Journal of Experimental Psychology: Human Perception and Performance 18, 645–668 (1992)

Kelso, J.A.S., DelColle, J., Schöner, G.: Action-perception as a pattern formation process. In: Jeannerod, M. (ed.) Attention and performance XIII, pp. 139–169. Erlbaum, Hillsdale (1990)

Kelso, J.A.S., DeGuzman, G.C., Holroyd, T.: Synergetic dynamics of biological coordination with special reference to phase attraction and intermittency. In: Haken, H., Koepchen, H.P. (eds.) Rhythms in physiological systems. Springer series in synergetics, vol. 55, pp. 195–213. Springer, Heidelberg (1991)

Kelso, J.A.S., Bressler, S.L., Buchanan, S., DeGuzman, G.C., Ding, M., Fuchs, A., Holroyd, T.: A phase transition in human brain and behavior. Physics Letters A 169, 134–144 (1992)

Kelso, J.A.S., Case, P., Holroyd, T., Horvath, E., Raczaszek, J., Tuller, B., Ding, M.: Multistability and metastability in perceptual and brain dynamics. In: Kruse, P., Stadler, M. (eds.) Ambiguity in mind and nature, pp. 159–185. Springer, Heidelberg (1995)

Koch, C.: Thinking about the conscious mind: A Review of Mind – An Introduction, by John Searle. Science 306, 979–980 (2004)

König, P., Engel, A.K., Roelfsema, P.R., Singer, W.: Coincidence detection or temporal integration: The role of the cortical neuron revisited. Trends in Neurosciences 19, 130–137 (1996)

Kostopoulos, G.K.: Involvement of the thalamocortical system in epileptic loss of consciousness. Epilepsia 42(30), 13–19 (2001)

Kuhlman, W.N.: Functional topography of the human Mu rhythm. Electroencephalography and Clinical Neurophysiology 44, 83–93 (1978)

Lehmann, D., Strik, W.K., Henggeler, B., Koenig, T., Koukkou, M.: Brain electric microstates and momentary conscious mind states as building blocks of spontaneous thinking: I. Visual imagery and abstract thoughts. International Journal of Psychophysiology 29(1), 1–11 (1998)

Liang, M., Zhou, Y., Jiang, T., Liu, Z., Tian, L., Liu, H., Hao, Y.: Widespread functional disconnectivity in schizophrenia with resting-state functional magnetic resonance imaging. Neuroreport 17(2), 209–213 (2006)

Mima, T., Matsuoka, T., Hallett, M.: Functional coupling of human right and left cortical motor areas demonstrated with partial coherence analysis. Neuroscience Letters 287, 93–96 (2000)

Müller, M.M., Bosch, J., Elbert, T., Kreiter, A., Sosa, M., Valdes-Sosa, P., Rockstroh, B.: Visually induced gamma-band responses in human electroencephalographic activity: A link to animal studies. Experimental Brain Research 112(1), 96–102 (1996)

Munk, M.H.J., Roelfsema, P.R., König, P., Engel, A.K., Singer, W.: Role of reticular activation in the modulation of intracortical synchronization. Science 272, 271–274 (1996)

Niznikiewicz, M.A., Kubicki, M., Shenton, M.E.: Recent structural and functional imaging findings in schizophrenia. Current Opinion in Psychiatry 16, 123–147 (2003)

Pauli, W.: Writings on physics and philosophy. In: Enz, C.P., von Meyenn, K. (eds.). Springer, Berlin (1994)

Perez Velazquez, J.L.: Brain, behaviour and mathematics: Are we using the right approaches? Physica D: Nonlinear Phenomena 212(3-4), 161–182 (2005)

Perez Velazquez, J.L., Wennberg, R.: Metastability of brain states and the many routes to seizures: Numerous causes, same result. In: Pandalai, S.G. (ed.) Recent research developments in biophysics, vol. 3, pp. 25–59. Transworld Research Network, Kerala (2004)

Pritchard, W.S.: The brain in fractal time: 1/f-like power spectrum scaling of the human electroencephalogram. International Journal of Neuroscience 66(1–2), 119–129 (1992)

Rodriguez, E., George, N., Lachaux, J.P., Martinerie, J., Renault, B., Varela, F.: Perception's shadow: Long-distance synchronization in the human brain activity. Nature 397, 430–433 (1999)

Schiff, N.D., Plum, F.: The role of arousal and "gating" systems in the neurology of impaired consciousness. Journal of Clinical Neurophysiology 17, 438–452 (2000)

Schöner, G., Kelso, J.A.S.: Dynamic pattern generation in behavioral and neural systems. Science 239, 1513–1520 (1988)

Sporns, O.: Complex neural dynamics. In: Jirsa, V.K., Kelso, J.A.S. (eds.) Coordination dynamics: Issues and trends, pp. 197–215. Springer, Berlin (2004)

Sporns, O., Kötter, R.: Motifs in brain networks. PLoS Biology 2, 1910–1918 (2004)

Steinmetz, P.N., Roy, A., Fitzgerald, P., Hsiao, S.S., Niebur, E., Johnson, K.O.: Attention modulates synchronized neuronal firing in primate somatosensory cortex. Nature 404, 187–190 (2000)

Talbot, M.: The baby lab: How Elizabeth Spelke peers into the infant mind, New York, pp. 90–101, September 4 (2006)

Tallon-Baudry, C., Bertrand, O., Fischer, C.: Oscillatory synchrony between human extrastriate areas during visual short-term memory maintenance. Journal of Neuroscience 21, RC177, 1–5 (2001)

Thorpe, S., Fize, D., Marlot, C.: Speed of processing in the human visual system. Nature 381, 520–522 (1996)

Tononi, G., Edelman, G.M.: Schizophrenia and the mechanisms of conscious integration. Brain Research Reviews 31, 391–400 (2000)

Tononi, G., Sporns, O., Edelman, G.M.: A measure for brain complexity: Relating functional segregation and integration in the nervous system. Proceedings of the National Academy of Science USA 91, 5033–5037 (1994)

Tononi, G., Sporns, O., Edelman, G.M.: Complexity and coherency: Integrating information in the brain. Trends in Cognitive Sciences 2, 474–484 (1998)

Tschacher, W., Dauwalder, J.P.: The dynamical systems approach to cognition: Concepts and empirical paradigms based on self-organization, embodiment and coordination dynamics. World Scientific, Singapore (2003)

Tuller, B., Case, P., Ding, M., Kelso, J.A.S.: The nonlinear dynamics of speech categorization. Journal of Experimental Psychology: Human Perception and Performance 20, 1–14 (1994)

Varela, F.J., Lachaux, J.-P., Rodriguez, E., Martinerie, J.: The brainweb: Phase synchronization and large-scale integration. Nature Reviews Neuroscience 2, 229–239 (2001)

Welsh, J.P., Ahn, E.S., Placantonakis, D.G.: Is autism due to brain desynchronization? International Journal of Developmental Neuroscience 23, 253–263 (2005)

Werner, G.: Metastability, criticality and phase transitions in brain and its models. Biosystems 90(2), 496–508 (2007)

Wright, J.J., Robinson, P.A., Rennie, C.J., Gordon, E., Bourke, P.D., Chapman, C.L., Hawthorn, N., Lees, G.J., Alexander, D.: Toward an integrated continuum model of cerebral dynamics: The cerebral rhythms, synchronous oscillation and cortical stability. Journal of Biological and Information Processing Systems 63(1-3), 71–88 (2001)

Part II: Volition and Consciousness: Are They Illusions?

7

Physiology of Volition

Mark Hallett, MD

Chief, Human Motor Control Section, NINDS
NIH, Building 10, Room 7D37
10 Center Dr MSC 1428
Bethesda, MD 20892-1428
hallettm@ninds.nih.gov

Summary. The idea of free will is a conscious awareness of the brain concerning the nature of the movement that it produces. There is no evidence for it to be a driving force in movement generation. This review considers the physiology of movement generation and how the concepts of willing and agency might arise. Both the anatomical substrates and the timing of events are considered. Movement initiation and volition are not necessarily linked, and one line of evidence comes from consideration of patients with disorders of volition. Movement is generated subconsciously, and the conscious sense of willing the movement comes later, but the exact time of this event is difficult to assess because of the potentially illusory nature of introspection. The evidence suggests that movement is initiated in frontal lobe, particularly the mesial areas, and the sense of volition arises as the result of a corollary discharge from premotor and motor areas likely involving the parietal lobe. Agency probably involves a similar region in the parietal lobe and requires both the sense of volition and movement feedback.

Keywords: free will, volition, agency, corollary discharge, frontal lobe, parietal lobe, premotor cortex, motor cortex.

1 Definition of Terms

What is volition? The common view is that it is the human process of choosing which movement to make and when to make it; it is often referred to as “free will.” I will review what is known about how the brain makes movement, and it is not clear that such a process has been identified. Indeed, to some extent it is not even clear how to recognize it. However, there is another aspect of volition which is certain, and that is that humans have the perception that their movements are freely chosen. This perception is an element of consciousness, a so-called quale, and even though consciousness itself is not understood, we are able to investigate the perception of volition. Considerable work is being done in this area. There are two aspects of volition. One is the sense of willing a movement to occur. The second is agency, the sense that the person caused the movement that just occurred. In this situation, the person is the agent. Willing can occur without a movement happening, but to have a sense of agency there needs to be both willing and the movement.

2 Disorders of Volition

While the perception of volition is common, it is not universal. There are many neurological disorders in which movements occur without volition or with a distorted sense of volition. These disorders are of great interest in themselves, but also indicate that movement genesis is not obligatorily linked to the perception of volition. Such movements could theoretically arise in two different ways. The process of movement genesis could be aberrant. Alternatively, the process of movement genesis might involve normal mechanisms, but the linkage to the perception is faulty. Thus, if there is an aspect of volition that does indeed choose movement, it can be separated from the aspect of volition which is perceived. A brief review of some of these disorders of volition will illustrate this point.

There is the problem of involuntary movements. Patients with Huntington disease have chorea, but often do not recognize their involuntary movements early in the course of their illness. When asked about a movement, patients will say that it was voluntary. Patients with tics often cannot say whether their movements are voluntary or involuntary. They find it easier to say that they can suppress their movements or that they just let them happen.

The symptom of the loss of ability to make, or initiate, voluntary movement is often called abulia or, in the extreme, akinetic mutism. The bradykinesia and akinesia of patients with Parkinson disease is a symptom complex of the same type, but milder. The alien hand phenomenon is the feeling that the hand does not belong to the person and is often characterized by unwanted movements that arise without any sense of their being willed. In one extreme form, called diagonistic dyspraxia, there is intermanual conflict, with the left hand performing actions contrary to actions performed by the right hand (Aboitiz et al. 2003; Scepkowski & Cronin-Golomb 2003).

Although psychogenic movements look voluntary, patients say that they are involuntary (Hallett et al. 2006). Their etiology is unknown, but they are thought to have a "psychiatric" origin, perhaps via a "conversion" mechanism. Similarly, in schizophrenia, their movements can look normal, but there is often the subjective impression of the patient that their movements are being externally (or alien) controlled.

3 A Model for Voluntary Movement

It thus becomes necessary to model voluntary movement as a process of movement generation coupled with a process of perception of volition. While it is parsimonious to model volition only on the perception side, we can leave open the possibility that there is another element of volition on the generation side. The simple model would consist of movement motivation, movement planning, movement execution, and movement perception. The perception component includes willing and agency.

In more detail, the model is illustrated in figure 7.1A. Movement begins with a motivation, and this leads to planning of a movement. When planned, it can be executed. The perceptual component is alerted to the upcoming movement from both the planning and execution modules by a feedforward or corollary discharge signal. The movement generates feedback to the perceptual component, which also includes error feedback from the movement to help insure its eventual accuracy. Thus it must also feed back to earlier modules to affect the motor command.

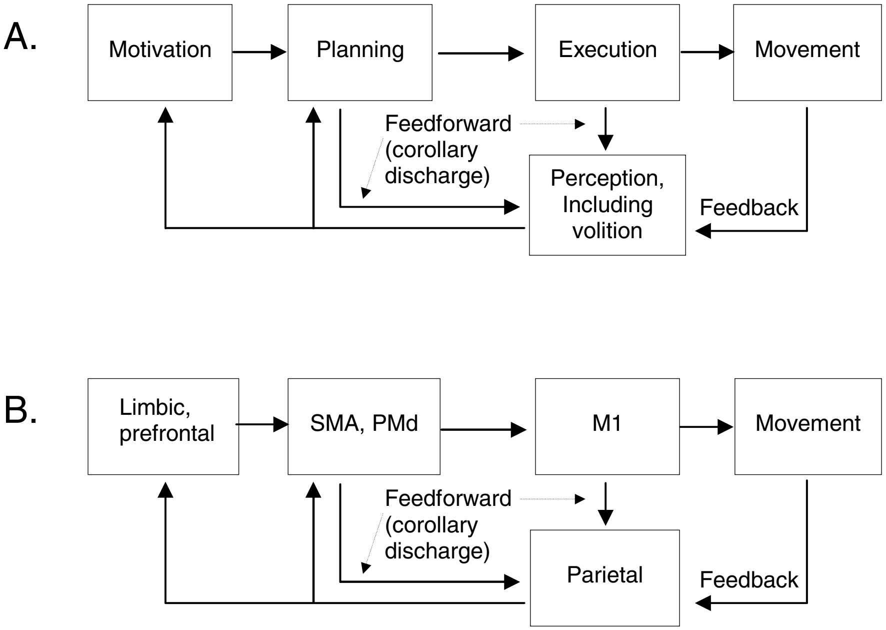

Fig. 7.1. A. Possible model of movement generation, volition, and agency. The blocks indicate functional activities and the arrows indicate time.
B. Neuroanatomical map with possible regional substrates for the functions diagrammed in part A. SMA supplementary motor area, PMd dorsal part of the premotor cortex, M1 primary motor cortex. Part B is modified from Figure 7 from Hallett (2007) with permission.

While the diagram suggests a beginning, it is important to realize that the brain is always working and always making various movements. The absence of movement is death. It might be better to think about the process as a continuous one. For each movement, there is not a "big bang." The relevant question, at any particular moment, is why was that specific movement made.

4 Motivation and Planning

Motivations for movement generally arise from limbic and frontal parts of the brain, in response to homeostatic drives, emotional forces, and desire for rewards. Hunger, thirst, warmth, sleepiness, love, sex drive, full bladder, seeking pleasure, and avoiding pain are forces that will lead to behavior. These influences focus onto the mesial motor areas, supplementary motor area (SMA), cingulate motor area (CMA), premotor cortex, dorsal and ventral (PMd, PMv), and dorsolateral prefrontal cortex (DLPFC) where movements are planned and initiated.

Functional imaging studies can reveal what regions are active with movement selection, that is, the "free" decision of *what* movement to make. Using blood flow positron emission tomography (PET), Deiber et al. have investigated movement selection in a series of studies. In the first study, normal subjects performed five different motor tasks consisting of moving a joystick on hearing a tone (Deiber et al. 1991). In the control task they always pushed it forwards (fixed condition), and in four other experimental tasks the subjects had to select between four possible directions of movement depending on instructions, including one task where the choice of movement direction was to be freely chosen and random. The greatest activation was seen in this latter task with significant increases in regional cerebral blood flow most prominently in the SMA. In a second study, normal subjects were asked to make one of four types of finger movements depending on instructions (Deiber et al. 1996). Details here were better controlled and included a rest condition. Of the numerous comparisons, the critical one for the discussion here is between the fully specified condition and the freely chosen, random movement. The anterior part of the SMA was the main area preferentially involved with the freely chosen movement.

Another aspect of movement selection is the choice of *when* to move. This was approached by Jahanshahi et al. using PET (Jahanshahi et al. 1995). Normal subjects, in a first task, were asked to make self-initiated right index finger extensions on average once every 3 s. A second task was externally triggered finger extension with the rate yoked to that of the self-initiated task. Greater activation of the right DLPFC was the only area that significantly differentiated the self-initiated movements from the externally triggered movements. In a follow-up PET experiment, measurements of regional cerebral blood flow were made under three conditions: rest, self-initiated right index finger extension at a variable rate of once every 2-7 s, and finger extension triggered by pacing tones at unpredictable intervals (at a rate yoked to the self-initiated movements). Compared with rest, unpredictably cued movements activated the contralateral primary sensorimotor cortex, caudal

SMA and contralateral putamen. Self-initiated movements additionally activated rostral SMA, adjacent anterior cingulate cortex and bilateral DLPFC.

A similar experiment was conducted by Deiber et al. using fMRI focusing on the frontal mesial cortex (Deiber et al. 1999). There were two types of movements, repetitive or sequential, performed at two different rates, slow or fast. Four regions of interest (pre-SMA, SMA, rostral cingulate motor area, CMAr, and caudal cingulate motor area, CMAc) were identified anatomically on a high-resolution MRI of each subject's brain. Descriptive analysis, consisting of individual assessment of significant activation, revealed a bilateral activation in the four mesial structures for all movement conditions, but self-initiated movements were more activating than visually triggered movements. The more complex and more rapid the movements, the smaller the difference in activation efficiency between the self-initiated and the visually triggered tasks; this indicated an additional processing role of the mesial motor areas involving both the type and rate of movements. Quantitatively, activation was more for self-initiated than for visually triggered movements in pre-SMA, CMAr and CMAc.

Stephan et al. (2002) used neuroimaging to identify structures that were activated with consciously made movements more than subconscious ones. Subjects were asked to tap their right index finger in time with different rhythmic tone sequences. One sequence was perfectly regular and others had deviations of the timing of the tones by 3, 7 and 20%. Only with the 20% variance were subjects aware of having to alter the timing of the tapping. When done at a subconscious level (3%), movement adjustments were performed employing bilateral ventral mediofrontal cortex. Awareness of change without explicit knowledge of the nature of change (7%) led to additional ventral prefrontal and premotor but not dorsolateral prefrontal activations. Only fully conscious motor adaptations (20%) showed prominent involvement of anterior cingulate and dorsolateral prefrontal cortex. The authors proposed that "these results demonstrate that while ventral prefrontal areas may be engaged in motor adaptations performed subconsciously, only fully conscious motor control which includes motor planning will involve dorsolateral prefrontal cortex." In another experiment, free selection of movement was contrasted with externally specified selection of movement (Lau et al. 2004b). The conclusion was that the DLPFC was associated with selection of either type, but the pre-SMA was specifically associated with the free selection.

The self-initiation of movement and conscious awareness of movement appear to involve mesial motor structures. As pointed out by Paus, the mesial motor structures and the anterior cingulate cortex in particular is a place of convergence for motor control, homeostatic drive, emotion, and cognition (Paus 2001).

5 Voluntary Movement without Prior Consciousness

It is clear that not all movement that people would say is voluntary is preceded by a conscious decision. Sometimes it is possible to recognize this by introspection. For example, when trying to make a vocal rapid response to a question, a person might recognize that he or she speaks the answer prior to being aware of the answer. This was proven with studies of movement triggered by a backwardly masked stimulus. Backward masking is the phenomenon where a large stimulus that quickly follows a small stimulus will block the small stimulus from being perceived.

Taylor and McCloskey demonstrated that voluntary movements could be triggered by backwardly masked visual stimuli (Taylor & McCloskey 1990). Large and small visual stimuli were presented to normal human subjects in two different experiments. In some trials, the small stimulus was followed 50 ms later by the large stimulus. In perception experiments, they demonstrated in this circumstance that the small stimulus was not perceived even with forced-choice testing. In reaction time (RT) experiments, the RTs for responses to the masked stimulus were the same as those for responses to the easily perceived, nonmasked stimulus. Hence, subjects were reacting to stimuli not perceived. In this circumstance, the order of events was stimulus-response-perception, and not stimulus-perception-response that would seem necessary for the ordinary view of free will.

Subsequently these authors extended this work by using large and small stimuli in two visual locations that signaled two different types of movement (Taylor & McCloskey 1996). Large and small stimuli were presented in either location, and in some trials, the small stimulus was followed 50 ms later by the large stimulus in both locations. In this circumstance, the small stimulus was "masked" by the large stimulus and could not be perceived on forced-choice testing. Despite not perceiving the test stimulus, subjects were able to select and execute the motor response appropriate for each location. The RTs for responses to the masked stimulus and to the same stimulus presented without masking were the same. The authors concluded that "this result implies that appropriate programs for two separate movements can be simultaneously held ready for use, and that either one can be executed when triggered by specific stimuli without subjective awareness of such stimuli and so without further voluntary elaboration in response to such awareness." In this situation, the order of events would have to be stimulus-"response selection"-response-perception.

Similar results have been obtained in experiments using weak and strong electric shock stimuli to the palm as a trigger for movement (MacIntyre & McComas 1996).

6 Timing of the Perception of Volition

The classic experiment first identifying the time of awareness of volition was reported by Libet, Gleason, Wright and Pearl (1983). Subjects sat in front of a clock with a rapidly moving spot and were told to move at will. Subsequently, they were asked to say what time it was (where the spot was) when they had the first subjective experience (the quale) of intending to act (this time was called W for "will"). They also were asked to specify the time of awareness of actually moving (this time was called M). There were two types of voluntary movements, one type was thoughtfully initiated and a second type was "spontaneous and capricious." As a control for the ability to successfully subjectively time events, subjects were also stimulated at random times with a skin stimulus and they were asked to time this event (called S). At the same time, EEG was being recorded and movement-related cortical potentials (MRCPs) were assessed to determine timing of activity of the brain.

The MRCP prior to movement has a number of components (Jahanshahi & Hallett 2003; Shibasaki & Hallett 2006). An early negativity preceding movement, the *Bereitschaftspotential* or BP, has two phases, an initial, slowly rising phase lasting from about 1500 ms to about 400 ms before movement, the early BP or BP1, and a later, more rapidly rising phase lasting from about 400 ms to approximately the time of movement onset, late BP, BP2, or "the negative slope" (NS'). The topography of the early BP is generalized with a vertex maximum. With the late BP the negativity begins to shift to the central region contralateral to the hand that is moving. The main contributors to the early BP are the premotor cortex and the supplementary motor area (SMA), both bilaterally (Toma et al. 2002). With the appearance of the late BP the activity of the contralateral primary motor cortex (M1) becomes prominent. With thoughtful, preplanned movements, the BP began about 1050 ms prior to EMG onset (the type I of Libet), and with more spontaneous movements, the BP began about 575 ms prior to movement (the type II of Libet) (Libet et al. 1982). The MRCP is a direct measure of activity in the brain that is related to the genesis of movement.

Subjects were reasonably accurate in determining the time of S, indicating that this method of timing of subjective experience was acceptable. W occurred about 200 ms prior to EMG onset and M occurred about 90 ms prior to EMG onset. The onset of the BP type I occurred about 850 ms prior to W, and the onset of the BP type II occurred about 375 ms prior to W (fig. 7.2A). The authors concluded "that cerebral initiation of a spontaneous, freely voluntary act can begin unconsciously, that is, before there is any (at least recallable) subjective awareness that a 'decision' to act has already been initiated cerebrally" (Libet et al. 1983).

These results have been reproduced by many others, so the basic data are really not in question. Haggard and Eimer looked carefully at the timing of W compared with BP onset and the onset of another measure, the lateralized readiness potential (LRP, the difference in the voltage of right and left central regions) in tasks where subjects moved either their right or left hands (Haggard & Eimer 1999). The LRP timing was similar to the late BP component indicating the onset of asymmetry of the cortical activity relating to the hand that will eventually move. The onset of the LRP proceeded W. Across subjects they found a better relationship between the timing of the onset of the LRP and W than between the onset of the BP and W, and suggested that the "processes underlying the LRP may cause our awareness of movement initiation." This work suggested that movement selection also precedes awareness.

In other work, Haggard et al. looked at the timing of M with respect to movement in more detail (Haggard et al. 1999). They looked at M in relation to the initiation of sequences of movements of various lengths. Sequences of longer length take a longer time to prepare for execution. In such circumstances, M occurs more in advance of the first movement of the sequence. This implies that the awareness of actions is associated with "some pre-motor event after the initial intention and preparation of action, but before the assembly and dispatch of the actual motor command to the muscles."

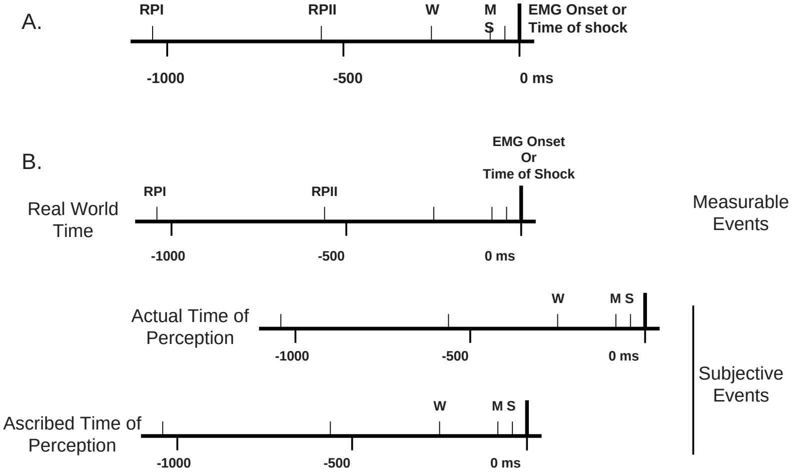

Fig. 7.2. A. Timing of subjective events and the *Bereitschaftspotential* (readiness potential, RP) with data from Libet et al. 1983. RPI is the onset of the *Bereitschaftspotential* with ordinarily voluntary movements, and RPII is the onset with movements made quickly with little forethought. W is the subjective timing of the will to move, M is the subjective timing of the onset of movement, S is the subjective timing of a shock to the finger. EMG onset or shock delivery is set at zero ms. Part A is Figure 2 from Hallett (2007) with permission.

B. Possible timing of subjective events in comparison to measurable events in the course of making voluntary movements. This is similar to figure 7.2A, but the subjective events and measurable events are plotted on separate time lines. The subjective events are plotted twice, once at the time they are ascribed to in real world time and once when they might actually have occurred. The latter is only hypothetical, but is necessarily in the right direction from the ascribed times. Part B is Figure 6 from Hallett (2007) with permission.

6.1 Criticisms of the "Libet Clock" Experiment

The interpretation of the data from the Libet clock experiments has been subject to much discussion and criticism. One issue is what really designates the intention to move. It could be argued that the decision to move is made when agreeing to do the experiment in the first place (Deecke & Kornhuber 2003; Mele 2006, 2007). The movements themselves are then a simple consequence of that earlier choice.

Moreover, there are data that indicate that the more specific the decision about future behavior, the more likely that the behavior will actually occur. Specifically, having an "implementation intention," a plan to implement a goal, is more effective than a general "goal intention" (Gollwitzer & Sheeran 2006). What is the nature of these decisions? These are "thinking," and thinking is another element of consciousness that we do not fully understand. However, thinking, like movement, is a manifestation of brain function, and a decision such as agreeing to do the experiment biases the probabilities of movement selection.

Libet himself argued that his results did not invalidate the concept of free will. His view was that the movement was indeed initiated subconsciously, but subject to veto once it reached consciousness (Libet 1999, 2006). This veto power could be considered free will. This is a somewhat unusual way of looking at the issue, and this power has been designated "free won't" (Obhi & Haggard 2004). Of course, "free won't" could also be initiated subconsciously and could be basically the same process as free will. For example, there is a cortical potential prior to relaxation of a tonic movement that is similar to the *Bereitschaftspotential* (Terada et al. 1995).

Another type of concern is the nature of subjective time and its variable relationship to real time (Eagleman et al. 2005). One aspect of this is that the subjective present is actually slightly in the real past. It takes time for sensory information to reach the brain, and these times are variable for different types of input. There has to be time to allow this information to be aligned for a unitary percept. Several experiments reveal some of the features of subjective time. In the flash-lag illusion, a flash is given together with a moving object in the same location. However, the moving object is seen to be where the moving object is approximately 80 ms after the flash. This appears to be due to a process of postdiction where the percept attributed to a specific time is modulated by what happens in the subsequent 80 ms (Eagleman & Sejnowski 2000). If someone presses a key regularly and sees a resultant flash at a particular interval, they get sufficiently linked such that if the key press to flash interval is shortened, persons get the sense that the flash occurs prior to the key press (Stetson et al. 2006). In another experiment, persons pressed a key and then heard a tone at variable intervals. The subjective time for these two events was determined and the interval between the keypress and tone was erroneously short when the real interval was relatively short and more accurate when the interval was longer (Haggard et al. 2002). This did not happen when the movement was caused by a TMS pulse. Hence, intention appears to bind the movement and consequence closer together.

Because of these problems with subjective time, we have approached the problem in a different way. In preliminary experiments, we asked subjects to make movements at freely chosen times while listening to tones occurring at random times (Matsuhashi & Hallett 2006). If a tone came after the thought to make a movement, but before the movement, the subject was to veto the movement. No introspective data are needed to interpret the data; this suggested that the time interval between intending to move and movement is longer than that of the Libet W, but still not as long as the MRCP. This experiment can be considered a study of "free won't."

6.2 Events in the Immediate Premovement Period

It is important to recognize that movement genesis is not a strictly linear process, specifically, movement does not obligatorily occur a fixed number of ms following the onset of the BP. The initiation process may vacillate depending on the various influencing factors. As noted earlier, Libet pointed out that the upcoming movement intention might be "vetoed" after it becomes conscious (Libet 1985, 1999). "Conscious vetoing of a conscious intention" can occur up until the point of "no return." The point of no return is ordinarily studied in a reaction time situation where a go stimulus is followed by a no-go stimulus, and is very close to the time of the expected movement (Mirabella et al. 2006).

The anatomical correlate of vetoing has been studied by Brass and Haggard (2007). Subjects were investigated with fMRI and were asked to periodically make voluntary hand movements. On some occasions they were asked to veto their movement (and not move) after having made the decision to move. Timing of the decision to move was determined by the Libet clock methodology. They found that a specific area in the fronto-median cortex was more strongly activated with veto compared to when they actually made the movement. In some ways this is similar to experiments that have been conducted with EEG that have identified a frontally predominant "nogo" potential when a prepared movement is not carried out (Fallgatter 2001).

A modified version of the Libet clock experiment has been done with fMRI (Soon et al. 2008). Subjects made movements of right or left finger at freely chosen times while watching a series of letters. They indicated the time of choice by indicating the letter they were seeing. Using a sophisticated analysis method, they were able to identify with up to 60% probability of the right or left choice as long as 10 seconds prior to the movement. The probability is not high, but might be what is expected 10 seconds prior to a movement. The brain likely starts planning early and the probabilities oscillate, but the probability of the final outcome may begin to climb well in advance of the actual movement.

6.3 Dissociating the Timing of the Sense of Volition and the Actual Movement

The timing of perception of W and M can be influenced by TMS over the pre-SMA delivered "immediately after the action" or 200 ms later (Lau et al. 2007). This had the effect of moving the W judgment earlier in time and the M judgment later in time. This effect was time specific and did not occur with stimulation over the primary motor cortex. There are a number of conclusions. Subjective timing of events that are felt to occur prior to the movement may be influenced after the movement. This poses another problem for the method of subjective timing, but also might be consistent with the possibility that the sense of W actually does occur after the movement. Indeed, as noted earlier, there must be a delay between any event in the real world and its perception. Perhaps the delay is sufficiently long so that the real time of W is after movement onset even if it is perceived to be before movement onset (fig. 7.2B). Moreover, these results are consistent with a role for the mesial motor areas in the subjective sense of volition, a topic that will be discussed later.

Other experiments show that, in a reaction time movement, the judgment of M is not fixed to the movement itself. A strong, single pulse TMS over the motor cortex during the reaction period of a reaction time movement can delay the execution of the movement without affecting its form. In experiments where the RT is delayed with TMS over the motor cortex, the judgment of when movement occurred is delayed less than the movement itself (Haggard & Magno 1999). A startling stimulus delivered at the time of the go signal can shorten reaction time. In this situation, again the M judgment does not move (Sanegre et al. 2004). These studies suggest that even movement awareness (M) and actual movement execution are processed by parallel pathways.

6.4 Sense of Volition Depends on Sense of Causality

Wegner argues that free will is an illusion derived from the relationship between one's thought and the movement itself (Wegner 2002, 2003, 2004). The thought must occur before the movement; it must be consistent with the movement; and there must not be another obvious cause for the movement. These features imply causality, that is, that the thought led to the movement. That W precedes M is critical for people to believe that a movement is voluntary. Patients with schizophrenia with passivity phenomena do not have the sense that they are in control of their movements. Preliminary work from our group has shown that there is a foreshortened interval between W and M (Pirio Richardson et al. 2006).

7 Anatomy of the Sense of Volition

The awareness of W, as well as of M, could well derive from feedforward signals (corollary discharges) (Poulet & Hedwig 2007) from the movement planning and the movement execution since all of this certainly occurs prior to movement feedback. Indeed, it has been demonstrated that movement feedback is not necessary (Frith 2002; Frith et al. 2000).

Since attention accentuates brain activity, it should be possible to help identify what areas are involved with intention by directing attention to intention itself. In the Libet clock experiments, attention is directed to intention in the W condition. Looking at the MRCPs in the W condition compared with the M condition showed a larger amplitude in the W condition (Sirigu et al. 2004). Using fMRI, the W condition (called the I condition in the paper) produced greater activation in the pre-SMA, right dorsal prefrontal cortex and left interparietal sulcus (Lau et al. 2004a). With connectivity analysis, the pre-SMA and prefrontal areas were correlated, but not the parietal area. The authors suggest that the pre-SMA is the critical area for the sense of intention. An alternate interpretation might be that the frontal area reflects the movement genesis and the parietal area reflects the sense of volition. Another experiment showed that attention to M compared with movement without attention yielded activation in the cingulate motor area (CMA) (Lau et al. 2006), another structure that should be involved in movement genesis. The authors noted that M was earlier in time when the CMA was more active, and that W was earlier in time when pre-SMA was more active. This suggests another difficulty in the subjective measurement of W and M, in that they depend on attention.

Evidence that the parietal lobe is relevant to the sense of voluntariness comes from experiments with the Libet clock in five patients with parietal lobe lesions (Sirigu et al. 2004). These patients were able to make voluntary movements with normal force although two of the patients had apraxia and one of these two also had a mild sensory disturbance. While their estimation of M was in the normal range, their estimate of W was a much smaller interval from EMG onset than normal: -55.0 ms compared with -239.2 ms. A cerebellar patient group was also investigated as another control group, and their performance was normal. These data suggest that the parietal lobe plays a part in the awareness of voluntary action and this awareness is delayed if the parietal cortex is damaged.

8 Anatomy of the Sense of Agency

An imaging study has investigated the sense of "agency," the feeling that leads us to attribute an action to ourselves rather than to another person (Farrer et al. 2003). For there to be agency, there has to be a match of the intentional command and movement feedback. The investigators used a device that allowed them to modify the subject's degree of control of the movements of a virtual hand presented on a screen. During a blood-flow PET study, they compared four conditions: (1) a condition where the subject had a full control of the movements of the virtual hand, (2) a condition where the movements of the virtual hand appeared rotated by 25 degrees with respect to the movements made by the subject, (3) a condition where the movements of the virtual hand appeared rotated by 50 degrees, and (4) a condition where the movements of the virtual hand were produced by another person and did not correspond to the subject's movements. In the inferior part of the parietal lobe, specifically on the right side, the less the subject felt in control of the movements of the virtual hand, the higher the level of activation (fig. 7.3). In the insula, the more the subject felt in control, the more the activation.

Evidence that the insula is relevant comes from a variety of sources. In an analysis of 27 stroke patients, the symptom of anosognosia for the contralateral limb was commonly associated with damage to the posterior insula (Karnath et al. 2005). The insula is a site of convergence of information about the physiological condition of all parts of the body, and can be considered the center for interoception (Craig 2003). This may help construct a sense of self (Damasio 2003). Indeed, the role of the insula might be to indicate the "body ownership" of a movement rather than its voluntary nature (Tsakiris et al. 2007). There seems to be more evidence, however, for parietal areas to be more relevant in the sense of agency.

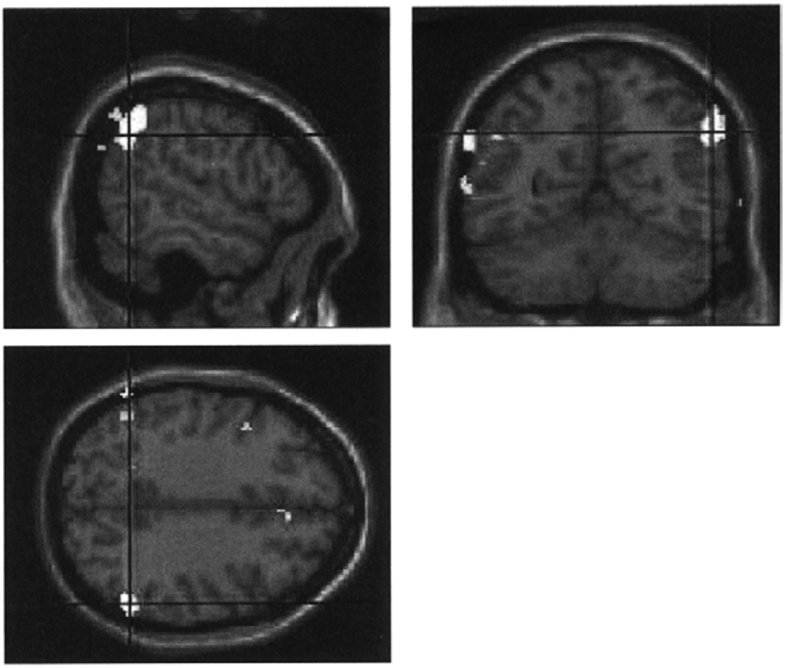

Fig. 7.3. PET scan showing the parietal region that was inversely correlated with the sense of agency. Subjects moved their hand and observed a virtual hand with variable relationship to their own movements, which gave varying sense of agency. Adapted from Farrer et al. 2003.

A similar experiment with joystick movements was carried out in which subjects were to decide whether a visual display of the movement was their own or a different similar one (David et al. 2007). The investigators focused attention on the extrastriate body area (EBA), a region defined by its "specific" activation with images of the human body. There were two findings of interest. The EBA was more active when the visual display was dissimilar to the subject's own movement. Connectivity was greater between the EBA and the posterior parietal cortex when the subject was correct about the self-other judgment.

One of the difficulties with these experiments is that there is both a sensorimotor mismatch and a loss of agency, and it is possible that the parietal activation is indicating only the former. This issue was directly investigated using several other experiments where subjects' movements were displayed back to them with various temporal delays (Farrer et al. 2008). Delays up to 400 ms led to mismatches, but subjects still had the sense of agency. On the other hand, delays of 800 ms or more lead to a loss of the sense of agency. For the shorter delays, the investigators used the placement of pegs into a pegboard, and for the longer delays, the investigators used finger wiggling. In both experiments, they identified similar areas around the right angular gyrus being most relevant.

If the right parietal area is relevant for agency, then disruption of its function with TMS might be able to confuse the determination of agency. An experiment of this type was carried out using finger pointing movements where subjects saw either their own movements or ones slightly deviated by a computer algorithm (Preston & Newport 2008). TMS made it more likely that the subject would decide that he was not the agent.

9 Conclusions

There is no evidence yet identified for free will as a force in the generation of movement, but there is increasing understanding about how qualia of several aspects of free will are generated. While the qualia of free will are powerful and common, they are subject to manipulation and illusion, and can go awry in brain disorders. With the evidence reviewed in this manuscript, it is possible to make tentative neuroanatomical assignments for the functions of movement generation, willing, and agency noted in figure 7.1A. These are diagrammed in figure 7.1B. Mapping the model onto brain structures indicates that movement is likely initiated in mesial motor areas (SMA including pre-SMA and CMA) and premotor cortex (PMd) which are in turn influenced by prefrontal and limbic areas. The movement command goes to primary motor cortex. Corollary discharges likely come from the SMA and PMd to parietal area, and these may be responsible for the sense of volition (W). Parietal and frontal areas maintain a relatively constant bidirectional communication. It is likely that this network of structures includes the insula. The sense of agency comes from the appropriate match of volition and movement feedback, likely also centered on the parietal area.

Acknowledgment

This work was supported by the Intramural Program of the NINDS, NIH. Since this is work of the U.S. government, there is no copyright. The text has been taken in part from Hallett (2007), but extensively revised and updated.

References

Aboitiz, F., Carrasco, X., Schroter, C., Zaidel, D., Zaidel, E., Lavados, M.: The alien hand syndrome: Classification of forms reported and discussion of a new condition. Neurological Sciences 24, 252–257 (2003)

Brass, M., Haggard, P.: To do or not to do: The neural signature of self-control. Journal of Neuroscience 27, 9141–9145 (2007)

Craig, A.D.: Interoception: The sense of the physiological condition of the body. Current Opinion in Neurology 13, 500–505 (2003)

Damasio, A.: Mental self: The person within. Nature 423, 227 (2003)

David, N., Cohen, M.X., Newen, A., Bewernick, B.H., Shah, N.J., Fink, G.R., et al.: The extrastriate cortex distinguishes between the consequences of one's own and others' behavior. NeuroImage 36, 1004–1014 (2007)

Deecke, L., Kornhuber, H.H.: Human freedom, reasoned will, and the brain: The Bereitschaftspotential story. In: Jahanshahi, M., Hallett, M. (eds.) The Bereitschaftspotential: Movement-related cortical potentials, pp. 283–320. Kluwer Academic/Plenum, New York (2003)

Deiber, M.P., Passingham, R.E., Colebatch, J.G., Friston, K.J., Nixon, P.D., Frackowiak, R.S.J.: Cortical areas and the selection of movement: A study with positron emission tomography. Experimental Brain Research 84, 393–402 (1991)

Deiber, M.P., Ibañez, V., Sadato, N., Hallett, M.: Cerebral structures participating in motor preparation in humans: A positron emission tomography study. Journal of Neurophysiology 75, 233–247 (1996)

Deiber, M.P., Honda, M., Ibañez, V., Sadato, N., Hallett, M.: Mesial motor areas in self-initiated versus externally triggered movements examined with fMRI: Effect of movement type and rate. Journal of Neurophysiology 81, 3065–3077 (1999)

Eagleman, D.M., Sejnowski, T.J.: Motion integration and postdiction in visual awareness. Science 287, 2036–2038 (2000)

Eagleman, D.M., Tse, P.U., Buonomano, D., Janssen, P., Nobre, A.C., Holcombe, A.O.: Time and the brain: How subjective time relates to neural time. Journal of Neuroscience 25, 10369–10371 (2005)

Fallgatter, A.J.: Electrophysiology of the prefrontal cortex in healthy controls and schizophrenic patients: A review. Journal of Neural Transmission 108, 679–694 (2001)

Farrer, C., Franck, N., Georgieff, N., Frith, C.D., Decety, J., Jeannerod, M.: Modulating the experience of agency: A positron emission tomography study. NeuroImage 18, 324–333 (2003)

Farrer, C., Frey, S.H., Van Horn, J.D., Tunik, E., Turk, D., Inati, S., et al.: The angular gyrus computes action awareness representations. Cerebral Cortex 18, 254–261 (2008)

Frith, C.D.: Attention to action and awareness of other minds. Consciousness and Cognition 11, 481–487 (2002)

Frith, C.D., Blakemore, S., Wolpert, D.M.: Explaining the symptoms of schizo¬phrenia: Abnormalities in the awareness of action. Brain Research Reviews 31, 357–363 (2000)

Gollwitzer, P.M., Sheeran, P.: Implementation intentions and goal achievement: A meta-analysis of effects and processes. Advances in Experimental Social Psychology 38, 69–119 (2006)

Haggard, P., Eimer, M.: On the relation between brain potentials and the awareness of voluntary movements. Experimental Brain Research 126, 128–133 (1999)

Haggard, P., Magno, E.: Localising awareness of action with transcranial magnetic stimulation. Experimental Brain Research 127, 102–107 (1999)

Haggard, P., Newman, C., Magno, E.: On the perceived time of voluntary actions. British Journal of Psychology 90(Pt 2), 291–303 (1999)

Haggard, P., Clark, S., Kalogeras, J.: Voluntary action and conscious awareness. Nature Neuroscience 5, 382–385 (2002)

Hallett, M.: Volitional control of movement: The physiology of free will. Clinical Neurophysiology 118, 1179–1192 (2007)

Hallett, M., Fahn, S., Jankovic, J., Lang, A.E., Cloninger, C.R., Yudofsky, S.C. (eds.): Psychogenic Movement Disorders: Neurology and Neuropsychiatry. Lippincott Williams & Wilkins/AAN Press, Philadelphia (2006)

Jahanshahi, M., Hallett, M. (eds.): The Bereitschaftspotential: Movement Related Cortical Potentials. Kluwer Academic/Plenum, New York (2003)

Jahanshahi, M., Jenkins, I.H., Brown, R.G., Marsden, C.D., Passingham, R.E., Brooks, D.J.: Self-initiated versus externally triggered movements, I: An investigation using measurement of regional cerebral blood flow with PET and movement-related potentials in normal and Parkinson's disease subjects. Brain: A Journal of Neurology 118, 913–933 (1995)

Karnath, H.O., Baier, B., Nagele, T.: Awareness of the functioning of one's own limbs mediated by the insular cortex? Journal of Neuroscience 25, 7134–7138 (2005)

Lau, H.C., Rogers, R.D., Haggard, P., Passingham, R.E.: Attention to intention. Science 303, 1208–1210 (2004a)

Lau, H.C., Rogers, R.D., Ramnani, N., Passingham, R.E.: Willed action and attention to the selection of action. NeuroImage 21, 1407–1415 (2004b)

Lau, H.C., Rogers, R.D., Passingham, R.E.: On measuring the perceived onsets of spontaneous actions. Journal of Neuroscience 26, 7265–7271 (2006)

Lau, H.C., Rogers, R.D., Passingham, R.E.: Manipulating the experienced onset of intention after action execution. Journal of Cognitive Neuroscience 19, 81–90 (2007)

Libet, B.: Unconscious cerebral initiative and the role of conscious will in voluntary action. Behavioral and Brain Sciences 8, 529–566 (1985)

Libet, B.: Do we have free will? Journal of Consciousness Studies 9, 47–57 (1999)

Libet, B.: The timing of brain events: Reply to the "Special Section". Journal of Consciousness and Cognition (edited by Pockett, S.) 15, 540–547 (2006)

Libet, B., Wright Jr., E.W., Gleason, C.A.: Readiness-potentials preceding unrestricted 'spontaneous' vs. pre-planned voluntary acts. Electroencephalography and Clinical Neurophysiology 54, 322–335 (1982)

Libet, B., Gleason, C.A., Wright, E.W., Pearl, D.K.: Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential): The unconscious initiation of a freely voluntary act. Brain: A Journal of Neurology 106, 623–642 (1983)

MacIntyre, N.J., McComas, A.J.: Non-conscious choice in cutaneous backward masking. Neuroreport 7, 1513–1516 (1996)

Matsuhashi, M., Hallett, M.: The timing of conscious thought into action (Abstract). Clinical Neurophysiology 117(suppl. 1), S96 (2006)

Mele, A.R.: Free will and luck. Oxford University Press, Oxford (2006)

Mele, A.R.: Decision, intentions, urges and free will: Why Libet has not shown what he says he has. In: Campbell, J.K., O'Rourke, M., Silverstein, H. (eds.) Causation and explanation, Topics in Contemporary Philosophy. MIT Press, Cambridge (2007)

Mirabella, G., Pani, P., Pare, M., Ferraina, S.: Inhibitory control of reaching movements in humans. Experimental Brain Research 174, 240–255 (2006)

Obhi, S.S., Haggard, P.: Free will and free won't. American Scientist 92, 358–365 (2004)

Paus, T.: Primate anterior cingulate cortex: Where motor control, drive and cognition interface. Nature Reviews Neuroscience 2, 417–424 (2001)

Pirio Richardson, S., Matsuhashi, M., Voon, V., Peckham, E., Nahab, F., Mari, Z., et al.: Timing of the sense of volition in patients with schizophrenia (Abstract). Clinical Neurophysiology 117(suppl. 1), S97 (2006)

Poulet, J.F., Hedwig, B.: New insights into corollary discharges mediated by identified neural pathways. Trends in Neurosciences 30, 14–21 (2007)

Preston, C., Newport, R.: Misattribution of movement agency following right parietal TMS. Social Cognitive and Affective Neuroscience 3, 26–32 (2008)

Sanegre, M.T., Castellote, J.M., Haggard, P., Valls-Sole, J.: The effects of a startle on awareness of action. Experimental Brain Research 155, 527–531 (2004)

Scepkowski, L.A., Cronin-Golomb, A.: The alien hand: Cases, categorizations, and anatomical correlates. Behavioral and Cognitive Neuroscience Reviews 2, 261–277 (2003)

Shibasaki, H., Hallett, M.: What is the Bereitschaftspotential? Clinical Neurophysiology 117, 2341–2356 (2006)

Sirigu, A., Daprati, E., Ciancia, S., Giraux, P., Nighoghossian, N., Posada, A., et al.: Altered awareness of voluntary action after damage to the parietal cortex. Nature Neuroscience 7, 80–84 (2004)

Soon, C.S., Brass, M., Heinze, H.J., Haynes, J.D.: Unconscious determinants of free decisions in the human brain. Nature Neuroscience 11, 543–545 (2008)

Stephan, K.M., Thaut, M.H., Wunderlich, G., Schicks, W., Tian, B., Tellmann, L., et al.: Conscious and subconscious sensorimotor synchronization – prefrontal cortex and the influence of awareness. NeuroImage 15, 345–352 (2002)

Stetson, C., Cui, X., Montague, P.R., Eagleman, D.M.: Motor-sensory recalibration leads to an illusory reversal of action and sensation. Neuron 51, 651–659 (2006)

Taylor, J.L., McCloskey, D.I.: Triggering of preprogrammed movements as reactions to masked stimuli. Journal of Neurophysiology 63, 439–446 (1990)

Taylor, J.L., McCloskey, D.I.: Selection of motor responses on the basis of unperceived stimuli. Experimental Brain Research 110, 62–66 (1996)

Terada, K., Ikeda, A., Nagamine, T., Shibasaki, H.: Movement-related cortical potentials associated with voluntary muscle relaxation. Electromyography and Clinical Neurophysiology 95, 335–345 (1995)

Toma, K., Matsuoka, T., Immisch, I., Mima, T., Waldvogel, D., Koshy, B., et al.: Generators of movement-related cortical potentials: fMRI-constrained EEG dipole source analysis. NeuroImage 17, 161–173 (2002)

Tsakiris, M., Hesse, M.D., Boy, C., Haggard, P., Fink, G.R.: Neural signatures of body ownership: A sensory network for bodily self-consciousness. Cerebral Cortex 17(10), 2235–2244 (2007)

Wegner, D.M.: The illusion of conscious will. MIT Press, Cambridge (2002)

Wegner, D.M.: The mind’s best trick: How we experience conscious will. Trends in Cognitive Sciences 7, 65–69 (2003)

Wegner, D.M.: Précis of The illusion of conscious will. Behavioral and Brain Sciences 27, 649–659 (2004); discussion 659–692

8

How We Recognize Our Own Actions

Sarah-Jayne Blakemore

Institute of Cognitive Neuroscience
University College London
17 Queen Square
London WC1N 3AR
United Kingdom
s.blakemore@ucl.ac.uk

Summary. This chapter first describes how predicting the sensory consequences of action contributes to the recognition of one's own actions. Second, the chapter discusses three symptoms in which this prediction mechanism is proposed to be impaired: the consequences of parietal lobe damage, passivity experiences associated with schizophrenia, and phantom limbs.

Keywords: action prediction, forward models, internal models, delusions of control, auditory hallucinations, fMRI, PET.

1 Forward Models and Prediction of Action

When you make an arm movement, you instantly recognize that movement as your own. It has been proposed that the recognition of action is achieved by predicting the sensory consequences of motor commands whenever movements are made (Frith et al. 2000). One way that the brain predicts the consequences of movement is by using a forward model. Forward models use the efference copy of motor commands to predict the sensory consequences of a movement (see fig. 8.1). This prediction is then compared with the actual feedback from the movement. When there is little or no discrepancy between the predicted and actual sensory consequences of a movement, the movement is classified as self-produced, and its consequences can be attenuated relative to external sensory events.

1.1 Why Can't You Tickle Yourself?

An example of this is the phenomenon that people cannot tickle themselves (Weiskrantz et al. 1971). We carried out a series of experiments to investigate why this is the case. In the first set of experiments, subjects were asked to rate the sensation of a tactile stimulus on the palm of their hand when the correspondence between self-generated movement and its sensory consequences was altered. Subjects moved a robotic arm with their left hand and this movement caused a second foam-tipped robotic arm to move across their right palm. By using this robotic interface so that the tactile stimulus could be delivered under remote control by the subject, delays of 100, 200, and 300 milliseconds were introduced between the movement of the left hand and the tactile stimulus on the right palm. The result was that the sensory stimulus no longer corresponds to what is predicted, so as the delay is increased the sensory prediction becomes less accurate (see fig. 8.1). The results showed that subjects rated self-produced tactile stimulation as being less tickly, intense, and pleasant than an identical stimulus produced by the robot (Blakemore et al. 1999). Furthermore, subjects reported a progressive increase in the tickly rating as the delay was increased.

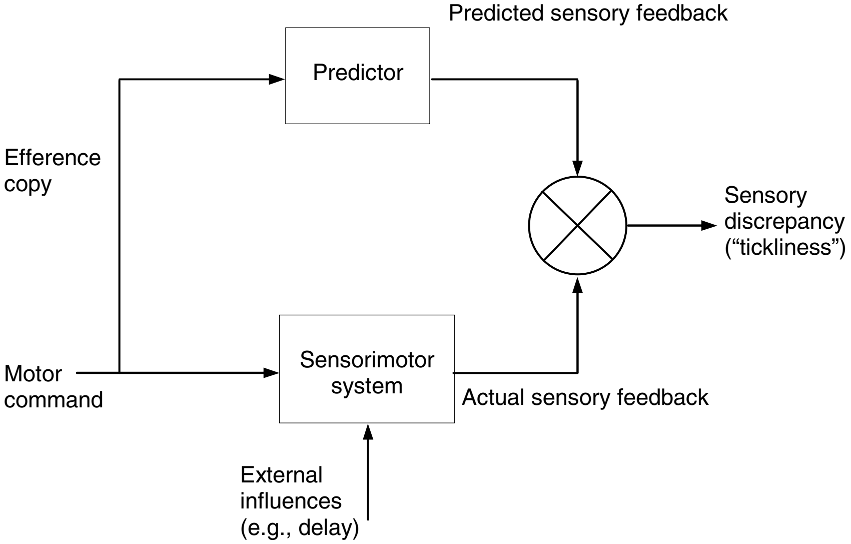
Predicted sensory feedback
Predictor
Efference copy
Sensory discrepancy ("tickliness")
Motor command
Sensorimotor system
Actual sensory feedback
External influences (e.g., delay)

Fig. 8.1. The Forward Model

These results suggest that the perceptual attenuation of self-produced tactile stimulation is due to precise sensory predictions. When there is no delay, a forward model correctly predicts the sensory consequences of the movement, so no sensory discrepancy ensues between the predicted and actual sensory information, and the motor command to the left hand can be used to attenuate the sensation on the right palm. As the sensory feedback deviates from the prediction of the model (by increasing the delay) the sensory discrepancy between the predicted and actual sensory feedback increases, which leads to a decrease in the amount of sensory attenuation.

In the second series of experiments, we investigated the neural basis of this phenomenon. In an fMRI study, subjects experienced tactile stimulation on their palm that was produced either by the subject himself or by the experimenter (Blakemore et al. 1998). The results showed an increase in activity of the secondary somatosensory cortex (SII) and the anterior cingulate cortex (ACC) when subjects experienced an externally produced tactile stimulus relative to a self-produced tactile stimulus. The reduction in activity in these areas in response to self-produced tactile stimulation might be the physiological correlate of the reduced perception associated with this type of stimulation. While the decrease in activity in SII and ACC might underlie the reduced perception of self-produced tactile stimuli, the pattern of brain activity in the cerebellum suggests that this area is the source of the SII and ACC modulation. In SII and ACC, activity was attenuated by all movement: these areas were equally activated by movement that did and that did not result in tactile stimulation. In contrast, the right anterior cerebellar cortex was selectively deactivated by self-produced movement which resulted in a tactile stimulus, but not by movement alone, and significantly activated by externally produced tactile stimulation. This pattern suggests that the cerebellum differentiates between movements depending on their specific sensory consequences.

A further experiment supported this hypothesis. When delays were introduced between the movement and its tactile consequences, cerebellar activity increased (Blakemore et al. 2001). The higher the delay, the higher was activity in the cerebellum. We suggested that the cerebellum is involved in signaling the discrepancy between predicted and actual sensory consequences of movements.

1.2 Actions Can Be Unavailable to Awareness

It has been proposed that sensory prediction is the basis of our awareness of action (Frith et al. 2000). There are a number of demonstrations that many aspects of action occur without awareness. In our tactile experiment described above in which delays were introduced between subjects' actions and their consequences (Blakemore et al. 1999), we asked subjects at the end of the experiment whether they had noticed the delays. None of the subjects claimed to have noticed the delays. This suggests that subjects are unaware of sensory discrepancies between the predicted and actual consequences of movement.

Further evidence that sensations associated with actual movements are unavailable to awareness comes from a study in which the sensory consequences of movement were made to deviate from subjects' expectations (Fourneret & Jeannerod 1998). In this study, the subjects' task was to draw a straight line on a computer screen. Subjects could not see their arm or hand and were given false feedback about the trajectory of their arm movement. Thus they had to make considerable deviations from a straight movement in order to achieve their goal. Verbal reports indicated that subjects were unaware that they were making deviant movements – they claimed to have made straight movements. These results suggest that we are aware of the movements we intend rather than the movements we actually make.

These studies suggest that, as long as actions more or less correspond to intentions, subjects are not aware of the action and its consequences, but instead are aware of the prediction.

2 Abnormalities in the Control and Awareness of Action

In the remainder of this chapter, I will briefly describe two examples of neurological experiences (the results of parietal lesions and phantom limbs) and an example of a psychiatric symptom (delusion of control/passivity) in which awareness of action is somehow changed. Further examples can be found in Frith et al. (2000).

2.1 Parietal Lobe Lesions

Damage to the parietal lobe often causes problems with the control and awareness of action. These cases suggest that the parietal lobe plays a role in producing a sense of agency and the conscious representation of actions, a proposal that has been supported by neuroimaging experiments (Frith et al. 2000). Patients with left parietal lobe damage confuse their hand movements with those of another agent (Sirigu et al. 1999). Further evidence for the parietal lobe's role in storing representations of movement come from brain damaged patients who lose feeling in the limb controlateral to the lesion. Wolpert, Goodbody, and Husain (1998) describe a patient who had a large cyst in the left parietal lobe and reported the experience of the position and presence of her right limbs fading away over seconds if she could not see them. Her experience of a constant tactile stimulus or a weight also faded away, but changes in such sensations could be detected. Slow reaching movements to peripheral targets with the right hand were inaccurate, but reaching movements made at a normal pace were unimpaired. In this case, there seemed to be a circumscribed problem with the representation of the current limb position in that it could not be maintained in the absence of changing stimulation. Wolpert, Goodbody, and Husain postulated that the parietal cortex is involved in both maintaining and updating an internal body state issued from sensory and motor signals.

2.2 Delusions of Control/Passivity Experiences Associated with Schizophrenia

Certain psychiatric symptoms are characterized by an inability to distinguish self- and externally produced actions. Many patients with schizophrenia describe "passivity" experiences in which actions, thoughts, or emotions are made for them by some external agent rather than by their own will. The experience of passivity might arise from a lack of awareness of the predicted limb position (Frith et al. 2000). The idea is that the forward model prediction somehow does not reach awareness in these patients. In the presence of delusions of control, the patient is not aware of the predicted consequences of a movement and is therefore not aware of initiating a movement. In parallel, the patient's belief system is faulty so that he interprets this abnormal sensation in an irrational way.

Several studies have shown that patients with delusions of control confuse self-produced and externally generated actions. Using the paradigm in which subjects see feedback of their own hand movement or that of the experimenter's hand making similar movements, Daprati and her colleagues found that schizophrenic patients with delusions of control are more likely than control subjects to confuse their hand with that of the experimenter (Daprati et al. 1997). These patients have difficulty distinguishing between correct visual feedback about the position of their hand and false feedback when the image of the hand they see is in fact that of another person attempting to make the same movements as the patient. One explanation for this is that the patients only have proprioceptive and visual feedback to rely on for recognition whereas normal control subjects are able additionally to compare the sensory prediction with the sensory feedback from the movement.

Evidence that this confusion between self and other in patients with delusions of control is a consequence of an abnormal sensory prediction comes from studies based on the finding that, normally, because a movement is predicted, its sensory consequences can be perceptually attenuated relative to external sensations (Blakemore et al. 1999). Patients with delusions of control do not show this perceptual attenuation of self-produced sensory stimulation (Blakemore et al. 2001). The normal perceptual attenuation of the sensory consequences of movement is accompanied by a reduction in activity in SII and ACC compared with an external sensory stimulus (Blakemore et al. 1998). If delusions of control are associated with an impairment in sensory prediction, we would expect to see no attentuation of the activity in sensory regions. This was similar to the result of a study in which schizophrenic patients with and without delusions of control were scanned while they performed a movement task (Spence et al. 1997). The presence of delusions of control was associated with overactivity in right inferior parietal cortex. Moreover, activity in this region returned to normal levels when the patients were in remission. Overactivity of the inferior parietal cortex might reflect a heightened response to the sensory consequences of the movements the patients were making during scanning, contributing to the feeling that movements are externally controlled.

2.3 Phantom Limbs

After amputation of a limb many patients experience a phantom limb: they still feel the presence of the limb although they know it does not exist (Ramachandran & Rogers-Ramachandran 1996). It has been suggested that neural plasticity plays a role in the experience of phantom limbs. Some patients report being able to move their phantoms voluntarily, while others experience their phantom as paralyzed and cannot move it even with intense effort. If the limb was paralyzed before amputation the phantom normally remains paralyzed. If not, then typically immediately after amputation patients feel that they can generate movement in the phantom. However, with time they often lose this ability. It has been suggested that the estimated position of a limb is not based solely on sensory information, but also on the stream of motor commands issued to the limb muscles. On the basis of these commands the forward model can estimate the new position of the limb before any sensory feedback has been received. If these commands lead to the prediction of movement then the phantom will be experienced as moving. However, the motor control system is designed to adapt to changing circumstances. Since the limb does not actually move there is a discrepancy between the predicted and the actual consequences of the motor commands. With time the forward models will be modified to reduce these discrepancies – the prediction will be altered so that no movement of the limb is predicted even when motor commands to move the limb are issued. Such adaptation in the forward models could explain why patients eventually lose the ability to move their phantoms.

Adaptation of the forward models would also explain how Ramachandran and Rogers-Ramachandran (1996) were able to reinstate voluntary movement of the phantom by providing false visual feedback of a moving limb corresponding to the phantom. This was achieved by placing a mirror in the midsaggital plain. With the head in the appropriate position it was possible for the patient to see the intact limb at the same time as the mirror reflection of this limb. For most patients moving their hand in this mirror box rapidly leads to the perception that they are now able to move the phantom limb again. It has been suggested that the false visual feedback supplied by the mirror box allowed the forward models to be updated. In consequence, efference copy produced in parallel with the motor commands now generated changes in the predicted position of the missing limb corresponding to what the patient had seen in the mirror.

3 Conclusion

In this chapter, I have summarized how the forward model can be used to predict the sensory consequences of action and have suggested that this prediction might be available to awareness. Two examples of neurological symptoms that are characterized by changes in the prediction process are the experiences after damage to the parietal lobe and phantom limbs in amputees. A delusion of control or passivity symptom is a psychiatric symptom that may be characterized by an impaired prediction process.

References

Blakemore, S.-J., Wolpert, D.M., Frith, C.D.: Central cancellation of self-produced tickle sensation. Nature Neuroscience 1, 635–640 (1998)

Blakemore, S.-J., Frith, C.D., Wolpert, D.W.: Spatiotemporal prediction modulates the perception of self-produced stimuli. Journal of Cognitive Neuroscience 11, 551–559 (1999)

Blakemore, S.-J., Smith, J., Steel, R., Johnstone, E., Frith, C.D.: The perception of self-produced sensory stimuli in patients with auditory hallucinations and passivity experiences: Evidence for a breakdown in self-monitoring. Psychological Medicine 30, 1131–1139 (2000)

Blakemore, S.-J., Frith, C.D., Wolpert, D.W.: The cerebellum is involved in predicting the sensory consequences of action. Neuroreport 12(9), 1879–1885 (2001)

Daprati, E., Franck, N., Georgieff, N., Proust, J., Pacherie, E., Dalery, J., Jeannerod, M.: Looking for the agent: An investigation into consciousness of action and self-consciousness in schizophrenic patients. Cognition 65, 71–86 (1997)

Fourneret, P., Jeannerod, M.: Limited conscious monitoring of motor performance in normal subjects. Neuropsychologia 36, 1133–1140 (1998)

Frith, C.D., Blakemore, S.-J., Wolpert, D.M.: Abnormalities in the awareness and control of action. Philosophical Transactions of the Royal Society of London: Series B, Biological Sciences 355(1404), 1771–1778 (2000)

Ramachandran, V.S., Rogers-Ramachandran, D.: Synaesthesia in phantom limbs induced with mirrors. Proceedings of the Royal Society of London: B 263, 377–386 (1996)

Sirigu, A., Daprati, E., Pradat-Diehl, P., Franck, N., Jeannerod, M.: Perception of self-generated movement following left parietal lesion. Brain: A Journal of Neurology 122, 1867–1874 (1999)

Spence, S.A., Brooks, D.J., Hirsch, S.R., Liddle, P.F., Meehan, J., Grasby, P.M.: A PET study of voluntary movement in schizophrenic patients experiencing passivity phenomena (delusions of alien control). Brain: A Journal of Neurology 120(11), 1997–2011 (1997)

Weiskrantz, L., Elliot, J., Darlington, C.: Preliminary observations of tickling one-self. Nature 230, 598–599 (1971)

Wolpert, D.M., Goodbody, S.J., Husain, M.: Maintaining internal representations: The role of the human superior parietal lobe. Nature Neuroscience 1, 529–533 (1998)

9

Volition and the Function of Consciousness

Hakwan C. Lau

Columbia University
Psychology Department 355D Sch Ext
1190 Amsterdam Avenue MC: 5501
New York, NY 10027
hadwan@gmail.com

Summary. What are the psychological functions that could only be performed consciously? People have intuitively assumed that many acts of volition are not influenced by unconscious information. These acts range from simple examples such as making a spontaneous motor movement, to higher cognitive control. However, the available evidence suggests that under suitable conditions, unconscious information can influence these behaviors and the underlying neural mechanisms. One possibility is that stimuli that are consciously perceived tend to yield strong signals in the brain, which makes us think that consciousness has the function of such strong signals. However, if we could create conditions where the stimuli could yield strong signals but not the conscious experience of perception, perhaps we would find that such stimuli are just as effective in influencing volitional behavior. Future studies that focus on clarifying this issue may tell us what the defining functions of consciousness are.

Keywords: volition, intention, Libet, functions of consciousness.

1 Introduction

Many acts of volition seem to require conscious effort. We consciously initiate spontaneous motor movements. We cancel planned actions at will. We deliberately avoid particular actions. We intentionally shift our action plans in order to pursue different goals. Sometimes, theorists say, these are the functions of consciousness, as if evolution has equipped us with the gift of consciousness just to perform these acts. Without consciousness, presumably, we would only be able to perform much simpler actions that are no more sophisticated than embellished reflexes.

In this chapter we review available evidence to see if these intuitive claims are empirically supported. Recent studies in cognitive neuroscience suggest that many of these complex processes can actually be performed without consciousness. Or at least, many of them can be directly influenced by unconscious information. This calls into question what is the true function of consciousness, if not to enable us to deliberate our actions. We end by discussing what is logically required for an experiment to demonstrate the true function of consciousness.

2 Spontaneous Motor Initiation

Motor actions that are made not in immediate or direct response to external stimuli can be said to be spontaneously initiated. These are also sometimes called self-paced or self-generated actions. For instance, one may choose to casually flex one's wrist while sitting in a dark room, out of one's own free choice and timing, not to react to anything in particular. Some philosophers have argued that in cases like this, it should seem obvious that the action is caused by one's conscious intention (Searle 1983). Whereas one may argue that fast reactions to external stimuli may be driven by unconscious reflexes (e.g., a runner leaping forward upon hearing the starting whistle), spontaneous actions do not seem to have any immediate cause but the conscious intention itself.

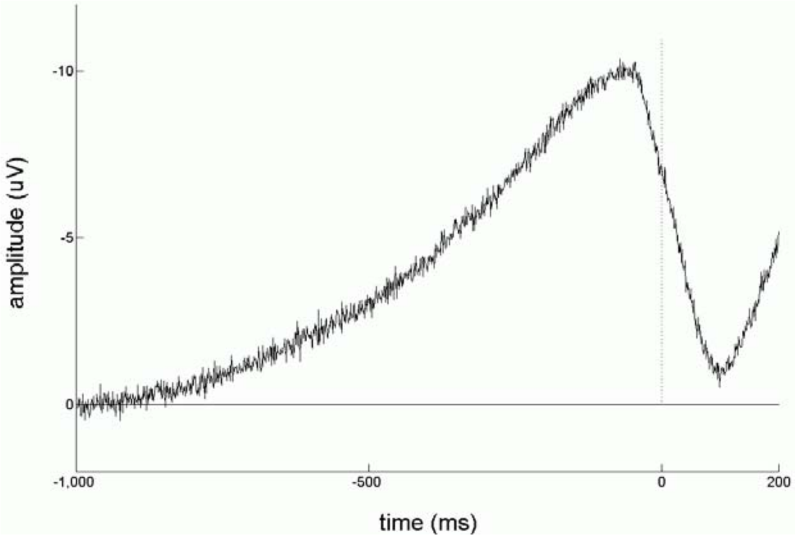

Fig. 9.1. A typical recording of the readiness potential (RP) preceding spontaneous movements. The RP is usually recorded at the top of the scalp, above medial frontal premotor areas. It gradually ramps up, beginning about 1–2 seconds before movement and peaking around the time of movement execution. Figure edited and adapted from Haggard and Eimer 1999.

However, it has been shown that there is preparatory activity in the brain that starts at as early as 1–2 seconds before spontaneous actions are executed. This aspect of one of the most perplexing findings in cognitive neuroscience was originally reported by Kornhuber and Deecke in the 1960s (Kornhuber & Deecke 1965). They placed electrodes on the scalp to measure electroencephalography (EEG) while subjects made spontaneous movements at their own timing. The EEG data that were time-locked to the point of motor execution (as measured by muscle contraction indicated by electromyography, EMG) were averaged over many trials, which produced an event-related potential (ERP) known as the *Bereitschaftspotential* (BP) or *readiness potential* (RP). The readiness potential slowly rises, peaking at around the point of action execution and starting from 1–2 seconds before that (fig. 9.1). The readiness potential is most pronounced at electrodes near the vertex (Cz in the EEG coordinate system), which is directly above the medial premotor areas (including the supplementary motor area, SMA, pre-supplementary motor area, pre-SMA, and the cingulate motor areas below them). It is generally believed that one major source of the readiness potential lies in the medial premotor areas (Ball et al. 1999; Erdler et al. 2000; Weilke et al. 2001; Cunnington et al. 2003).

The demonstration of the readiness potential behavior calls into question whether spontaneous movements are really caused by the preceding conscious intentions. Intuitively, conscious intentions seem to cause motor actions almost immediately – it seems to take much less time than 1–2 seconds. This could mean that the brain starts to prepare for the actions way before we consciously initiate them.

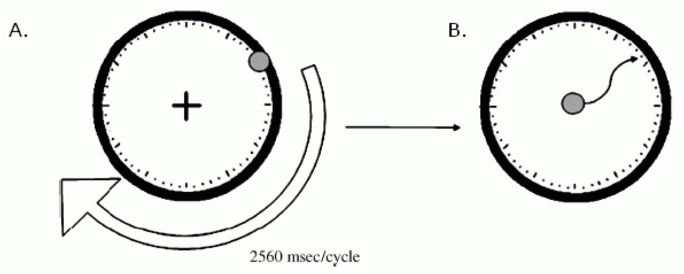

Fig. 9.2. The Libet clock paradigm
A. The subject views a dot rotating slowly (2.56 seconds per cycle) around a clock face and waits for an urge to move to arise spontaneously. When the urge arrives, the subject makes a movement (e.g., a key press).
B. After making the movement, the subject estimates the earliest time at which the intention to move was experienced. To carry out this time estimate, the subject moves the dot to the position on the clock face corresponding to the time when intention was first felt. In a common control condition, the subject uses the clock to estimate the time of movement rather than the onset of intention. Figure edited and adapted from Lau et al. 2007.

Benjamin Libet and colleagues empirically studied the timing of the conscious intention in relation to the readiness potential and the action (Libet et al. 1983a). To measure the onset of conscious intention, he invented a creative but controversial paradigm which is sometimes called the “Libet clock paradigm.” In those studies, subjects watched a dot revolving around a clock face at a speed of 2.56

second per cycle, while they flexed their wrist spontaneously (fig. 9.2). After the action was finished, subjects were required to report the location of the dot when they “first felt the urge” to produce the action, that is, the onset of intention. The subject might say it was at 3 o’clock or 4 o’clock position when they first felt the intention, for instance. This way the subjects could time and report the onset of their intention, and the experimenter could then work out when the action was actually produced, and hence the temporal difference between the two. Libet and colleagues reported that subjects on average report the onset of intention to be about 250 ms before motor execution.

Many people feel uncomfortable with the fact that the onset of the readiness potential seems to be so much earlier than the onset of intention, and some have tried to explain away the gap. Libet and colleagues have tried to study the onset of the readiness potential more carefully, discarding trials which might have been “contaminated” by pre-planning of action well before the action, as reported by the subjects. By only looking at the trials where the actions were supposed to be genuinely spontaneous, Libet and colleagues reported that the onset of the readiness potential is only about 500 ms before action execution (Libet et al. 1983a). However, this is still clearly earlier than the reported onset of intention. And by discarding so many trials, it may be that the analysis just lacked the power to detect an earlier onset.

Some have argued that the onset of the readiness potential might be an artifact due to the averaging needed to produce the ERP (Miller & Trevena 2002). However, Romo and Schultz (1987) have recorded from neurons in the medial premotor areas while monkeys made self-paced movements. It was found that these neurons in fact fired as early as 2.6 seconds before movement onset.

Others have argued that the readiness potential may not reflect the specific and causal aspects of motor initiation. However, as mentioned earlier, it is likely that the readiness potential largely originates from the medial premotor areas. Lesion to these areas can abolish the production of spontaneous actions (Thaler et al. 1995). These areas also contain neurons that code specific action plans (Shima & Tanji 1998; Tanji & Shima 1996). Further, when people use the Libet clock paradigm to time their own intentions, there is attentional modulation of activity in the medial pre-SMA (Lau et al. 2004), as if people were reading information off the area which is likely to be a source of the readiness potential.

The Libet clock method has also received considerable criticism. It involves timing across modalities, and could be susceptible to various biases (Libet 1985; Gomes 2002; Joordens et al. 2002; Klein 2002; Trevena & Miller 2002). However, it is unlikely that all these biases are in the direction that would help to narrow the gap between the onsets of the readiness potential and intention. Some have actually suggested that the different biases may point to different directions and thus just cancel each other out (Klein 2002). Also, in the original experiments by Libet and colleagues, there were control conditions that tested for the basic accuracy of the clock. They asked subjects to use the clock to time either the onset of movement execution, or in another condition to time the onset of tactile stimuli. Since the actual onsets of these events are objectively measurable, they could estimate the subjective error of onset reports produced by the clock method. They found the error to be in the order of about 50 ms – much smaller than the gap between the onsets of the readiness potential and intention.

The basic results of Libet and colleagues have also been replicated in several different laboratories (e.g., Lau et al. 2004; Haggard & Eimer 1999). In general, the same pattern is found, namely, that the onset of intention is either around or later than 250 ms before action execution; this seems to confirm our intuition that conscious intentions seem to be followed by motor actions almost immediately. In fact, given that the readiness potential could start as early as 1–2 seconds before action execution, it is hard to imagine how the onset of intention could coincide with or precede the readiness potential, unless one thinks of intention as a kind of prior intention (Searle 1983), such as the general plan that is formed at the beginning of the experimental session when the subject agrees to produce some actions in the next half an hour or so. We shall discuss this kind of higher-cognitive "intention" later in the chapter. However, the intention we are concerned with here is the immediate "urge" to produce the motor action (Libet et al. 1982).

Taken together, the evidence suggests that conscious intention, that is, the immediate feeling of motor initiation, is unlikely to be the "first unmoved mover" in triggering spontaneous motor movements. It is likely to be preceded by unconscious brain activity that may contribute to action initiation. What, then, is conscious intention for?

3 Conscious Veto?

Libet's interpretation of the timing-of-intention results is that although intention may not be early enough to be the first cause of action, the fact that it is before action execution means that it could still be part of the causal chain. Maybe the decision to move is initiated unconsciously, but the awareness of intention may allow us to "veto," that is, to cancel the action.

This seems to be a possibility. Libet and colleagues (Libet et al. 1983b) as well as other researchers (Brass & Haggard 2007) have performed experiments in which subjects prepare for an action and then cancel it in the last moment, just before it is executed. The fact that we have the ability to "veto" an action seems beyond doubt. The question, however, is whether having the conscious intention is critical. Can the choice of veto be preceded by unconscious activity, just as the intention to act is preceded by the readiness potential? Or maybe sometimes actions are unconsciously vetoed, even without our awareness?

Some recent evidence suggests that conscious intention may not facilitate a veto. As mentioned earlier, when people were using the Libet clock to time the onset of their intentions, there was attentional modulation of activity in the pre-SMA (Lau et al. 2004). These data have been subsequently further analyzed (Lau et al. 2006), and it has been shown that subjects who showed a large degree of attentional modulation tended to also report the onset of intention to be early. One interpretation could be that attention biases the judgment of onset to be earlier. It was found in another experiment that this was also true when people used the Libet clock to time the onset of the motor execution. The higher the level of fMRI activity modulated by attention, the earlier subjects reported the onset to be, even though on average subjects reported the onsets to be earlier than they actually were; this means a bias to the negative (i.e., early) direction produced more erroneous rather more precise reports. In general, the principle of attentional prior entry (Shore et al. 2001) suggests that attention to an event speeds up its perception and negatively biases the reported onset. If this were true in the case of the Libet experiments, this could mean that attention might have exaggerated the 250 ms onset; that is, had subjects not been required to attend to their intentions in order to perform the timing tasks, the true onset of conscious intention may well be much later than 250 ms prior to action execution. This calls into question whether we have enough time to consider the veto.

Another study reported that some patients with lesion to the parietal cortex reported the onset of intention to be as late as 50 ms prior to action execution (Sirigu et al. 2004). If the awareness of intention allows one to veto actions, one might expect these patients to have much less time to consciously evaluate spontaneous intentions and cancel the inappropriate ones. This could be quite disastrous to daily life functioning. Yet there were no such reports about these patients.

Finally, in another study (Lau et al. 2007), single pulses of transcranial magnetic stimulation (TMS) were sent to the medial premotor areas (targeting the pre-SMA). Again, subjects were instructed to produce spontaneous movements and to time the onset of intentions and movement execution using the Libet clock. Surprisingly, although TMS was applied *after* motor execution, it has an effect on the reported onsets. No matter whether TMS was applied immediately after action execution or with a 200 ms delay, the stimulation exaggerated the temporal distance between the reported onsets of intention and movement, as if people reported a prolonged period of conscious intending. One interpretation may be that TMS injected noisy activity into the area and the intention monitoring mechanism did not distinguish this from endogenously generated activity that is supposed to represent intention. However, what is crucial is the fact that the reported onsets can be manipulated even after the action is finished. This seems to suggest that our awareness of intention may be constructed after the fact, or at least not completely determined before the action is finished. If conscious intentions are not formed before the action, they certainly cannot play any role in facilitating veto, let alone causing it.

This interpretation may seem wild, but it is consistent with other proposals. For instance, Wegner (2002) has suggested that maybe conscious will is an illusion. The sense of agency is often inferred post hoc, based on many contextual factors. Wegner cites experiments to support these claims. One example is a study on "facilitated communication" (Wegner, Fuller & Sparrow 2003). Subjects (playing the role of "facilitators") were asked to place their fingers on two keys of a keyboard, while a confederate (playing the role of "communicator") placed his or her fingers on top of those of the subject. Subjects were given headphones with which they listened to questions of varying difficulty. Confederates were given headphones as well, and subjects were led to believe that the confederates would be hearing the same questions, although in fact the confederates heard nothing. Subjects were told to detect subtle, unconscious movements in the confederate's fingers following each question. When such movements were detected, the subject should press the corresponding key in order to answer on the confederate's behalf. It was found that subjects answered easy questions well above chance levels. If they had performed the task strictly according to the instructions, however, they should have performed at chance. Therefore, subjects must have been directing their own key presses. Nonetheless, they attributed a significant causal role for the key presses to the confederate. The degree to which subjects answered easy questions correctly was not correlated with the degree to which they attributed causal responsibility to confederates, suggesting that the generation of action and attribution of action to an agent are independent processes.

To summarize, although theorists have speculated that the awareness of intention may play some role in allowing us to cancel or edit our actions, considerable doubt has been cast on this by recent empirical evidence.

4 Exclusion and Inhibition

Another kind of situation that seems to require conscious deliberation involves the need to avoid a particular action or response. This is related to "vetoing" as described above, except that the action being inhibited is not necessarily self-paced, and may be specified externally. One example would be to perform stem completion while avoiding a particular word. So for instance, the experimenter may ask the subjects to produce any word starting with letter D (i.e., completing a "stem"), but avoid the word "dinner." So subjects can produce "dog," "danger," "dear," etc., but if they produce the word "dinner," it would be counted as an error. This is called the exclusion task (Jacoby et al. 1992).

One interesting aspect of the exclusion task is that people can perform well only if they clearly perceive and remember the target of exclusion (i.e., the word "dinner" in the foregoing example). If the target of exclusion is presented very briefly and followed by a mask, such that it was only very weakly perceived, people may fail to exclude it (Debner & Jacoby 1994; Merikle et al. 1995). In fact, they tend to produce exactly the word they should be avoiding with higher likelihood than if they were not presented with the word at all. It has been argued that this exclusion failure phenomenon is the hallmark of unconscious processing (Jacoby et al. 1992). The weak perception of the target probably produced a representation for the word, but because the signal is not strong enough to reach the level of conscious processing, we are unable to inhibit the corresponding response.

In addition to the intuitive appeal, the notion that consciousness is required for exclusion is also supported by a case study of a blindsight patient (Persaud & Cowey 2008). Subject GY has a lesion to the left primary visual cortex (V1), and reports that most of his right visual field is subjectively blind. However, in forced-choice situation he can discriminate simple stimuli well above chance level in his "blind" field (Weiskrantz 1986, 1997). In one study he was required to perform an exclusion task (Persaud & Cowey 2008), that is, to say the location (up or down) where the target was *not* presented. Whereas he could do this easily in the normal field, he failed the task when stimuli were presented to his blind field. Note that he was significantly worse than chance in the blind field, as if the unconscious signal drove the response directly and inflexibly, defying exclusion control. This seems to support the account that consciousness is required for exclusion.

The general idea that inhibition requires consciousness seems to be supported by other studies too, including those that do not employ the exclusion paradigm. One study tested the subjects' ability to ignore distracting moving dots, while doing a central task that has nothing to do with the distractors (Tsushima et al. 2006). It was found that if the motion of the distractor was above the perceptual threshold, people could ignore the dots and inhibit the distraction successfully. Somewhat paradoxically, when the motion was below the perceptual threshold, people could not ignore the dots and were distracted. The results from brain imaging seem to suggest that when the motion of the stimuli was strong, it activated the prefrontal cortex, and triggered it to suppress the motion signal. When the motion of the stimuli was below the perceptual threshold, however, the signal failed to trigger the inhibitory functions in the prefrontal cortex, and therefore the motion signals were not suppressed and thus remained distracting.

However, the notion that flexible control or inhibition of perceptual signals requires consciousness is not without its critics (Snodgrass 2002; Haase & Fisk 2001; Visser & Merikle 1999). One problem becomes clear when we consider the motion distractor example above. "Conscious signal" here seems to be the same thing as a strong signal, driven by larger motion strength in the stimuli. Obviously, signals have to be strong enough to reach the prefrontal cortex in order to trigger the associating executions functions. Do unconscious stimuli fail to be excluded because we are not conscious of them, or is it just because the signal is not strong enough? Or, are the two explanations one and the same? We will come back to this issue in the final section of the chapter.

Other researchers have reported evidence that seems to support unconscious inhibition. For instance, in one study (Snodgrass & Shevrin 2006) people were asked to detect visually presented words. In certain conditions, some subjects showed detection performance that was significantly *worse* than chance. These words were presented so briefly that typically detection performance would be near chance. We usually take chance-level as the objective threshold for conscious perception. Below chance-level performance could be taken as evidence that the subjects did not consciously perceive the words. And yet, if they had no information at all regarding the words, performance should be exactly at chance rather than below. It seems that these subjects were actively suppressing the words.

These are unusual cases and are somewhat hard to interpret. We take chance-level as the objective threshold for conscious perception because when people perform at chance, it indicates that they do not have the explicit information regarding the target of perception. However, if people perform significantly below chance, it means that somehow they have the information regarding the detection, which violates the very logic we adopt to label perception unconscious. But in any case, the stimuli were supposed to be really weak, and it is intriguing that some subjects seem to be automatically suppressing the words. Are we to take these somewhat unusual cases as evidence for rejecting the notion that exclusion or inhibition requires consciousness? It seems that, logically, if we claim that a certain function *requires* consciousness, we should predict there will never be a case where one could perform such function unconsciously. How seriously are we to take this logic? Should we reject functions as requiring consciousness on the basis of a single experiment? We will return to this argument in the last section of the chapter.

5 Top-Down Cognitive Control

So far we have discussed acts of volition that are relatively simple, like starting a motor movement, or avoiding a particular action. Sometimes we also voluntarily prepare for a set of rules or action plans in order to satisfy a more abstract goal in mind. For instance, a telephone ring may usually trigger a particular action, for example, to pick up the phone. However, when one visits friends at their homes, one may deliberately change the mapping between the stimulus (telephone ring) and action: that is, it would be more appropriate to sit still, or ask the host to pick up the phone, rather than picking it up yourself. This volitional change of stimulus-response contingency is an example of top-down cognitive control.

It has been suggested that top-down cognitive control may require consciousness (Dehaene & Naccache 2001). The idea is that unconscious stimuli can trigger certain prepared actions, as demonstrated in studies in subliminal priming (Kouider & Dehaene 2007). However, the preparation or setting up of the stimulus-response contingency may require consciousness.

However, recent studies suggest that this might not be true, in the sense that unconscious information seems to be able to influence or even trigger top-down cognitive control too (Mattler 2003; Lau & Passingham 2007). In one study subjects had to prepare to make a phonological or semantic judgment, based on the orientation of a figure they saw (fig. 9.3). In every trial, if they saw a square, they had to prepare to judge whether an upcoming word has two syllables (e.g., “table”) or not (e.g., “milk”). If they saw a diamond, they had to prepare to judge whether an upcoming word refers to a concrete object (e.g., “chair”) or an abstract idea (e.g., “love”). In other words, they had to perform top-down cognitive control based on the instruction figure (square or diamond). However, before the instruction figure was presented, there was actually an invisible prime figure, which could also be a diamond or a square. It was found that the prime could impair subjects’ performance when it suggested the alternative (i.e., wrong) task to the subjects. One could argue that this was only because the prime distracted the subjects on a perceptual level, and did not really trigger cognitive control. However, the experiment was performed in the fMRI scanner, and the brain recordings suggest that when being primed to perform the wrong task, subjects used more of the wrong neural resources too (Lau & Passingham 2007). That is, areas that are more sensitive to phonological or semantic processing showed increased activity when the explicit instruction figure made subjects perform the phonological and semantic tasks respectively. The invisible primes also seem to be able to trigger activations in task-sensitive areas. This seems to suggest that they can influence or exercise top-down cognitive control.

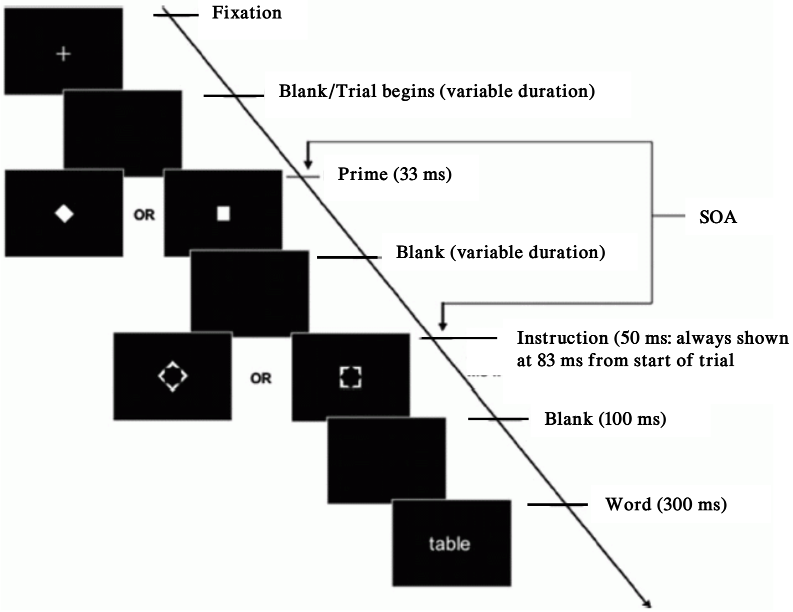
Fixation
+
Blank/Trial begins (variable duration)
Prime (33 ms)
OR
SOA
Blank (variable duration)
Instruction (50 ms: always shown at 83 ms from start of trial
OR
Blank (100 ms)
Word (300 ms)
table

Fig. 9.3. Experimental paradigm of Lau and Passingham (2007). Subjects view briefly presented words and perform either a phonological task (is the word one syllable or two syllables?) or a semantic task (does the word name something concrete or abstract?). Before word presentation, subjects are instructed which task to perform on a given trial by a visual symbol (a square for the phonological task, or a diamond for the semantic task). The symbolic instruction itself acts as a metacontrast mask for an earlier prime, also a square or a diamond. Because the prime is briefly presented and masked, it is not consciously perceived. On half of trials, the prime is congruent with the instruction, and on the other half, incongruent. Behavioral and imaging results suggest that the unconscious primes affected top-down task switching. When primes were incongruent with instructions, accuracy fell, reaction time increased, and brain regions corresponding to the task indicated by the prime were partially activated (all relative to the prime-congruent condition). But when the stimulus onset asynchrony (SOA) between prime and instruction was lowered, such that primes became visible, the priming effect was not evident. This double dissociation suggests that the interference of incongruent primes on task switching cannot be attributed to conscious processing. Figure adapted from Lau and Passingham 2007.

Another study examines how unconscious information affects our high-level objectives by focusing on how the potential reward influences our level of motivation (Pessiglione et al. 2007). Subjects squeezed a device to win a certain amount of money. The harder they squeezed, the more money they would win. However, the size of the stake in question for a particular trial was announced in the beginning by presenting the photo of a coin. The coin could either be a British pound (~2 U.S. dollars) or a penny (~2 U.S. cents), and it signified the monetary value of the maximal reward for that trial. Not surprisingly, people squeezed harder when the stakes were high, but interestingly, the same pattern of behavior was observed even when the figure of the coin was masked such that subjects reported not seeing it. This suggests that unconscious information can influence our level of motivation as well.

If unconscious information alone is sufficient to exercise all these sophisticated top-down control functions, why do we need to be conscious at all?

6 How to Find the True Function of Consciousness?

The foregoing is not meant to be an exhaustive review of all studies on the potential functions of consciousness. We select some examples from a few areas that are particularly related to volition, and discuss what role consciousness may play. It may, of course, be that there are other psychological functions that require consciousness.

Yet, one cannot help but feel that there is some inherent limitation to this whole enterprise of research. If we claim that a certain function requires consciousness, we are making the claim that the function should never be able to be performed unconsciously. In principle, it would only take a single experiment to falsify this claim. This explains why this review may seem biased, in that we focus on studies that show the power of the unconscious, rather than studies demonstrating what functions definitely require consciousness. In principle, falsifying the claim that a certain function requires consciousness is straightforward. But this is not the case for demonstrating functions that do require consciousness.

One can of course try to show that subjects could normally do a task if the relevant information is consciously perceived. And then one tries to "knock-out" the conscious perception for such information, and try to show that the task could no longer be performed. But how would one know that in "knocking-out" the conscious perception, one does not "knock-out" too much? One typically suppresses conscious perception by visual masking, by using brief presentation, by distracting the subject, by applying transcranial magnetic stimulation, by pharmacological manipulations, etc. But all of these could potentially impair the unconscious as well as the conscious signal. Perhaps in cases where the perception has been rendered unconscious, the signal is just no longer strong enough to drive the function in question? This would mean that, in principle, it would be possible for a future study to find the optimal procedure or setup to just render the information unconscious, without reducing the signal strength too much. And in that case the subjects may be able to perform the task in question. This would falsify our claim.

This means that in looking for functions that require consciousness, we need to adopt some different strategies. One potentially useful approach is to try to demonstrate something akin to a "double dissociation." When conscious perception is suppressed, we often find that a sophisticated function (e.g., top-down cognitive control) can no longer be performed, though some simpler function (e.g., priming for a prepared motor response) may still be activated by unconscious information. From the foregoing discussion, one could see that this may not be as surprising or informative as it seems. It could be merely that the unconscious signal is too weak to drive the relatively sophisticated function. A demonstration of the opposite would, however, be much more convincing: If after suppression of conscious perception, the subjects can still perform a rather sophisticated function, but fail to perform a simple function, this would suggest that the simple function really requires consciousness. In this case, it could not be that the suppression of conscious perception has taken away too much of the signal strength, because if that were the case then the subjects should not be able to perform the relatively sophisticated function (fig. 9.4).

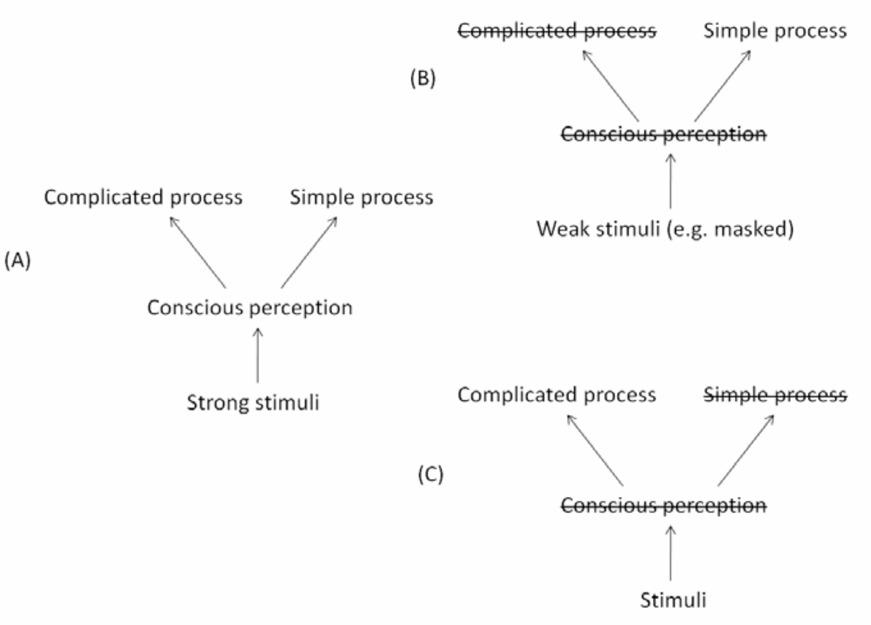

Fig. 9.4. (a) The normal situation for conscious perception. Stimuli are strong enough to drive processes of different complexity. (b) A typical situation for unconscious perception. Stimuli are weak such that complicated processes are no longer activated, though simple processes can still be triggered. It could be argued that this is not surprising as we may expect that complicated processes require a stronger signal. (c) A potentially more informative situation. If one could find a stimulus that is not consciously perceived, but yet is sufficiently strong to trigger a complicated process, then the relatively simple process that the stimulus does not drive would seem to critically depend on consciousness.

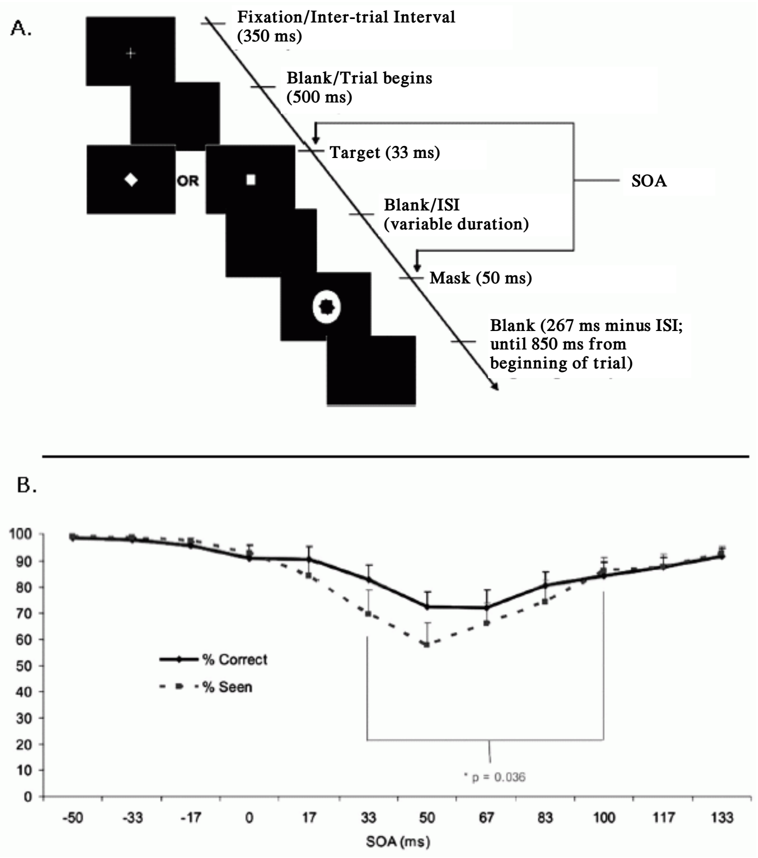

Fig. 9.5. Inducing "relative blindsight" in normal observers using metacontrast masking.

A. Metacontrast masking paradigm. The subject is presented with a visual target (in this case, either a square or diamond). Afterwards, a metacontrast mask is presented. The mask differentially affects discrimination accuracy and visual awareness of the target as a function of stimulus onset asynchrony (SOA).

B. Discrimination accuracy and visual awareness as a function of metacontrast mask SOA. The metacontrast mask creates a characteristic U-shaped function of performance vs. SOA. At shorter and longer SOAs, discrimination accuracy is high, but it dips at intermediate SOAs. The same is true for visual awareness, but the shape of the awareness masking function is not perfectly symmetrical with respect to the performance masking function. That is, there are certain SOAs at which forced choice performance is matched, but visual awareness differs significantly (e.g., as illustrated in the SOAs of 33 ms and 100 ms in fig. 9.5B). Such performance-matched conditions could be used to investigate the functions of consciousness. If some task can be performed better in the condition of higher subjective visibility, it can plausibly be said to require visual awareness. Because forced-choice discrimination accuracy is matched across the two conditions, the superior performance of the task in the high visibility condition cannot be attributed to a difference in signal strength. Figure adapted from Lau and Passingham 2006.

An alternative approach may be to directly match for signal strength between the conscious and the unconscious conditions. This might seem quite difficult because conscious signals may seem to be strong in general. However, as discussed above, blindsight subjects can perform forced-choice discrimination on visual stimuli well above chance, even when they claim that conscious awareness is missing. Forced-choice performance is often taken as an objective estimate of signal strength; the detection theoretical measure d' is mathematically just the signal-to-noise ratio. In blindsight subject GY, for whom only half of the visual field lacks awareness, we can imagine presenting weak stimuli to the normal visual field such that forced-choice performance would match that in the blind field (Weiskrantz et al. 1995). This way we can test if certain functions cannot be performed based on information presented to the blind field, which may shed light on when consciousness is required.

One may argue that blindsight patients are rare and the way their brains process visual information may not be generalizable to intact brains. However, there are other paradigms whereby in normal subjects one could match for forced-choice performance and yet produce a difference in the level of conscious awareness. For instance, one study (Lau & Passingham 2006) used metacontrast masking to create similar conditions where forced-choice discrimination accuracy for the visual targets were matched, and yet the subjective reports of how often subjects saw the identity of the targets differed (fig. 9.5). One could imagine presenting these stimuli to subjects and seeing if they drive a certain function with different levels of effectiveness. If the subjects perform better in the condition where subjective conscious awareness of the stimuli is more frequent, one could argue that this function is likely to depend critically on consciousness.

7 Conclusion

Acts of volition are accompanied by a sense of conscious effort or intention. The fact that we feel the conscious effort is not in doubt. What is less clear is whether the processes underlying the conscious experience directly contribute to the execution of the actions, in a way that is not accomplished by unconscious processes just as effectively. The general picture seems to be that many sophisticated functions can be performed unconsciously or driven by unconscious information.

Does this mean that consciousness has no special function at all? The answer is not yet clear. It is likely that some psychological functions do require consciousness – that is, can never be performed unconsciously – but experiments have not yet been able to convincingly pin them down.

Experimenters will have to overcome the following problem. If we assume that conscious perception is always accompanied by stronger and longer-lasting signals that are more effective than unconscious signals in propagating themselves throughout the brain, then certainly, consciousness would have the functions of these strong signals. However in studies of blindsight (Weiskrantz et al. 1995) as well as in normals (Lau & Passingham 2006) it has been shown that signal strength as indicated by forced-choice performance is not always one and the same as conscious awareness. Therefore, future studies may need to focus on identifying the functions that really cannot be performed unconsciously, even when the signal strength is sufficiently strong. This may help to reveal the true function of consciousness.

References

Ball, T., Schreiber, A., Feige, B., Wagner, M., Lücking, C.H., Kristeva-Feige, R.: The role of higher-order motor areas in voluntary movement as revealed by high-resolution EEG and fMRI. NeuroImage 10(6), 682–694 (1999)

Brass, M., Haggard, P.: To do or not to do: The neural signature of self-control. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience 27(34), 9141–9145 (2007)

Cunnington, R., Windischberger, C., Deecke, L., Moser, E.: The preparation and readiness for voluntary movement: A high-field event-related fMRI study of the Bereitschafts-BOLD response. NeuroImage 20(1), 404–412 (2003)

Debner, J.A., Jacoby, L.L.: Unconscious perception: Attention, awareness, and control. Journal of Experimental Psychology: Learning, Memory, and Cognition 20(2), 304–317 (1994)

Dehaene, S., Naccache, L.: Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. Cognition 79(1–2), 1–37 (2001)

Erdler, M., Beisteiner, R., Mayer, D., Kaindl, T., Edward, V., Windischberger, C., et al.: Supplementary motor area activation preceding voluntary movement is detectable with a whole-scalp magnetoencephalography system. NeuroImage 11(6), 697–707 (2000)

Gomes, G.: The interpretation of Libet's results on the timing of conscious events: A commentary. Consciousness and Cognition 11(2), 221–230 (2002); discussion 308–313, 314–325

Haase, S.J., Fisk, G.: Confidence in word detection predicts word identification: Implications for an unconscious perception paradigm. The American Journal of Psychology 114(3), 439–468 (2001)

Haggard, P., Eimer, M.: On the relation between brain potentials and the awareness of voluntary movements. Experimental brain research. Experimentelle Hirnforschung. Expèrimentation cérébrale 126(1), 128–133 (1999)

Jacoby, L.L., Lindsay, D.S., Toth, J.P.: Unconscious influences revealed: Attention, awareness, and control. The American Psychologist 47(6), 802–809 (1992)

Joordens, S., van Duijn, M., Spalek, T.M.: When timing the mind one should also mind the timing: Biases in the measurement of voluntary actions. Consciousness and Cognition 11(2), 231–240 (2002); discussion 308–313

Klein, S.: Libet's research on the timing of conscious intention to act: A commentary. Consciousness and Cognition 11(2), 273–279 (2002); discussion 304–325

Kornhuber, H., Deecke, L.: Hirnpotentialänderungen bei Willkurbewegungen und passiven Bewegungen des Menschen: Bereitschaftspotential und reafferente Potentiale. Pflügers Archive 284, 1–17 (1965)

Kouider, S., Dehaene, S.: Levels of processing during non-conscious perception: A critical review of visual masking. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 362(1481), 857–875 (2007)

Lau, H.C., Passingham, R.E.: Relative blindsight in normal observers and the neural correlate of visual consciousness. Proceedings of the National Academy of Sciences of the United States of America 103(49), 18763–18768 (2006)

Lau, H.C., Passingham, R.E.: Unconscious activation of the cognitive control system in the human prefrontal cortex. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience 27(21), 5805–5811 (2007)

Lau, H.C., Rogers, R.D., Haggard, P., Passingham, R.E.: Attention to intention. Science 303(5661), 1208–1210 (2004)

Lau, H.C., Rogers, R.D., Passingham, R.E.: On measuring the perceived onsets of spontaneous actions. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience 26(27), 7265–7271 (2006)

Lau, H.C., Rogers, R.D., Passingham, R.E.: Manipulating the experienced onset of intention after action execution. Journal of Cognitive Neuroscience 19(1), 81–90 (2007)

Libet, B.: Unconscious cerebral initiative and the role of conscious will in voluntary action. Behavioral and Brain Sciences 8, 529–566 (1985)

Libet, B., Wright, E.W., Gleason, C.A.: Readiness-potentials preceding unrestricted 'spontaneous' vs. pre-planned voluntary acts. Electroencephalography and Clinical Neurophysiology 54(3), 322–335 (1982)

Libet, B., Gleason, C.A., Wright, E.W., Pearl, D.K.: Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. Brain: A Journal of Neurology 106(Pt. 3), 623–642 (1983a)

Libet, B., Wright, E.W., Gleason, C.A.: Preparation- or intention-to-act, in relation to pre-event potentials recorded at the vertex. Electroencephalography and Clinical Neurophysiology 56(4), 367–372 (1983b)

Mattler, U.: Priming of mental operations by masked stimuli. Perception & Psychophysics 65(2), 167–187 (2003)

Merikle, P.M., Joordens, S., Stolz, J.A.: Measuring the relative magnitude of unconscious influences. Consciousness and Cognition 4(4), 422–439 (1995)

Miller, J., Trevena, J.A.: Cortical movement preparation and conscious decisions: Averaging artifacts and timing biases. Consciousness and Cognition 11(2), 308–313 (2002)

Persaud, N., Cowey, A.: Blindsight is unlike normal conscious vision: Evidence from an exclusion task. Consciousness and Cognition 17(3), 1050–1055 (2008)

Pessiglione, M., Schmidt, L., Draganski, B., Kalisch, R., Lau, H., Dolan, R.J., et al.: How the brain translates money into force: A neuroimaging study of subliminal motivation. Science 316(5826), 904–906 (2007)

Romo, R., Schultz, W.: Neuronal activity preceding self-initiated or externally timed arm movements in area 6 of monkey cortex. Experimental Brain Research. Experimentelle Hirnforschung. Expérimentation cérébrale 67(3), 656–662 (1987)

Searle, J.R.: Intentionality: An essay in the philosophy of mind. Cambridge University Press, Cambridge (1983)

Shima, K., Tanji, J.: Both supplementary and presupplementary motor areas are crucial for the temporal organization of multiple movements. Journal of Neurophysiology 80(6), 3247–3260 (1998)

Shore, D.I., Spence, C., Klein, R.M.: Visual prior entry. Psychological Science: A Journal of the American Psychological Society/APS 12(3), 205–212 (2001)

Sirigu, A., Daprati, E., Ciancia, S., Giraux, P., Nighoghossian, N., Posada, A., et al.: Altered awareness of voluntary action after damage to the parietal cortex. Nature Neuroscience 7(1), 80–84 (2004)

Snodgrass, M.: Disambiguating conscious and unconscious influences: Do exclusion paradigms demonstrate unconscious perception? The American Journal of Psychology 115(4), 545–579 (2002)

Snodgrass, M., Shevrin, H.: Unconscious inhibition and facilitation at the objective detection threshold: replicable and qualitatively different unconscious perceptual effects. Cognition 101(1), 43–79 (2006)

Tanji, J., Shima, K.: Supplementary motor cortex in organization of movement. European Neurology 36(suppl. 1), 13–19 (1996)

Thaler, D., Chen, Y.C., Nixon, P.D., Stern, C.E., Passingham, R.E.: The functions of the medial premotor cortex, I: Simple learned movements. Experimental Brain Research. Experimentelle Hirnforschung. Expérimentation cérébrale 102(3), 445–460 (1995)

Trevena, J.A., Miller, J.: Cortical movement preparation before and after a conscious decision to move. Consciousness and Cognition 11(2), 162–190 (2002) (discussion 314–325)

Tsushima, Y., Sasaki, Y., Watanabe, T.: Greater disruption due to failure of inhibitory control on an ambiguous distractor. Science 314(5806), 1786–1788 (2006)

Visser, T.A., Merikle, P.M.: Conscious and unconscious processes: The effects of motivation. Consciousness and Cognition 8(1), 94–113 (1999)

Wegner, D.M.: The illusion of conscious will. MIT Press, Cambridge (2002)

Wegner, D.M., Fuller, V.A., Sparrow, B.: Clever hands: Uncontrolled intelligence in facilitated communication. Journal of Personality and Social Psychology 85(1), 5–19 (2003)

Weilke, F., Spiegel, S., Boecker, H., von Einsiedel, H.G., Conrad, B., Schwaiger, M., et al.: Time-resolved fMRI of activation patterns in M1 and SMA during complex voluntary movement. Journal of Neurophysiology 85(5), 1858–1863 (2001)

Weiskrantz, L.: Blindsight: A case study and implications. Oxford University Press, Oxford (1986)

Weiskrantz, L.: Consciousness lost and found: A neuropsychological exploration. Oxford University Press, Oxford (1997)

Weiskrantz, L., Barbur, J.L., Sahraie, A.: Parameters affecting conscious versus unconscious visual discrimination with damage to the visual cortex (V1). Proceedings of the National Academy of Sciences of the United States of America 92(13), 6122–6126 (1995)

Part III: Broader Understandings of Volition and Consciousness

10

Conscious Willing and the Emerging Sciences of Brain and Behavior

Timothy O’Connor

Department of Philosophy
Indiana University
Sycamore 026
Bloomington, IN 47405
toconnor@indiana.edu

Summary. Recent studies within neuroscience and cognitive psychology have explored the place of conscious willing in the generation of purposive action. Some have argued that certain findings indicate that the commonsensical view that we control many of our actions through conscious willing is largely or wholly illusory. I rebut such arguments, contending that they typically rest on a conflation of distinct phenomena. Nevertheless, I also suggest that traditional philosophical accounts of the will need to be revised: a raft of studies indicate that control over one’s own will among human beings is limited, fragile, and – insofar as control depends to an extent on conscious knowledge – admitting of degrees. I briefly sketch several dimensions along which freedom of the will may vary over time and across agents.

Keywords: Free will, choice, consciousness, self-awareness, intention, willing, reasons, automaticity, Libet.

There is a wide panoply of philosophical views concerning the role of the will in human freedom of choice and action. Philosophers aim to give schematic accounts of the conditions that a person’s willing and carrying out a course of action must satisfy to be appropriately accounted as free and subject to moral praise or blame. For my purpose here, it is enough to note only the main families of views, and these only briefly. Since at least the seventeenth century, it has been common to divide philosophical accounts of freedom in terms of whether or not they regard freedom as compatible with causal determinism in the natural order. On “compatibilist” views, choice is a *mechanism* by which an agent’s predominant desires and intentions, along with corresponding beliefs concerning how they may be realized, ineluctably initiate appropriately matching actions. On most compatibilist views, what distinguishes will or choice from other natural mechanisms is that it is, in John Fischer’s term, *reasons-responsive*, a mechanism that is activated by the agent’s own goals and desires. Some will add further requirements to this unadorned picture. For example, Harry Frankfurt and others maintain that a will that is free must have a hierarchical structure, such that the first-order desires that cause one’s choices are ratified by a higher-order willing that one be moved by the first-order desire – a willing that one be the sort of person who is regularly moved to act by that sort of motivation. (Addicts might be people who typically do *not* will to be moved by the desires that in fact move them towards satisfying their addiction, and thus are judged not to be free in these willings.) On all compatibilist views, simple or adorned, the capacity to form choices or acts of will is embedded in a seamless causal flow of events constituting the universe’s history. (Any causal indeterminacy in the decision-making process is at best irrelevant to what makes for freedom and at worst potentially undermining.)

The development and refinement of compatibilist views continues apace. But the last couple of decades has also seen a resurgence of theorizing about the will along “incompatibilist” lines. All incompatibilists who deem free will to be possible maintain that choice is some sort or other of indeterministic capacity by which I “directly” determine my choice while guided by reasons. This power of direct determination, or choice, is then elucidated in different ways: first, as a nondeterministic variant of the compatibilist account on which it is the power of one’s beliefs, desires, and prior intentions to cause one’s choices; second, as a sui generis noncausal variety of active power; or third, as a sui generis agent-causal power (a primitive capacity of a person to form an executive intention to act in a specific way). The way reasons are thought to guide the exercise of such a power of choice is either efficiently causal (albeit nondeterministic) in nature, or it is purely teleological, or it is some hybrid of these.

While we see a wide range of views, we should not lose sight of the fact that most of them are trying to articulate the very same features of freedom that seem compelling from a subjective, pretheoretical point of view. We may mark this underlying commonality by observing that, on all these views, the freedom and responsibility of human beings requires our having and regularly exercising three very general capacities: (1) an awareness of and sensitivity to reasons (including especially moral reasons) for actions; (2) the ability to weigh and even to critically probe our own desires and intentions and to reevaluate on occasion our overarching goals; and (3) the ability to choose, based on reasons, which action we shall undertake on a given occasion.

Perhaps the most striking feature in philosophical discussion of free will in very recent years is the rise of free will *skeptics*, who wish to challenge all the accounts indicated above. This skeptical view strikes at the core datum of traditional accounts of the will by maintaining that the experience of conscious will is illusory: we do not freely control *any* of our actions via the conscious exertion of will, however this may be characterized. (Their more measured cousins, the *revisionists*, recommend that we drastically revise the ordinary understanding of free will so as to make it suited to the likes of you and me, who [according to the psalmist] were created a little – say the revisionists, a *lot* – lower than the angels.) An older skeptical view (one that continues to this day) was rooted in a judgment that the very idea of freedom of the will resists coherent articulation. According to the new-wave skeptics, however, the problem is thought to be empirical: the efficacy of conscious will indicated by our subjective experience of agency runs counter to mounting evidence from the sciences that in one way or another touch upon volition and the genesis of human action. Such a conclusion has been endorsed by some of the relevant scientists themselves and thereby has begun to enter the popular imagination.

In what follows, I will offer reasons for resisting this empirical argument for the skeptical view. I will first try to show that there are conceptual confusions underlying some of the planks in the skeptic's case. I will then make a start at confronting the challenge that remains, once the confusions are identified and set aside.

1 The Sciences of Volition and Agency: Empirical Challenges to Free Will on Three Fronts

I will begin by summarizing some of the main empirical findings of relevance to the existence and nature of free will in human beings. These come from three broad research areas: neuroscience, the study of clinical mental disorders, and social psychology.

Neuroscience

Over the past three decades, neuroscience has delivered a variety of fascinating and sometimes surprising results concerning human action. Here I will consider (1) cases where actions are artificially produced but give rise to *unwitting confabulations by agents concerning their own agency*, (2) cases where *unperceived environmental stimulation of the brain can significantly influence seemingly free choices*, and lastly (3) the much-discussed studies concerning the *timing of preparatory brain activity vis-à-vis the experience of agency*.

In the confabulation cases, neurosurgeons operating on the brains of conscious patients were able to induce behaviors such as the raising of a hand by electrically stimulating motor-control areas of the brain (Delgado 1969; Gazzaniga 1994). We have very good reason to believe that the patients did not form a conscious choice to move their arm. Yet when asked immediately afterwards why they moved their arms, they tended to confabulate rational explanations – apparently sincerely – such as "I was trying to get your attention."

In the external brain stimulation studies, subjects were asked to freely choose between moving either the left or right index finger when signaled. While they awaited the signal, a large magnet was secretly moved across the motor area of either the left or right side of the brain. It turns out that subjects showed an increased tendency to move the finger contralateral to the side that was stimulated, while believing that that they were choosing voluntarily and without any discernible external influence (Brasil-Neto et al. 1992).

Finally, and most famously, there are Benjamin Libet's (1985) studies on the timing of the experience of willing (studies that have since been refined by Patrick Haggard, H.C. Lau, and others). Libet devised a study in which people are asked to wiggle their finger within a short interval of time (thirty seconds or so). The experimenter instructs them to do so whenever they wish – though spontaneously, not by deciding the moment in advance. Throughout, they are to watch a special clock with a very fast-moving dial (a beam of light) and note its location at the precise moment at which they felt the "urge" or "wish" to move the finger. During the experiment, a device measures electrical activity on the agent's scalp. Libet discovered that a steady increase in this activity (dubbed the "readiness potential," or RP) consistently preceded the time the agents cited as when they experienced the will to move. By averaging results over hundreds of experiments, Libet determined that the RP preceded the "experience of will" by an average of some 400 milliseconds, a significant interval in the context of neural activity. Libet and others concluded from this result that "conscious will" is not the initiator of voluntary acting but instead a *consequence* of an unconscious physical process that also (and according to some hypotheses, independently) triggers the action.

Clinical Mental Disorders

I turn next to clinical medical disorders. There are quite a number of abnormal clinical phenomena, but two deserve special attention in considering the relation of the will to purposive action. In anarchic (or "alien") hand syndrome, complex movements of the person's hand are carried out in a smooth way towards the execution of an apparent goal (taking someone else's glass of water, say), yet the person claims not to have intended them, and they are generally unwanted, causing embarrassment. (Many of us think here of Peter Sellers' title character in Stanley Kubrick's film *Dr. Strangelove*.) And some schizophrenic patients report the perception that other agents are controlling their actions, even though there is good reason to believe that the actions in fact issue from intentions of the agent, albeit unconscious ones. In both kinds of cases, agents are acting purposively and under no discernible control by an external agent or source, yet they lack, in some sense, the feeling of being the authors of their actions. Some free will skeptics draw attention to these cases because the "sense of freedom," or the sense of being in control of what one does, is arguably a fundamental basis of our belief that we are in fact free to choose, in some measure. And their thought is that these and other cases make plausible the idea that the sense of agency is not in fact anything like a direct perception of our own agency, even in normal cases: it may have a distinct physiological source from the causal pathway that leads to purposive choice and action.

Social Psychology

Finally, we will consider the more low-tech studies in social psychology. In his provocative and entertaining book, *The Illusion of Conscious Will* (2002), psychologist Daniel Wegner elaborates a number of findings that show that people in general are profoundly susceptible to the *inducement of false beliefs regarding one's agency*. In one study, participants are invited to move a computer mouse in concert with someone who is a secret confederate of the experimenter. They are told to choose freely where on the screen to move the cursor, although in fact the confederate is gently forcing the selection. When, just before the mouse is moved, subjects hear in their headphones a recording of a word (e.g., "swan") corresponding to the confederate's chosen target, they have an increased tendency to report that *they* acted intentionally in making the selection. Other studies have shown that prompting subjects to have hostile thoughts about someone prior to the person's appearing to have suffered negative consequences makes the subjects more inclined to view themselves as responsible for the outcome. (For example, you tell the subject to think negative thoughts about a person before performing a voodoo curse on a voodoo doll and have the object of the curse feign a headache.) Likewise, envisioning a positive outcome before a favorite team's televised sporting event leads to increased feelings that one has somehow influenced the outcome. Running in the opposite direction, persistent testimony of others can lead people to believe falsely that they previously performed an action. And yet other studies have shown malleability in one's illusory sense of direct control over *another* person's actions.

These are all examples of *induced false beliefs regarding agency*. A second type of finding in social psychology is the surprising degree to which often *unrecognized situational factors influence individual moral choices*. A variety of studies indicate that the percentage of subjects willing to help someone in need will vary significantly or even dramatically depending on such factors as how many other persons were perceived to be in a position to help, how busy the subjects were at the time, and even whether ambient noise or odors differed from normal levels (Doris 2002; Doris & Stich 2006). It appears that these situational factors may in some cases be better predictors of behavior than the subject's general character traits, as self-reported. The challenge that these findings may pose to human freedom, in the view of some (Nahmias 2007 and forthcoming), is twofold: (1) our actions may be heavily influenced by arbitrary situational factors that we do not control and whose influence we are often unaware of; and (2) ever since Aristotle, it is common to ground human responsibility in our making choices that over time help to form a character out of which much subsequent behavior flows. The skeptical claim is that general moral character is not as significant in explaining our behavior as this presupposes.

2 Some Conceptual Tools in Aid of Deflating Some of the Challenges

We now have before us some of the empirical results thought to threaten the view that human beings are freely directing their actions via conscious willings. In thinking through the several cases, the distinctness of a number of agency-related concepts needs to be born in mind. Consider, then, each of the following:

First, three items in the category of *will or desire*:

- *Minimally voluntary action*: an action that unfolds "automatically," rather than being consciously willed, while coinciding with one or more of one's desires or intentions, conscious or unconscious, and lacking signs of external or internal compulsion.
- *Willing, or conscious forming of an intention to act*: a purposive and executive, or action-initiating, event. Note that these come in two sorts, one that is present-directed (deciding to act here and now) and one that is future-directed (deciding now to act at some particular future time or upon recognition of an appropriate stimulus).
- *Urge or want*: a felt desire to perform an action that one may or may not satisfy.

Next, two sorts of *belief*:

- *Belief concerning one's action*: We often have beliefs concerning what we are about to do (and why), beliefs concerning what we are now doing (and why), and beliefs concerning what we have recently done (and why). It is an empirical question the extent to which these types of belief are aligned.
- *Belief concerning the wider causal import of one's basic actions*: we also have beliefs concerning the more or less immediate effects of our basic actions (those mental actions or bodily movements that we directly control). These are generally inferred from observational clues, though rarely is the inference consciously made.

Finally, two sorts of *experience*:

- *Experience of willing or intention-formation*: when we consciously decide a course of action, there is an "actish phenomenal quality" (Ginet 1990) – we experience ourselves as willing or intending our basic actions. (*Occasionally*, as when we struggle to reach a difficult decision, this involves the further quality of *effort*.) It is both an empirical and a theoretical question as to the relationship of this experience to the willing/intention itself.
- *General "sense of authorship":* It seems correct to say that there is a comparatively persisting yet phenomenally less distinct experience of being the author of what we do, an experience that coincides with any sort of activity of which one is minimally consciously aware. This has no evident conceptual or even

evidential relationship to the spontaneous, natural belief that our actions are metaphysically free, though it may help causally to sustain that belief.

It should be fairly uncontroversial that each of these categories pick out real and distinct phenomena (with the possible exception of the controverted category of conscious willing itself, which is in any case *conceptually* distinct from each of the others). Yet the attentive philosopher who reads recent studies in the psychology and neuroscience of volition will frequently be hard-pressed to map references he encounters to "experience of will" and "conscious decision," terms that generally go undefined or underdefined, onto one or another of these categories with any confidence. In any case, when we do bear these distinctions firmly in mind, we can readily see that *some* of the empirical findings I canvassed do not pose any obvious threat to the belief in human freedom.

2.1 Ex Post Facto Confabulation

Gazzaniga's neurosurgical patients who gave reasons for the arm movements that were artificially induced were clearly confabulating after the fact, perhaps due to a natural psychological mechanism that attempts to impose coherence between one's beliefs and observations concerning one's own movements. And social psychologists have shown that one can cause someone to confabulate the very occurrence of an action in order to produce coherence with the seemingly sincere testimony of others claiming to have observed such an action. But the scientist who notes these cases of unwitting, ex post facto formation or revision of actional beliefs should not be tempted to be partly complicit in them by agreeing that there was, after all, an *experience* of willing that fits the fabricated version of events! What we have here are unremarkable instances of our occasional penchant for forming false memories, rather than cases of illusory experiences of will, which is the skeptic's purported target.

2.2 Erroneous Beliefs Concerning the Wider Effects of One's Actions

Consider next the studies indicating that what is termed a "sense of agency" involving environmental outcomes one is not in fact controlling, such as another person's falling ill or a sports team's performance, can be induced *to some degree* by having the subject perform actions as simple as intentionally having certain thoughts. Whether in these cases there is the same "sense of agency" as generally accompanies ordinary intentional action is very doubtful. But what clearly is occurring is that subjects are being caused to form false beliefs concerning the *effects* of their basic actions. That we can easily be led to increased inclinations towards such beliefs absent good evidence is disconcerting, but again it seems besides the point when the question is our control over our own basic actions via the will.

2.3 *Automatisms*: Minimally Voluntary Actions Unaccompanied by Any Feeling of Agency

These strange cases include alien hand syndrome and the schizophrenic belief that someone else is controlling one's behavior from within. We might also include nonclinical, episodic instances among perfectly normal subjects, such as the table-turning phenomena among nineteenth-century spiritualists. (People seat themselves around a table with fingertips lightly touching the edge. As they await a message from a dead person, sensitive instruments reveal that, one by one, they begin ever so slightly to rotate the table – while quite sincerely believing that they are entirely passive observers of this event.) Such cases strike closer to the skeptical target insofar as they show that a feeling of agency is not necessary to *minimally voluntary* agency. However, such cases should be seen against the backdrop of the more pervasive fact of *automaticity*: for a great deal of our behavior, there is a feeling of authorship, yet we do not directly will or intend the actions in a conscious way. (Routine or learned skilled behaviors are the clearest instances, and they also underscore the usefulness of automaticity. As William James observed, we walk best along a beam the less we attend to the position of our feet upon it.) This pervasive automaticity has long been absorbed into our pretheoretical understanding of ourselves as agents, and, accordingly, it is not these sorts of actions that we take directly to involve conscious will. The automatisms of anarchic hand and severe forms of schizophrenia are unusual and striking only because the unconsciously generated action is at odds with the agent's standing intentions and is unaccompanied by the usual "sense of authorship." That subconscious processes can generate intentional behavior with these characteristics says nothing about the truth or falsity of our pretheoretic conception of ourselves as consciously and freely willing what we do on *other* occasions, just when we take ourselves to so will. They are not evidence in favor of some skeptical model of human agency over against a model on which we enjoy significant freedom because their data are not contrary to what our ordinary pretheoretic understanding would expect.

Instead of making the case in terms of a supposed *disconfirmation* of the ordinary view, on which the commonsense "hypothesis" is shown to predict consequences that are contrary to fact, a skeptic might try to argue alternatively that they are simply *better explained* by an alternative model, such as the one Wegner proposes, on which all actions are generated by subconscious mechanisms into which "conscious will" does not enter. I will not try to develop and respond to this form of the skeptical argument in any detail. But I note that it involves the difficult and contentious matter of whether, and how deeply, we are rationally entitled to start empirical inquiry with a strong presumption that our conviction that we (sometimes) direct our own actions via conscious willing is correct. The unavoidability of unargued yet essential starting assumptions is a familiar fact of life to philosophers, whose training, if not everyday practice, involves the contemplation of all manner of radically skeptical hypotheses about the world and our knowledge of it. (What evidence can show that my lifelong perceptual experience is not a continuous dream, or the product of some nonveridical source such as the *Matrix*?) It is a different matter for scientists, of course, whose research is entirely unperturbed and unguided by such fanciful philosophical queries. This is a problem only insofar as it may induce in some scientists a false understanding of the scientific enterprise as resting on no foundational assumptions whatsoever. It is commonly recognized that science must presuppose that experience is not wholly illusory – there is a world external to our minds, roughly of the sort experience suggests; that our foundational forms of reasoning, deductive and inductive, are sound; and that laws that hold in the observed parts of the universe can reasonably be generalized to the unobserved. It is less often recognized that scientists must presume the fundamental reliability (not perfect, but to a high degree) of their *methods of intervention* in testing theories. That is to say, if we are to accept the data that are collectively adduced in a given domain, we must presume that it resulted from individual processes that, *inter alia*, reflected the actual intentions of those who collected it. Were it not the case that in the conduct of an experiment, scientists are reliably aware of what they are doing and why they are doing it, we would have little reason to accept the significance of their reports.

2.4 Libet Cases

I now turn to Libet's findings (replicated by others) that in his experimental set-up there is a smooth build-up of electrical activity in the motor cortex a few hundred milliseconds prior to the time the subjects indicated as the onset of the conscious urge or wish to perform the movement. Let's begin by noting that in the experiment a subject agrees at the outset to perform a specific action within a short interval of time. All that is left to be determined by the agent is the precise time of its occurrence. Though there is more that needs remarking on, we should not ignore a very obvious fact here: in agreeing to cooperate with the experiment as described, *the agent has already decided to perform a specific action*. In our terms, she forms the future-directed intention to Ø, for a specific action-type Ø (say, flicking her wrist or wiggling her finger). No evidence is adduced that there is a slow build-up towards a readiness potential before *this* sort of decision. Libet, however, will say that this is inconsequential. For the agent goes on to form another, ostensibly free, intention to perform the action here and now, and his choice is shown to have a neural antecedent associated with the movement itself.

But is this really so – is this choice of timing a prototypical instance of a willing that is experienced as wholly spontaneous? As Al Mele (1997) points out, in describing the instructions that the experimenter gives to the subject, Libet (1985) uses several terms interchangeably: "urge," "desire," "wish," and "intention." The subject is asked to note the precise timing of such an urge/desire/wish/intention for each of several movements that will be performed in a single setting. This experimental set up and set of instructions, it seems to me, invites an interpretation differing sharply from Libet's own. First, the action is hardly a "spontaneous" one. Even though the subjects are instructed not to pre-plan the timing of their actions, and instead wait for the urge/desire to do so, the action *type* itself is pre-planned, and even its timing is to a significant degree, as they are instructed to make the movement within a thirty-second or so interval of time. Secondly, by asking the subjects to carefully introspect to pinpoint the timing of the impulse to move, the experimenter is inviting the subject to adopt the role of observer in relation to his own conscious experience, and specifically to wait for an unplanned *urge* to occur. This certainly encourages a passive posture. Having decided that one will move, one looks for the *urge* to do so in order to act upon it. I suggest that such a preformed intention to act upon the right internal "cue" initiates an unconscious process that promotes the occurrence (or perhaps *evolution*) of a conscious state of desire or intention that is not actively formed. In context, the state's default is to trigger the pre-planned activity, absent a last-second "veto" by the agent. In another study, Libet confirms the possibility of such vetoes subsequent to the experience of such an urge.

Given the plausibility of this alternative interpretation – on which the subject's actions are not *prototypical* spontaneous conscious willings – Libet's pioneering studies do not have anything like clearly negative implications for human freedom. To be sure, they have evidently opened up a fruitful field of neuroscientific inquiry and are helpful in pointing to the need for greater attention to the precise phenomenology of different instances of willing or desiring to act. One wonders, for example, whether some or all of the subjects experience a growing anticipation that they are about to act, one that is not experienced vividly at first. Clearly, conscious intentional states generally (as with our perceptual states) have phenomenal aspects that we cannot readily articulate and, importantly, our awareness of them comes in degrees, a point I shall now emphasize.

3 Philosophical and Scientific Models of the Will: Towards an Interface

To this point I have mostly thrown cold water on debunking-free-will claims that are based on the interesting and varied recent studies on the will. Nonetheless, I think that such findings do point to the need for fine-tuning of philosophical models of the will. Philosophers, especially those of us who are incompatibilists, are given to simple and idealized conceptions of human freedom. But the ever-accumulating empirical data and emerging partial theories present a messy picture that underscores the fragility of our freedom, and some elements of that picture cannot be readily mapped onto those idealized philosophical theories. One way in which philosophical models tend to over-idealize has already been remarked upon: given the pervasiveness of automaticity, the freedom and responsibility of much of what we do must be thought of as "inherited" from the comparatively few *directly free* choices that we make. While some recent philosophers have incorporated that fact into their thinking about the will, it is still not widely enough appreciated, so that philosophers often write as if we are constantly making explicit and considered choices.

In this final section, I want to suggest one other needed accommodation in our philosophical theorizing: a greater place for *conscious knowledge* in our account of freedom of will. Making that accommodation leads pretty directly to a striking consequence: though freedom of the will requires a baseline capacity of choice, the nature of which is the subject of much traditional controversy I will not adjudicate here, the *freedom* this capacity makes possible is, nevertheless, a property that comes in *degrees* and can vary over time within an individual.[1]

Some of the studies summarized above highlight ways in which we can be significantly influenced in our decision-making by circumstantial factors without being aware that this is so. And it is a commonplace that our own motivations at times can be largely or entirely hidden to us. In such cases, I suggest, our freedom is diminished – not evaporated entirely, but diminished, as it is not possible for us to subject our unconscious motivations or guiding beliefs to critical scrutiny and decide what to do in a more reflective way.

But awareness of such factors comes in degrees, along more than one dimension. Here are three such dimensions, all of which seem freedom-relevant:

- degree of awareness of *y*, for each influencing factor *y* (desire, intention, belief, or circumstance);
- the relative "portion" of one's total motivational structure (the totality of influencing factors) of which one is aware;
- the degree of likelihood that an unconscious influence *y* is a factor that one would reflectively endorse, were one to become aware of *y*'s influence. (It is one thing to be partly moved to act by a presently unconscious desire or intention that is well integrated into my character. It is another to be motivated by a factor – such as the level of ambient noise – whose relevance I would repudiate if asked.)[2]

There is at least one more degreed dimension to our motivations that I think bears importantly on our freedom. The medieval philosopher Robert Grosseteste in one place invites us to imagine God's creating an angel that exists for a single instant only.[3] (Here I must ask my readers more at home in science than in such

[1] The philosophical account of freedom into which I myself wish to absorb these elements is an agent causal theory. Many philosophers think that the challenges of squaring free will with the emerging sciences of brain and behavior are more severe for the agent-causal view than for the other views, even other indeterminist views. But this is doubtful. All such views seem to require a strongly antireductionist assumption with respect to mental states and the capacity of choice vis-à-vis the complex physical states that subserve them. Once the other views' robustly emergent ontological commitments are made explicit, the difference in strength of assumptions made by the models is diminished, as it concerns only what sort, not whether, a top-down causal factor must be introduced.

[2] A corollary to the freedom-relevance of these conditions is that one can increase one's freedom by becoming more aware of typical unconscious influences on human decision-making. This point is noted by Nahmias (forthcoming), who also emphasizes the relevance of conscious knowledge to freedom of will.

[3] See his *On the Freedom of the Will* (translated into English by Lewis (1991). Grosseteste's views influenced views influenced (via Henry of Ghent) the arch-champion of freedom of the will, John Duns Scotus.

philosophical flights of fancy to bear patiently with me.) In that instant, we are to imagine, the angel exercises his freedom by an instantaneous act of the will. In order to circumvent quite sensible worries about the coherence of Grosseteste's thought experiment, let us stretch out the angel's life to a few seconds. We will also suppose that the angel springs into existence with a fully developed psychology (of the typical angelic sort), complete with a bunch of pseudo-memories (false apparent memories) of a long, character-shaping history. And, as in Grosseteste's telling, he comes with a disposition to decide some matter straightaway. Finally, let us imagine that he is deciding between a plurality of alternatives, each having some attraction to him, is *fully* aware of his own motivations, and has the capacity to determine himself to do any of them. (Philosophers may insert here their favored account of this "self-determining" capacity.) In short, his will appears to be significantly free. Here is my question: in making his one and only choice, is our angel (whom we may call "Angel*o*") just as free as an intrinsically identical counterpart ("Angel*a*") who at the time in question is intrinsically identical to Angelo, while differing dramatically *historically*: unlike Angelo, Angela really does have a history, one filled with many prior choices that have partly shaped her present inclinations and intentions? At the time at which they both chose, where they happen to coincide perfectly in their inclinations and capacities, are they equally *free*?

It seems not. The reason is that for Angela, unlike Angelo, the very factors that shape her choice were to some extent of her own making. Like Angelo, she began with a set of psychological and behavioral dispositions that were merely "given." But over time, as she habitually made certain choices, her psychological makeup reflected less and less this "givenness" and more and more something that is her own free creation.

And so, too, for ourselves. We come into the world with powerful tendencies that are refined by the particular circumstances in which we develop. All of these facts are for us merely "given." They determine which choices we have to make and which options we will consider (and how seriously) as we arrive at a more reflective age. Now, most of us are fortunate enough not to be impacted by traumatic events that will forever limit what is psychologically possible for us, and, on the positive side, are exposed to a suitably rich form of horizon-expanding opportunities. Where this is so, the framework of considerations that structures our choices increasingly reflects our own prior choices. And, in this way, our freedom grows over time.

For a further reason to think this is right, consider a scenario involving an agent much as we take ourselves to be – except that his psychology is regularly manipulated, altering some of his preferences and the strengths of others. Owing to the marvelous, wireless neural-intervention technology of the late twenty-first century, all this occurs while he remains wholly oblivious. Even if his capacity to choose remains robust, it seems clear that we must judge his freedom, his autonomy, to be diminished.[4] The integrity of the self-formation process is a component of freedom,

[4] Al Mele (1995), ch. 9, makes this point, and Clarke (2003, pp. 16n.4, 77) concurs. For a dissenting view, see Daniel Dennett (2003) pp. 281–87.

or of freedom of the most valuable sort. This conclusion is reflected in corresponding judgments about moral responsibility. If Angelo and Angela each contemplate a morally significant matter, are inclined to a degree to both a virtuous and a vicious action, and choose the virtuous one, Angela seems the more praiseworthy. The action is *hers* to a greater degree.

Thus, a fourth neglected dimension of freedom is historical: the degree to which motivation *y* is a product of the agent's own past free choices.[5] I suspect that there is a corresponding condition, harder to state, with respect to *beliefs* that also shape one's choices: I am freer to the extent to which I am not being influenced by beliefs that have been formed by defective mechanisms, or even by importantly false beliefs that have arisen through nonculpable misuse of normally functioning mechanisms. (The hallucination-generated beliefs in schizophrenics are an extreme case of what I have in mind here.)

It is an open empirical question the extent to which any given individual, or human beings in general, realize these conditions – just as it is an open question whether and when the basic capacity to choose, to which philosophers give most of their attention, is present and regularly exercised in a way necessary for true freedom of choice. And so empirical sciences of brain and behavior potentially have a lot to teach us about the *extent* and *scope* of human freedom, once we recognize the mistake of thinking that it is an all-or-nothing matter. In truth, human freedom is always limited, fragile, and variable over time and across agents. It is the sort of thing that comes in degrees – a fact that should inform not only our philosophical-cum-scientific theorizing, but also our moral understanding and assessment of one another.

Acknowledgments

I have benefited greatly from discussion of ideas advanced in this chapter with several members of the "Top-Down Causation and Volition" symposium organized by the John Templeton Foundation in Yosemite National Park in April 2007, including Daniel Wegner, Cristoph Koch, Itzak Fried, Chris Frith, and Sean Spence. A version of this chapter was given as a plenary lecture at the 2007 meetings of the American Catholic Philosophical Association. I received helpful feedback and criticisms from Alex Pruss, Timothy Noone, and Robert Kane. In 2008, I read a revised version to a cognitive science seminar of the Institut Jean Nicod (École Normale Supérieure) in Paris; the 2008 German Congress of Philosophy in Essen; a 2008 conference at the University of Valencia, and colloquia at Denison and Purdue. I thank all the audiences on these occasions, especially the (insistently skeptical) Joëlle Proust, Max Kistler, Neil Roughley,

[5] As Neil Roughley pointed out to me, the specific effects of past decisions are relevant to their bearing on the freedom of future choices. We can imagine scenarios, for example, in which early choices cause one to fixate on certain goals, without one's having foreseen or intended this consequence, and this would clearly diminish rather than enhance one's future freedom. I'll not attempt here to develop an account of the historical dimension of freedom that captures this nuance.

Michael Bergmann, and Dana Nelkin. Some lines in the text were taken from my 2005 article and adapted for the present chapter.

References

Brasil-Neto, J.P., Pascual-Leone, A., Valls-Sole, J., Cohen, L.G., Hallett, M.: Focal transcranial magnetic stimulation and response bias in a forced choice task. Journal of Neurology, Neurosurgery, and Psychiatry 55, 964–966 (1992)

Clarke, R.: Libertarian accounts of free will. Oxford University Press, New York (2003)

Delgado, J.M.R.: Physical control of the mind: Toward a psychocivilized society. Harper & Row, New York (1969)

Dennett, D.: Freedom evolves. Viking, New York (2003)

Doris, J.: Lack of character: Personality and moral behavior. Cambridge University Press, Cambridge (2002)

Doris, J., Stich, S. (2006), Moral psychology: Empirical approaches. In: Stanford encyclopedia of philosophy, `http://plato.stanford.edu/entries/moral-psych-emp/`

Fischer, J.M.: The metaphysics of free will. Blackwell, Oxford (1994)

Frankfurt, H.: Alternative possibilities and moral responsibility. Journal of Philosophy 66, 829–839 (1969)

Gazzaniga, M.S.: Consciousness and the cerebral hemispheres. In: Gazzaniga, M.S. (ed.) The cognitive neurosciences, pp. 1391–1399. MIT Press, Cambridge (1994)

Ginet, C.: On action. Cambridge University Press, Cambridge (1990)

Haggard, P.: Conscious intention and motor cognition. Trends in Cognitive Sciences 9, 290–295 (2005)

Haggard, P., Clark, S., Kalogeras, J.: Voluntary action and conscious awareness. Nature Neuroscience 5, 382–385 (2002)

James, W.: The principles of psychology, vol. 2. Henry Holt & Co., New York (1890)

Lau, H.C., Rogers, R.D., Ramnani, N., Passingham, R.E.: Willed action and attention to the selection of action. NeuroImage 21, 1407–1415 (2004)

Lau, H.C., Rogers, R.D., Passingham, R.E.: On measuring the perceived onsets of spontaneous actions. Journal of Neuroscience 26, 7265–7271 (2006)

Lau, H.C., Rogers, R.D., Passingham, R.E.: Manipulating the experienced onset of intention after action execution. Journal of Cognitive Neuroscience 19, 81–90 (2007)

Libet, B.: Unconscious cerebral initiative and the role of conscious will in voluntary action. The Behavioral and Brain Sciences 8, 529–566 (1985)

Mele, A.: Autonomous agents. Oxford University Press, New York (1995)

Mele, A.: Strength of motivation and being in control: Learning from Libet. American Philosophical Quarterly 34, 319–332 (1997)

Nahmias, E.: Autonomous agency and social psychology. In: Marraffa, M., Caro, M., Ferretti, F. (eds.) Cartographies of the mind: Philosophy and psychology in intersection, pp. 169–185. Springer, Dordrecht (2007)

Nahmias, E.: The psychology of free will. In: Prinz, J. (ed.) Oxford handbook on philosophy of psychology. Oxford University Press, Oxford (forthcoming)

O'Connor, T.: Persons and causes: The metaphysics of free will. Oxford University Press, New York (2000)

O'Connor, T.: Freedom with a human face. Midwest Studies in Philosophy 29, 207–227 (2005)

Wegner, D.: The illusion of conscious will. MIT Press, Cambridge (2002)

Wegner, D., Wheatley, T.: Apparent mental causation: Sources of the experience of will. American Psychologist 54, 480–491 (1999)

11

Contemplative Neuroscience as an Approach to Volitional Consciousness

Evan Thompson

Department of Philosophy
University of Toronto
170 St. George Street, 4th floor
Toronto, ON M5R 2M8
Canada
evan.thompson@utoronto.ca

Summary. This chapter presents a methodological approach to volitional consciousness for cognitive neuroscience based on studying the voluntary self-generation and self-regulation of mental states in meditation. Called contemplative neuroscience, this approach views attention, awareness, and emotion regulation as flexible and trainable skills, and works with experimental participants who have undergone training in contemplative practices designed to hone these skills. Drawing from research on the dynamical neural correlates of contemplative mental states and theories of large-scale neural coordination dynamics, I argue for the importance of global system causation in brain activity and present an "interventionist" approach to intentional causation.

Keywords: volition, consciousness, neurodynamics, neurophenomenology, contemplative neuroscience, meditation.

In this chapter I sketch a methodological approach to volitional consciousness for cognitive neuroscience based on studying the voluntary self-generation and self-regulation of mental states in meditation. Called contemplative neuroscience, this approach views attention, awareness, and emotion regulation as flexible and trainable skills, and works with experimental participants who have undergone extensive training in contemplative practices designed to hone these skills. My discussion here is premised on the following three working assumptions (Lutz et al. 2007; Lutz et al. 2008):

- Advanced contemplative practitioners can generate new data that would not exist without contemplative mental training. These include various conscious states and processes occurring during contemplative practices as well as longer lasting traits that may be brought about by these practices.
- Advanced contemplative practitioners can reliably reproduce and maintain specific aspects or types of conscious processes. These include focused attention, one-pointed concentration, and various types of meta-cognitive awareness. This ability to stabilize conscious processes makes them easier to investigate experimentally.
- Advanced contemplative practitioners can give precise first-person descriptions of conscious mental states. This information is relevant for refining psychological taxonomies of the full range of conscious states and for interpreting neuroimaging data about conscious processes.

1 Toward a Neurophenomenology of Volition

Most cognitive neuroscientists studying consciousness maintain that self-reports of experience provide indispensable evidence about conscious processes (Jack & Roepstorff 2002). Although scientists often acknowledge the conceptual distinction between phenomenal consciousness (subjective experience) and access consciousness (cognitive access and reportability) (Block 1997), they usually argue that there is no clear scientific criterion for defining a process as conscious apart from its reportability (e.g., Dehaene & Naccache 2001). Nevertheless, as psychologists have discussed, being able to gain access to experience and report it is a cognitive capacity in its own right. When asked to report and describe our experiences – for example, our experiences of conscious will (Wegner 2002, 2004) – we need to introspect in order to become explicitly aware of our experience. In other words, we must consciously represent to ourselves our experience, thereby engaging meta-consciousness or meta-awareness, a distinct form of meta-cognition (Schooler 2002).

It is well known that self-reports requiring introspection or meta-awareness are subject to various biases, especially when subjects are asked or allowed to report what they take to be the causes of their experiences (Nisbett & Wilson 1977; see also Hurlbert & Heavey 2001; Schooler 2002). Yet even when subjects are discouraged from offering reasons or explanations, and are encouraged simply to describe their experiences as carefully as possible, a variety of interrelated difficulties present themselves (Petitmengin 2006; Schooler 2002; Schooler & Schreiber 2004).

One difficulty has to do with the instability of attention. Our attention tends to jump rapidly from one thing to another. As William James observed, in the case of voluntary attention, it takes considerable effort to sustain attention on a given object (James 1985, p. 91). Such mental effort seems especially present when the target of attention happens to be our experiences or mental processes and the type of attention is accordingly endogenous attention requiring top-down cognitive control.

A second difficulty is lack of awareness of experience. Not only does our attention jump around, but we usually have little or no explicit awareness of this attentional instability. Mind-wandering is a familiar case. Engaged in some task such as reading or writing, our attention wanders and we become lost in spontaneously arising thoughts and memories. Although we are conscious and have a variety of experiences with specific contents, we have little or no explicit awareness of those experiences as they occur. If we catch ourselves daydreaming, we then become meta-conscious of our subjective mental activity. In such cases there is a "temporal dissociation" between consciousness and meta-consciousness, "in which the triggering of meta-consciousness causes one to assess aspects of experience that had previously eluded explicit appraisal" (Schooler 2002, p. 340).

A third difficulty is that introspection or meta-awareness can change experience, particularly if the experience is nonverbal and verbalization is required (Schooler 2002). For example, directing attention towards an emotion and trying to verbalize it can change its phenomenal character in a variety of ways (Lambie & Marcel 2002).

Finally, a related difficulty is the possibility for misrepresentation that gets introduced when one has to represent the contents of consciousness in meta-consciousness. Schooler (2002) calls this sort of dissociation between consciousness and meta-consciousness "translational dissociation": "If meta-consciousness requires re-representing the contents of consciousness, then, as with any recoding process, some information could get lost or become distorted in the translation" (Schooler 2002, p. 342).

As scientists aiming for a better understanding of consciousness, we thus find ourselves in the following situation. On the one hand, self-reports of experience are indispensable and arguably the main source of evidence for the presence of a given conscious process. On the other hand, gaining cognitive access to experience and being able to report and describe it with precision are abilities that present their own challenges and presumably vary across individuals (as readers of novels already know).

Given this situation, we might conjecture that individuals who strive to develop a high degree of intimacy with and control over their own subjective mental processes through contemplative training of attention and meta-awareness can provide more detailed and accurate self-reports about the contents of consciousness (Lutz et al. 2008). This conjecture is one of the working assumptions of contemplative neuroscience (Lutz et al. 2007).

Volitional consciousness provides a potential case study for testing this conjecture. There has been little sustained investigation of the phenomenology of volition (even in the tradition of Phenomenology), particularly of the sorts of volitional experiences philosophers appeal to when they defend one or another account of free will (Nahmias et al. 2004). At the same time, psychologists such as

Wegner (2002, 2004) have concluded that the experience of conscious will as causally efficacious is illusory, though careful attention to the phenomenology of volition and agency does not seem to support this conclusion (Bayne 2006). Clearly, more phenomenological work needs to be done to investigate the experience of volition. Yet how should such investigation proceed?

Nahmias and colleagues propose that we should investigate the "folk phenomenology of free will," suggesting that "reports gathered from laypersons will be minimally tainted by *philosophical* theory" (Nahmias et al. 2004, p. 172).[1] They note the need for introspective reports from subjects who are not trained in theoretical debates (unlike the nineteenth- and early twentieth-century Introspectionists). Yet they also realize that obtaining self-reports about the experience of volition encounters difficulties of the sort discussed above:

> Our goal is to understand the phenomenology of free will in a way that informs the theoretical debate without tainting the phenomenology itself. But one might argue that, without training, the phenomenology of free will is either too difficult to apprehend or to describe or both. So, even if there *is* a folk phenomenology of free will, we may be unable to get systematic descriptions of it from folk who have yet to be trained in some relevant way.… For this purpose, it would be helpful for psychologists, perhaps guided by the questions raised in the philosophical debate, to reconsider the basic introspectionist project of gathering first-person reports, even offering some guidance about how to attend to conscious phenomena, while avoiding the introspectionists' tendency to train subjects in the theoretical debates. If the data from subjects' reports can be triangulated with behavioural and neuropsychological data, all the better. (Nahmias et al. 2004, p. 172)

Here we need to distinguish between training in theoretical debates and training in mental skills involving attention to and awareness of conscious processes. In calling for psychologists to offer guidance in how to attend to conscious phenomena, Nahmias and colleagues acknowledge that such attention can be guided and thus trained. They recommend the phenomenological interview, which uses open-ended questions to guide subjects to describe their experience while directing them away from trying to explain it. Such guidance is a form of training, for subjects learn to attend to their experience without trying to rationalize it or explain its causes (Petitmengin 2006).

This procedure of combining first-person phenomenological investigation, second-person phenomenological interviews, and third-person behavioral and neurophysiological measures is central to the approach known as neurophenomenology (Lutz et al. 2002; Lutz & Thompson 2003; Varela 1996). Thus, one way to make headway in understanding volitional experience and its relation to the brain is to pursue a neurophenomenology of volitional consciousness.

Contemplative neuroscience builds on neurophenomenology by proposing to supplement phenomenological reports from laypersons with reports from

[1] These authors are careful to say *minimally tainted* rather than *untainted.* As phenomenologists such as Husserl, Heidegger, and Merleau-Ponty have discussed, commonsense understanding typically contains a large amount of philosophical sediment. Indeed, it has been claimed that our modern Western conception of the will is a philosophical idea invented by Augustine (e.g., Murphy 2006, p. 14).

individuals who engage in the contemplative training of attention and meta-awareness (Lutz et al. 2007). Consider the Theravada Buddhist practice known as mindfulness of intention, often practiced during walking meditation. One cultivates an attention to and awareness of how the arising of an intention or volition precedes every movement.[2] Usually intentions or volitions arise without this sort of awareness, and hence almost always lead automatically to action. When one notices their arising, however, one gains the ability to choose whether to act on them or not. One might think that this mental practice would interrupt the flow of action, but practitioners report that it actually helps one to reinhabit the flow of everyday action with heightened attunement and less mindless automaticity. This type of practice thus seems well suited for phenomenological investigations of the experience of volition and agency, and individuals accomplished in this practice might be able to provide information about conscious volitional processes unavailable to untrained individuals.

2 Meditation and Neurodynamics

From the perspective of dynamical neuroscience, transient conscious states are embodied in large-scale dynamical patterns of temporally coordinated neural activity across selective brain regions and areas (see Cosmelli et al. 2007 for a review). The voluntary generation of mental states in meditation seems to be no exception. In a recent study, Lutz and colleagues found distinct dynamical patterns of electrical brain activity recorded at the scalp in advanced Tibetan Buddhist meditators compared with novice practitioners during the voluntary generation of a specific kind of meditative state (Lutz et al. 2004).[3] The brain waves of the long-term meditators showed high-amplitude gamma oscillations (25–42 Hz) and

[2] The Sanskrit or Pali word *cetana* can be translated as either intention or volition. In Buddhist psychology *cetana* is one of the so-called constant mental factors that is present in every moment of consciousness and that functions to direct a mental state towards its object.

[3] The advanced meditators had undergone training for 10,000 to 50,000 hours over time periods ranging from 15 to 40 years. The novices had undergone training for 1 week before the data collection. The meditative state generated by the practitioners is known as nonreferential compassion. Tibetan Buddhists define compassion as the deep wish that others be happy and free from suffering. Usually this wish is directed towards a specific person or group, but nonreferential compassion, though necessarily other-directed, does not have a particular target. It is described as a state of being in which an unconditional feeling of loving-kindness and compassion pervades the whole mind, while awareness rests calmly and stably in complete openness without focusing on any particular object. This sheer awareness is considered to be invariant across all modes of consciousness regardless of the particular contents of consciousness. In nonreferential compassion meditation, the aim is to settle one's mind in this fundamental awareness while cultivating an intense feeling of compassion. This sort of "objectless meditation" is thus different from types of meditation that require concentration on an object (such as one's breath or a mental image). The cultivation of nonreferential compassion is thought to transform the mind in profound ways, by lessening fixation on self, counteracting afflictive states of mind (such as hatred and jealousy), creating a general sense of well-being and an unrestricted mental readiness and availability to help others, and counteracting mental dullness in meditation practice.

phase-synchrony over lateral frontoparietal electrodes during meditation. The ratio of gamma frequency activity to slow frequency activity (4–13 Hz) was higher in the baseline resting state before meditation for the adepts compared with the novices, and this ratio increased sharply during meditation and remained higher after meditation. In other words, brain activity in the resting state before meditation was already significantly different between the two groups, and this difference increased markedly during meditation. Furthermore, when the adepts entered the 30-second resting-state periods between the 60-second meditation periods, their brain activity did not return to the initial premeditation baseline resting state, but instead displayed an ongoing baseline that reflected the previous meditation session.

These findings suggest that meditation induces short-term changes in neural activity and may bring about long-term changes in the brain. Nevertheless, this study cannot tell us whether contemplative training brings about these neural activity patterns or whether they reflect preexisting individual differences. A more recent longitudinal study, however, which examined the effects of three months of intensive training in insight meditation, showed that such training leads to increased control over the distribution of limited neural resources in attention (Slagter et al. 2007).[4] This study provides direct evidence that systematic contemplative mental training affects the brain.

Another question concerns the significance of the high amplitude gamma oscillations seen during meditation (or the significance of gamma synchrony more generally in relation to conscious processes). Can such measures be related in any way to conscious experience?

To address this sort of question it would help to have more information about what is going on subjectively in people's minds from moment to moment as they engage some mental process. Working with highly trained contemplatives could be a real advantage here. These individuals spend years honing their capacities of attention, concentration, and meta-awareness, so it stands to reason they can describe their own subjective experience more precisely than can individuals who lack this kind of mental training.

Following this line of thought, Lutz and colleagues in a subsequent study (Lutz et al. 2006) asked long-term contemplatives to describe their subjective experience during meditation. Furthermore, instead of imposing already established terms from Western psychology, they asked the adept practitioners to report on their experiences using descriptive terms from their own tradition of contemplative theory and practice. Specifically, they asked the practitioners to report on the quality of "clarity" in their meditation.

[4] This study examined the effects of intensive meditation training on the attentional blink: When two targets are presented in close temporal proximity and at the same location in a sequence of visual stimuli, the second target is often not seen. This effect is thought to result from competition between the two targets for limited attentional resources. The study found that intensive meditation training resulted in a smaller attentional blink and reduced brain resource allocation to the first target, as reflected by a smaller P3b waveform (an event-related potential thought to index the allocation of attentional resources).

Tibetan Buddhists use the term “clarity” to refer to the subjective intensity of the meditative state (see Lutz et al. 2007). Using the metaphor of light, they describe clarity as the luminosity or brilliance of the state. “Stability,” on the other hand, shields the light or flame of clarity from flickering or going out. “Stability” refers both to the degree to which one stays in the meditative state, instead of being perturbed out of it, and the ease with which one regains the state if dislodged from it. Clarity and stability contrast respectively with dullness and excitation: A dull meditative state lacks clarity and an excited meditative state lacks stability. Thus a meditative state might be unstable – one gets repeatedly bumped out of the state – but nonetheless clear or intensely experienced when one is in it. Or it could be stable but dull. In inexperienced meditators, clarity and stability tend to work against each other: The greater the stability, the more likely the meditative state is dull, the extreme being that one simply falls asleep; and the greater the intensity, the more likely one becomes excited and distracted, losing stability. The ideal meditative state finds a perfect balance between clarity and stability, so that neither dullness nor excitation impedes the mind.

When Lutz and colleagues asked the contemplative adepts to report on a scale from 1 to 9 any ongoing change in the clarity of their meditation (where 9 was defined as the peak of clarity they believed could be reached the day of the experiment), they found a strong correlation, over a time-course of several dozens of seconds, between self-reports of increasing clarity and the emergence of high-amplitude gamma activity, particularly in frontal regions. Hence the gamma activity observed during this type of meditation seems closely related to the meditative state’s phenomenal quality of clarity.

This way of using phenomenological information to help interpret neuroimaging data about conscious processes provides an example of the third working assumption of contemplative neuroscience mentioned at the beginning of this chapter and of the neurophenomenological approach more generally.

3 Emergence

The brain patterns seen in these and other neurodynamical studies of conscious processes are *emergent* in the following sense: They characterize the behavior of neural networks as complex (metastable) systems; they arise spontaneously given the local couplings among the network’s components and the way those couplings are globally constrained and regulated; and they do not belong to any of the system’s components taken singly or severally (Thompson & Varela 2001).

Elsewhere I have argued that this sort of emergence may involve forms of nonseparability – the emergence of dynamic wholes that supersede or subsume their parts in irreducibly relational structures – and downward causation – the alteration of local behavior by global relational patterns (Thompson 2007).

I also argue, however, that the term “downward causation” is a misnomer. Complex-system causality is not a matter of a higher level acting downwards on a lower level. Rather, the whole entangled system moves at once and always as a result of both local interactions and the way the system's global organization shapes the local interactions (see Thompson 2007 for further discussion).

How might we conceptualize intention or volition in relation to this kind of complexity? Scott Kelso has proposed that an intention corresponds to an order parameter of the system's dynamics – a collective variable that constrains the system's behavior, either by stabilizing or destabilizing it (Kelso 1995, pp. 141–46). As examples he gives the influence of intention on the dynamics of bimanual coordination (e.g., in-phase and anti-phase finger-tapping) and the intentional perceptual reversal of ambiguous figures (Kelso 1995, pp. 218–25). In this connection it is worth noting that Oliva Carter and colleagues (Carter et al. 2005) found that long-term Tibetan Buddhist meditators can measurably alter normal fluctuations in conscious state induced by binocular rivalry (a kind of bistable perception). Specifically, complete perceptual stability for several minutes was induced by focused attention (one-pointed concentration) type meditation. Another example of the intentional modulation of dynamical neural activity is voluntarily affecting the course of an epileptic seizure by using cognitive countermeasures to prevent or interrupt it (Thompson & Varela 2001; see also Le Van Quyen & Petitmengin 2002; Petitmengin et al. 2006).

This model of the neurodynamics of intention may also be applicable to the voluntary generation of mental states in meditation. Although highly speculative, the general idea would be that contemplative mental training creates new types of global order parameters for the neural coordination dynamics underlying various conscious processes. The voluntary generation of mental states in meditation would thus correspond to inducing such order parameters in the brain.

4 Volition as Intervention

In this chapter I have taken a methodological approach and have not addressed the explanatory gap between consciousness and the brain or the philosophical problem of mental causation. Not much headway can be made on the first problem until we know more about both the phenomenology of conscious processes and the functioning of the brain as a complex system, and no progress is possible on the second problem if we do not get past the dichotomous concepts of the mental and the physical (inherited from Descartes) (Thompson 2007). For this reason, we need ways of conceptualizing psychological and biological causation that avoid these concepts. Here an "interventionist" account of "upward" and "downward" causation can help.[5]

According to the interventionist theory of causation, for X to be a cause of Y is for intervening on X to be a way of intervening on Y. One way to intervene on biological events is to intervene on psychological events: Actively triggering a

[5] I owe this idea to the philosopher Michel Bitbol (2004). See Woodward (2003) for general discussion of the interventionist theory of causation, and Campbell (2007) for application to psychology.

change in one's mental states by purely psychological means (contemplative mental training, emotion regulation, psychotherapy) may result in short-term and long-term changes to neural activity patterns, immune system function, hormonal patterns, and so on. And one way to intervene on psychological events is to intervene on biological events: Actively triggering a change in one's biological states by purely biobehavioral means (drugs, transcranial magnetic stimulation) may result in short-term and long-term changes to one's mental states. In this view, as Michel Bitbol remarks:

> Making sense of upward and downward causation does not require a metaphysical distinction between the higher and basic levels of organization. Neither a substantial distinction, as in genuine dualism, nor a distinction between properties or structures as in the currently popular picture. It is enough to assume a duality of modes of access, or modes of intervention. If one intervenes at a higher level of organization, some effects of this action can then be detected by a mode of access specifically aimed at the lower level. This is downward causation. Conversely, if one intervenes at a microscopic level, some effects of this action can then be detected by a mode of access specifically aimed at a higher level of organization. This is upward causation. (Bitbol 2004)

This formulation allows us to say in a perfectly coherent way that contemplative experience acts downwardly on the brain by providing a distinct way of psychologically intervening on neurobiological processes.

5 Conclusion

Let me close with William James's observation that the essence of volition is effort of attention: "attention with effort is all that any case of volition implies. The essential achievement of the will, in short, when it is most 'voluntary,' is to attend to a difficult object and hold it fast before the mind" (James 1985, p. 317). For James, the question of psychological fact (rather than metaphysical speculation) in the free-will controversy "relates solely to the amount of effort of attention which we can at any time put forth. Are the duration and intensity of this effort fixed functions of the object, or are they not?" (p. 323). The line of thought pursued in this chapter suggests that sustained attention (which for advanced contemplatives may no longer be effortful), intention, and volition are not fixed functions of the object, but endogenously generated mental events that guide and control neural activities.

References

Bayne, T.: Phenomenology and the feeling of doing: Wegner on the conscious will. In: Pockett, S., Banks, W.P., Gallagher, S. (eds.) Does consciousness cause behavior? An investigation of the nature of volition. MIT Press/Bradford Books, Cambridge (2006)

Bitbol, M.: Downward causation: Concept and experience. Lecture presented at the conference De l'Autopoièse à la Neurophénomenologie/From Autopoiesis to Neurophenomenology: Un hommage à Francisco Varela/A Tribute to Francisco Varela, June 18-20, Paris, France (2004)

Block, N.: On a confusion about a function of consciousness. In: Block, N., Flanagan, O., Güzeldere, G. (eds.) The nature of consciousness: Philosophical debates, pp. 375–416. MIT Press/A Bradford Book, Cambridge (1997)

Campbell, J.: An interventionist approach to causation in psychology. In: Gopnik, A., Schulz, L. (eds.) Casual learning: Psychology, philosophy, and computation. Oxford University Press, New York (2007)

Carter, O.L., Presti, D.E., Callistemon, C., Ungerer, Y., Lui, G.B., Pettigrew, J.D.: Meditation alters perceptual rivalry in Tibetan Buddhist monks. Current Biology 15, R412–R413 (2005)

Cosmelli, D., Lachaux, J.-P., Thompson, E.: Neurodynamical approaches to consciousness. In: Zelazo, P.D., Moscovitch, M., Thompson, E. (eds.) The Cambridge handbook of consciousness. Cambridge University Press, New York (2007)

Dehaene, S., Naccache, L.: Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. Cognition 79, 1–37 (2001)

Hurlbert, R.T., Heavey, C.L.: Telling what we know: Describing inner experience. Trends in Cognitive Sciences 5, 400–403 (2001)

Jack, A.I., Roepstorff, A.: Introspection and cognitive brain mapping: From stimulus-response to script-report. Trends in Cognitive Sciences 6, 333–339 (2002)

James, W.: Psychology: The briefer course. University of Notre Dame Press, Notre Dame (1985)

Kelso, J.A.S.: Dynamic patterns: The self-organization of brain and behavior. MIT Press, Cambridge (1995)

Lambie, J.A., Marcel, A.J.: Consciousness and the varieties of emotion experience: A theoretical framework. Psychological Review 109, 219–259 (2002)

Le Van Quyen, M., Petitmengin, C.: Neuronal dynamics and conscious experience: An example of reciprocal causation before epileptic seizures. Phenomenology and the Cognitive Sciences 1, 169–180 (2002)

Lutz, A., Thompson, E.: Neurophenomenology: Integrating subjective experience and brain dynamics in the neuroscience of consciousness. Journal of Consciousness Studies 10, 31–52 (2003)

Lutz, A., Lachaux, J.-P., Martinerie, J., Varela, F.J.: Guiding the study of brain dynamics by using first-person data: Synchrony patterns correlate with ongoing conscious states during a simple visual task. Proceedings of the National Academy of Sciences USA 99, 1586–1591 (2002)

Lutz, A., Greischar, L.L., Rawlings, N.B., Ricard, M., Davidson, R.J.: Long-term meditators self-induce high-amplitude gamma synchrony during mental practice. Proceedings of the National Academy of Sciences USA 101, 16369–16373 (2004)

Lutz, A., Francis, A.D., Davidson, R.J.: Changes in the tonic high-amplitude gamma oscillations during meditation correlate with long-term practitioners' verbal reports. In: Association for the Scientific Study of Consciousness Annual Meeting, poster presentation (2006)

Lutz, A., Dunne, J.D., Davidson, R.J.: Meditation and the neuroscience of consciousness: An introduction. In: Zelazo, P.D., Moscovitch, M., Thompson, E. (eds.) The Cambridge handbook of consciousness. Cambridge University Press, New York (2007)

Lutz, A., Slagter, H., Dunne, J.D., Davidson, R.J.: Attention regulation and monitoring in meditation. Trends in Cognitive Sciences 12, 163–169 (2008)

Murphy, N.: Bodies and souls. Cambridge University Press, New York (2006)

Nahmias, E., Morris, S., Nadelhoffer, T., Turner, J.: The phenomenology of free will. Journal of Consciousness Studies 11, 162–179 (2004)

Nisbett, R.E., Wilson, T.D.: Telling more than we can know: Verbal reports on mental processes. Psychological Review 84, 231–259 (1977)

Petitmengin, C.: Describing one's subjective experience in the second person: An interview method for the science of consciousness. Phenomenology and the Cognitive Sciences 5, 229–269 (2006)

Petitmengin, C., Baulac, M., Navarro, V.: Seizure anticipation: Are neurophenomenological approaches able to detect preictal symptoms? Epilepsy and Behavior 9, 298–306 (2006)

Slagter, H.A., Lutz, A., Greischar, L.L., Francis, A.D., Nieuwenhuis, S., Davis, J., Davidson, R.J.: Mental training affects distribution of limited brain resources. PLoS Biology 5/6, e138 (2007), http://www.plosbiology.org

Schooler, J.: Re-representing consciousness: Dissociations between experience and meta-consciousness. Trends in Cognitive Sciences 6, 339–344 (2002)

Schooler, J., Schreiber, C.A.: Experience, meta-consciousness, and the paradox of introspection. Journal of Consciousness Studies 11, 17–39 (2004)

Thompson, E.: Mind in life: Biology, phenomenology, and the sciences of mind. Harvard University Press, Cambridge (2007)

Thompson, E., Varela, F.J.: Radical embodiment: Neural dynamics and consciousness. Trends in Cognitive Sciences 5, 418–425 (2001)

Varela, F.J.: Neurophenomenology: A methodological remedy for the hard problem. Journal of Consciousness Studies 3, 330–350 (1996)

Wegner, D.M.: The illusion of conscious will. MIT Press/Bradford Books, Cambridge (2002)

Wegner, D.M.: Précis of The illusion of conscious will. Behavioral and Brain Sciences 27, 649–692 (2004)

Woodward, J.: Making things happen: A theory of causal explanation. Oxford University Press, New York (2003)

12

Free Will and Top-Down Control in the Brain

Chris D. Frith

Wellcome Centre for Neuroimaging at University College London
12 Queen Square
London WC1N 3BG
United Kingdom
cfrith@fil.ion.ucl.ac.uk

CFIN, University of Aarhus
Aarhus University Hospital
Nørrebrogade 44, Building 30, 8000-Århus C
Denmark

Summary. I suggest that the physiological basis of free will, the spontaneous and intrinsic selection of one action rather than another, might be identified with mechanisms of top-down control. Top-down control is needed when, rather than responding to the most salient stimulus, we concentrate on the stimuli and actions relevant to the task we have chosen to perform. Top-down control is particularly relevant when we make our own decisions rather then following the instructions of an experimenter. Cognitive neuroscientists have studied top-down control extensively and have demonstrated an important role for dorsolateral prefrontal cortex and anterior cingulate cortex. If we consider the individual in isolation, then these regions are the likely location of will in the brain. However, individuals do not typically operate in isolation. The demonstration of will in even the simplest laboratory task depends upon an implicit agreement between the subject of the experiment and the experimenter. The top of top-down control is not to be found in the individual brain, but in the culture that is the human brain's unique environmental niche.

Keywords: top-down, control, attention, prefrontal cortex, anterior cingulate cortex, culture.

In this chapter I shall explore the relationship between free will and the concept of top-down control as used by cognitive psychologists. The concept of top-down control is formulated as a contrast with bottom-up control. If the choice I make is entirely determined by all the forces that are currently impinging upon me, then this would be an example of bottom-up control. By contrast, if I choose my action independently of external forces, then this would be an example of top-down control. And, in making such a choice, I would be exerting free will.

1 Top-Down and Bottom-Up Processes in Attention

Bottom-up and top-down mechanisms of control have been extensively studied in the context of selective attention (Itti et al. 2005). Selective attention addresses the problem of sensory overload. Our senses are constantly bombarded with stimuli. If we had to pay attention to all the different sights and sounds with which we are surrounded, we would not be able to function. Somehow, through selective attention, we manage to ignore most of what is happening around us and attend only to what is important. As a result only a very small proportion of the information striking our senses affects our behavior or impinges upon our conscious awareness. How does the brain achieve this selection?

An influential account of selective attention has been proposed by Desimone and Duncan (1995). The problem can be framed in terms of a competition between the many stimuli that are striking our senses. How is it that only a few of these stimuli win this competition for our attention? Desimone and Duncan proposed that there were two fundamental mechanisms. The first is a bottom-up process of free competition between stimuli through which the strongest stimulus wins. The second is top-down process by which this competition is biased in advance in favor of a particular type of stimulus so that this type of stimulus will win even when it is not the strongest. This top-down process is typically implemented following the instruction to perform a task in which some stimuli are relevant while others should be ignored. For example, the participant might be instructed to respond only to stimuli that appear on the left.

Behavioral and physiological studies show that the brain contains a relatively simple mechanism by which mutual interactions between the many competing stimuli ensure that only one or a few stimuli win the competition for the control of behavior and awareness. The idea is that each sensory channel inhibits all the others. This means that, as information passes up through the central nervous system (CNS), the stronger channels get stronger while the weaker channels get weaker, until only the strongest survives. It is this strongest survivor that determines our next action, for example, by moving our eyes towards the source of the stimulus. It is also this strongest survivor that enters conscious awareness. This is a bottom-up process, since the final outcome is determined solely by the sensory input: the bottom of the CNS. Through this mechanism our attention will be attracted by a bright flashing light or a loud noise, that is, a stimulus that is intense and unexpected. This is an automatic process over which we have little voluntary control. These very salient stimuli will capture our attention whether we like it or not.

But we are not just the slaves of our senses. We also have some voluntary control over our attention. We can deliberately focus our attention on one particular class of stimuli. Indeed, experiments on the focussing of attention have been a mainstay of cognitive psychology. In the covert attention paradigm, for example, participants are instructed to look straight ahead, but to focus their attention on, say, the left visual field and report when a target appears in that location (Posner et al.1982). Participants can also be asked to look out for a particular class of targets, for example, faces, rather than to focus on a particular location in space. To do such tasks we have to exert top-down control. In this case our behavior is not simply determined by the sensory input. Indeed, we have to inhibit bottom-up responses to stimuli that are highly salient, but irrelevant to the task we have set ourselves. The defining characteristics of top-down control are, in psychological terms, first, that we only respond to stimuli that are relevant to the task being performed, even if they are not the most salient; second, that this is a voluntary process that requires mental effort to be maintained. If our concentration lapses we will make mistakes and respond to the wrong stimulus.

The distinction between bottom-up and top-down control is also clear at the physiological level. Consider two paradigms used in early brain imaging studies of the visual system. In the studies of Zeki and his colleagues (Lueck et al. 1989) participants were shown classes of stimuli that differed on only one feature, such as color or motion. When colored stimuli were compared with black and white stimuli activity was seen in V4, the color area. This is an example of bottom-up processing. The participants simply and passively viewed the stimuli. The change in brain activity was caused by a change in the stimulus (from black and white to color). Shortly afterwards Corbetta and his colleagues (Corbetta et al. 1991) presented participants with the same stimulus array on all trials, but asked them to attend to different aspects of it: color on some trials, motion on others. In this experiment the participants had to actively attend to one feature rather than another. If the participants were attending to color rather than motion then there was more activity in V4. This is an example of top-down processing. The activity was not simply caused by the stimulus array, since the stimulus array was not different. It was only the participants’ focus of attention that changed.

2 Psychological and Physiological Definitions of Top-Down Control

At the physiological level bottom-up and top-down processes can be defined in terms of neural connections. Bottom-up processes are feed forward, for example, from primary visual cortex (V1) to visual association areas such as V4 and V5. Top-down processes are feedback, for example, from frontal or parietal cortex to sensory regions of the brain. Anatomically, the majority of reciprocal connections between cortical areas are symmetric with respect to the cortical layers in which connections originate and terminate. The axons in ascending pathways typically terminate in layer 4 of the cortex, while in descending or feedback pathways axons tend to avoid layer 4, terminating in layer 1 or layer 6. This has led to the hypothesis of an anatomical hierarchy, with some connections representing forward (ascending) pathways and their reciprocal counterparts representing feedback pathways (Felleman & Van Essen 1991). Top-down effects can be defined in terms of this anatomical hierarchy. However, the physiological definitions do not always map onto psychological definitions. Bottom-up in the psychological sense will always map onto feedforward connections, but the situation is more complex for top-down processes. Top-down processes in the psychological sense will map onto feedback connections, but it is possible to have feedback connections involved in a process that is bottom-up in psychological terms.

This was the case in the study of Macaluso and colleagues (2000). The participants' task was to detect visual targets that appeared on the left or on the right of fixation in an unpredictable manner. Tactile stimuli were also presented in the same location in space. These were irrelevant since their occurrence gave no information about where the visual targets would be. Nevertheless, when a tactile stimulus happened to occur in the same spatial location as a visual target, there was enhanced activity in extra-striate cortex (lingual gyrus). This is a uni-modal visual area which only receives feedforward signals from visual regions of the brain. The tactile stimulus can only have produced enhancement of activity in this area of the brain via a top-down, feedback signal (in the physiological sense) probably transmitted from multimodal areas in parietal cortex. This is not top-down in the psychological sense because the effect was automatic and entirely driven by the tactile stimulus. The effect occurs even if we are not concentrating on the tactile stimuli. As we shall see later, top-down control in the psychological sense is probably associated with feedback signals originating in frontal cortex.

The concept of top-down control in selective attention captures a key aspect of free will: the requirement that the choice of what to attend to is not solely determined by the stimuli that impinge upon us. Instead the choice of what to attend to is intrinsic to the person doing the attending. We have psychological markers that define tasks that require top-down control and we know that, at the physiological level, feedback connections are necessary, but not sufficient, for top-down control.

3 Free Will in the Brain: Where Is the Top in Top-Down Control?

Selective attention is not the ideal domain for studying free will since what participants do in experiments on covert attention cannot be directly observed. There is no overt behavior associated with attending to one thing rather than another. However, the same ideas can readily be applied to situations in which people make overt responses. For example, if I lift my right forefinger because it is pricked with a pin, that would be an example of bottom-up control of the finger lifting response. On the other hand, in order to lift my left forefinger every time my right forefinger was touched (unless I had a great deal of practice) I would have to exert top-down control. These two routes to the control of action are illustrated in figures 12.1 and 12.2.

The distinction between these two ways of controlling action are reflected in the central nervous system. When external cues are available for the selection of an appropriate action the lateral premotor cortex is engaged. In contrast, the internal selection of actions engages medial regions, in particular medial premotor cortex (supplementary motor area) and anterior cingulate cortex (Mueller et al., 2007; Passingham 1987).

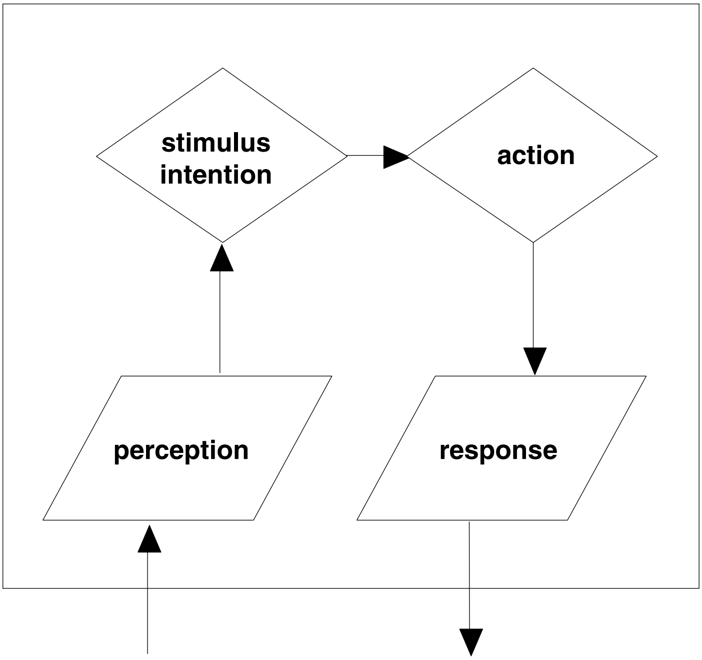

Fig. 12.1. Stimulus-driven action (bottom-up control): The response made is determined by the stimulus, the stimulus intention.

Top-down control, in the psychological sense, is defined as being voluntary rather than automatic. However, in the experiments I have referred to so far, the participants were simply doing what they were told to do by the experimenter, for example, "attend to the faces on the right." Their behavior was voluntary only in that they agreed to take part in the experiment and they continued to follow the instructions of the experimenter for the duration of the experiment. To study truly willed behavior the participants must decide for themselves what to do rather than simply follow instructions. This has been achieved in an experimental setting by allowing participants to make their own choices about what to do. But these choices must be selected from a very limited set of possibilities. For example, Libet and colleagues instructed participants to lift the index finger of their right hand, but gave them the freedom to make this response "whenever they felt the urge" (Libet et al. 1983). They could choose when to make the response, but not which response to make. In my own study of will (Frith et al. 1991) the participants had to lift their left or right index finger in response to a cue, but were free to choose which finger they would lift on each trial. They could choose which of two responses to make, but not when to make the response. In these experiments on "willed action," the requirement to choose the response or the time reliably activates the dorsolateral prefrontal cortex and anterior cingulate cortex in addition to the regions activated when subjects simply respond to cues (Jahanshahi & Frith 1998).

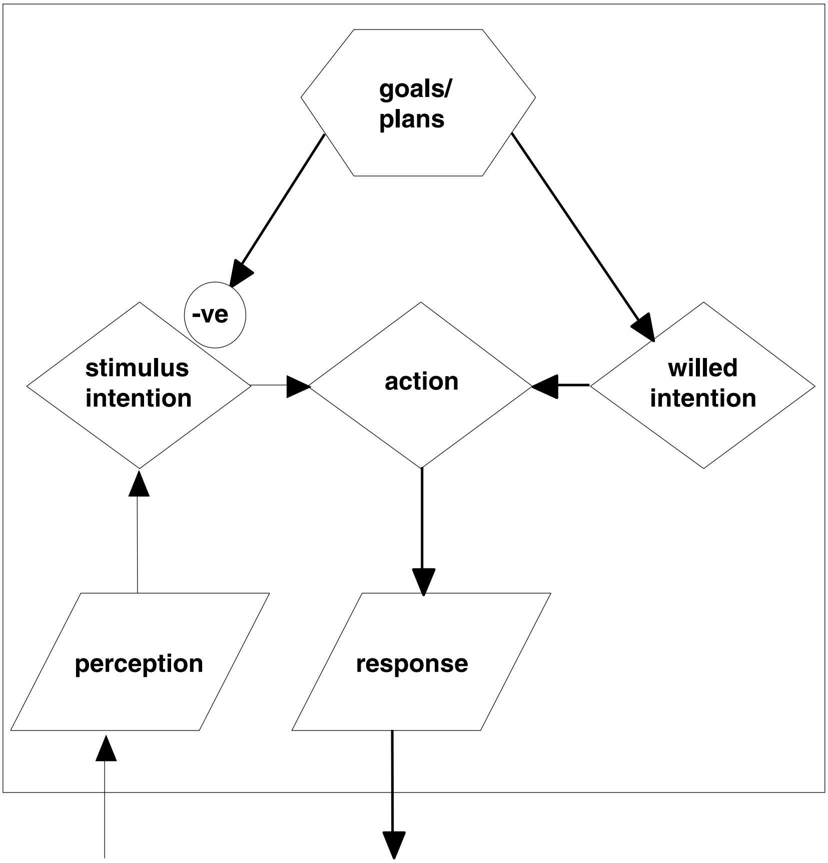

Fig. 12.2. Willed action (top-down control): The action is determined by goals and plans (willed intentions). Stimulus intentions are over-ridden. (The abbreviation "-ve" for "negative" implies inhibition.)

The term *top-down control* and the associated diagram shown in figure 12.2, imply that there is something at the top of this control process. There is a location or a system in the brain from which intrinsic control ultimately derives. Do our brain imaging results show that dorsolateral prefrontal cortex and/or the anterior cingulate are at the top, the source of top-down signals, and therefore the location of will in the brain? This conclusion is not entirely far fetched. The prefrontal cortex is the region of the brain that is most developed in humans in comparison to other primates (Semendeferi et al. 2001). Extensive damage to this region can lead to a syndrome in which the patient seems to be a slave to environmental stimuli, simply responding in a stereotyped manner to whatever he finds in front of him. One such patient quite inappropriately got undressed and climbed into bed, simply because he saw a bed with the cover turned down (Lhermitte 1983). The behavior of such a patient seems to be controlled solely by bottom-up processes. Such a patient can reasonably be described as lacking free will. In contrast, patients with lesions to anterior cingulate cortex may manifest a form of akinesia in which they fail to initiate any spontaneous voluntary actions (Tibbetts 2001). On recovery such patients report that there was nothing they really wanted to do. This can also be understood as a lack of will.

4 Action Initiation and Action Selection

There are two critical voluntary processes involved in these willed action tasks that depend upon prefrontal cortex. First, the participant has to select which of many possible actions to perform. Second, she has to turn this intention into an action, such as lifting her finger. How can her mental state cause a physical movement? Let us first consider this second problem. We do not find the same difficulty in understanding the bottom-up control of action, in which the action is initiated by a physical stimulus. In this case the cause and the effect are both in the physical domain. The energy in a stimulus activates a sense organ and this energy initiates the transmission of information through a series of neurons until the appropriate muscles are activated, again using physical energy. I suggest that the same process occurs with willed actions. The role of dorsolateral prefrontal cortex is not in initiating an action, but in selecting an action from among various alternatives. Without a prefrontal cortex we simply perform the actions automatically elicited by the stimuli around us: picking up the glass, putting on the spectacles, climbing into the bed (Lhermitte 1983). With an intact prefrontal cortex we can "sculpt this response space," inhibiting all the actions that are not appropriate in the current context (Frith 2000). In the extreme case we can inhibit all but one action. This is what we can achieve using our will. But this one remaining action is initiated, not by an act of will, but automatically by some stimulus in the environment. Top-down processes constrain actions, but they do not need to initiate them.

However, is the process of selecting appropriate actions necessarily at the top of the control process and, therefore, the key to free will in the brain? We explored this idea by asking subjects to make internally generated response selections (random number generation) at different rates (Jahanshahi et al. 2000). As the rate increased activity in dorsolateral prefrontal cortex also increased. But when the rate reached about 1 number per second, the task became too difficult and participants produced sequences that were less random (i.e., adjacent pairs of numbers; 1–2, 7–8, etc.). This failure occurs because of the conflict between the need to respond quickly and the need for more time to reject inappropriate numbers. When such conflict occurs there is a need for a higher-level control system to come into play in order to set the appropriate priorities. In terms of this analysis dorsolateral prefrontal cortex is clearly not at the top of the control system. I reach this conclusion because, when randomness began to fail at high rates, activity in dorsolateral prefrontal cortex declined, presumably because response selection had to be switched off in order to make the next response in time. The only region that showed an increase of activity when the rate became too high for randomness was the anterior cingulate cortex. So can we conclude that anterior cingulate cortex is at the top of a hierarchy of top down control in the brain?

I have characterized willed action tasks as tasks in which the participant has to decide for herself which response she is going to make. One criticism of this approach is that the choices involved in these tasks are so very trivial, lifting the left or the right index finger, for example. Would we get the same results if people were making more important decisions? There are a number of more recent studies in which people have to make less trivial decisions. Participants have been presented with moral dilemmas and asked to choose the appropriate course of action (Greene & Haidt 2002). For example, they have been confronted with the classic question in moral philosophy concerning whether it is right to take one life in order to save five. The problem for these experiments is that the participants are not directly confronting the dilemmas. They are merely indicating what they think they should to do in such situations.

In contrast, real choices have to be made in the many experiments investigating economic interactions involving trust and reciprocity. Here the choices made by the participants will directly determine how much monetary reward they will obtain. Decisions in such interactions are influenced by concepts such as trust, fairness, and altruism (Fehr & Camerer 2007). Although the content of the decisions is very different in these various moral and economic studies, the neural systems implicated are very similar (Greene et al. 2004; Sanfey & Chang 2008). There are essentially two competing systems involved in such decision making: a largely automatic emotional system, which we might call bottom-up, and a cognitive control system, often associated with reasoning, which we might call top-down. Just as in the very simple willed action tasks I discussed above, this top-down control system involves dorsolateral prefrontal cortex and anterior cingulate cortex.

So, even with more realistic and important choices, these seem to be the areas that are at the top in the brain's top-down control system. Even so, I remain unconvinced that we have identified where will is in the brain. There are at least two remaining problems for identifying the top in top-down control. The first concerns the physiology of the system. The diagram shown in figure 12.1, while very simplistic, is nevertheless supposed to be a realistic flow diagram which works equally well at the psychological and the physiological level. All the boxes have arrows indicating inputs and outputs. The one exception is the box at the top labeled goals/plans. As is appropriate for the module at the top of the hierarchy of control, this box only has outputs. Nothing must control the controller. The experiments I have reviewed above suggest that this box could also be labeled dorsolateral prefrontal cortex/anterior cingulated cortex. The problem is that there are no brain areas that have only outputs and no inputs (Semir Zeki, personal communication). So, on these grounds the search for the top of the system must fail.

So is my search for the top in top-down control in the brain ill conceived? Where is the top of top-down control if it is not located in the brain? In the final section of this chapter I will consider the role of social context in experiments on willed action.

5 Will as a Social Endeavor

Behavior in the experiments on moral dilemmas and those on trust and reciprocity has an obvious social component. Our response to moral dilemmas and the degree to which we trust others are both affected in the long term by culture and upbringing (e.g., Morewedge & Clear 2008). In the short term our behavior in such experiments is strongly affected by the presence of an audience. We are more altruistic and more moralistic when we believe we are being observed (e.g., Kurzban et al. 2007). In these situations there are strong cultural constraints on the exercise of free will. At first sight, such constraints are not so obvious in the willed action experiments in which people simply choose when to lift their finger. Are these situations in which we can look at free choice unconstrained by culture and social pressures?

If we analyze these situations more closely we find that the participant is not quite so free (Roepstorff & Frith 2004). For example, in Libet's experiment a participant will know that the excuse that he never had the urge to lift his finger would not be acceptable. Once he has agreed to take part in a willed action experiment, the instruction to "respond whenever you have the urge to do so" has to be interpreted very carefully. It would not appropriate to respond only a couple of times or to respond regularly every second. Likewise the instruction "to press which ever button you like" must also be carefully interpreted. It would clearly not be appropriate to choose the same response on every trial. This instruction must mean something along the lines of: "behave as if you were a free agent, choosing your responses in such a way that I cannot easily predict what you are going to do next." In other words the participant is setting himself the complex and highly constrained task of trying to behave like a free agent. The best way to do this is to try to produce a sequence that is fairly random in terms of choice and timing.

So what happens if we explicitly ask the participant to produce a random sequence? As with the willed action tasks, activity is observed in dorsolateral prefrontal cortex. Generating a random sequence is something we find very difficult. We cannot make responses truly at random. Instead we have to think in terms of patterns of responding and try to avoid making the obvious patterns: not choosing the same response every time, avoiding simple alternations, avoiding double alternations, etc. To generate random response sequences we have to apply complex constraints during response selection. It is for the purpose of applying such constraints that dorsolateral prefrontal cortex is recruited (Jahanshahi et al. 2000). Dorsolateral prefrontal cortex is associated with willed actions because this brain region constrains the actions we select. By an act of will we can elect to do one thing rather than another. We can inhibit our selfish instincts and resist the temptation of an immediate reward (Knoch & Fehr 2007). But we can also inhibit our instinct to share and justify this self-interested behavior to ourselves (Valdesolo & DeSteno 2008). If we consider will in relation to an isolated individual, then indeed will can be located in prefrontal cortex. This is the region of the brain that enables us not simply to be slaves to bottom-up processes.

However, when we consider the implicit agreement between the subject and the experimenter in these simple willed action tasks, we see that the constraints by which the subject chooses the appropriate responses are imposed by the social context. The constraints on action applied by dorsolateral prefrontal cortex are determined by the social context and derive from the inputs from brain regions concerned with the analysis of social context. In a willed action task our subject is obeying the instructions of the experimenter to behave like a free agent. She is certainly exerting top-down control, but are we justified in considering this to be an example of free will?

My analysis of the willed action tasks has demonstrated the importance of the interaction between the experimenter and the participant. The production of the willed responses by the participant derives from an implicit agreement between the experimenter and the participant about what is the purpose of the experiment (Jack & Roepstorff 2002). Once we consider the participant embedded in the social setting, rather than in isolation, we see that will emerges from this social interaction. The top-down constraints that permit acts of will come from outside the individual brain. Through the interactions of many brains, humans create the culture from which higher cognitive functions, including will and consciousness, emerge. If we are to understand the neural basis of free will, we must take into the account the brain mechanisms that allow minds to interact.

References

Corbetta, M., Miezin, F.M., Dobmeyer, S., Shulman, G.L., Petersen, S.E.: Selective and divided attention during visual discriminations of shape, color, and speed: Functional anatomy by positron emission tomography. Journal of Neuroscience 11, 2383–2402 (1991)

Desimone, R., Duncan, J.: Neural mechanisms of selective visual attention. Annual Review of Neuroscience 18, 193–222 (1995)

Fehr, E., Camerer, C.F.: Social neuroeconomics: The neural circuitry of social preferences. Trends in Cognitive Sciences 11, 419–427 (2007)

Felleman, D.J., Van Essen, D.C.: Distributed hierarchical processing in the primate cerebral cortex. Cerebral Cortex 1, 1–47 (1991)

Frith, C.D.: The role of dorsolateral prefrontal cortex in the selection of action as revealed by functional imaging. In: Monsell, S., Driver, J. (eds.) Control of cognitive processes, pp. 549–565. MIT Press, Cambridge (2000)

Frith, C.D., Friston, K., Liddle, P.F., Frackowiak, R.S.: Willed action and the prefrontal cortex in man: A study with PET. Proceedings: Biological Sciences/The Royal Society 244, 241–246 (1991)

Greene, J.D., Haidt, J.: How (and where) does moral judgment work? Trends in Cognitive Sciences 6, 517–523 (2002)

Greene, J.D., Nystrom, L.E., Engell, A.D., Darley, J.M., Cohen, J.D.: The neural bases of cognitive conflict and control in moral judgment. Neuron 44, 389–400 (2004)

Itti, L., Rees, R., Tsotsos, J.K. (eds.): Neurobiology of attention. Academic Press, Burlington (2005)

Jack, A.I., Roepstorff, A.: Introspection and cognitive brain mapping: From stimulus-response to script-report. Trends in Cognitive Sciences 6, 333–339 (2002)

Jahanshahi, M., Frith, C.D.: Willed action and its impairments. Cognitive Neuropsychology 15, 483–533 (1998)

Jahanshahi, M., Dirnberger, G., Fuller, R., Frith, C.D.: The role of the dorsolateral prefrontal cortex in random number generation: A study with positron emission tomography. NeuroImage 12, 713–725 (2000)

Knoch, D., Fehr, E.: Resisting the power of temptations: The right prefrontal cortex and self-control. Annals of the New York Academy of Sciences 1104, 123–134 (2007)

Kurzban, R., DeScioli, P., O'Brien, E.: Audience effects on moralistic punishment. Evolution and Human Behavior 28, 75–84 (2007)

Lhermitte, F.: 'Utilization behaviour' and its relation to lesions of the frontal lobes. Brain: A Journal of Neurology 106(Pt. 2), 237–255 (1983)

Libet, B., Gleason, C.A., Wright, E.W., Pearl, D.K.: Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential): The unconscious initiation of a freely voluntary act. Brain: A Journal of Neurology 106, 623–642 (1983)

Lueck, C.J., Zeki, S., Friston, K.J., Deiber, M.P., Cope, P., Cunningham, V.J., Lammertsma, A.A., Kennard, C., Frackowiak, R.S.: The colour centre in the cerebral cortex of man. Nature 340, 386–389 (1989)

Macaluso, E., Frith, C.D., Driver, J.: Modulation of human visual cortex by crossmodal spatial attention. Science 289, 1206–1208 (2000)

Morewedge, C.K., Clear, M.E.: Anthropomorphic god concepts engender moral judgment. Social Cognition 26, 182–189 (2008)

Mueller, V.A., Brass, M., Waszak, F., Prinz, W.: The role of the preSMA and the rostral cingulate zone in internally selected actions. NeuroImage 37, 1354–1361 (2007)

Passingham, R.E.: Two cortical systems for directing movement. In: Motor areas of the cerebral cortex, Ciba Foundation Symposium, vol. 132, pp. 151–161. Wiley, Chichester (1987)

Posner, M.I., Cohen, Y., Rafal, R.D.: Neural systems control of spatial orienting. Philosophical Transactions of the Royal Society of London: Series B, Biological Sciences 298, 187–198 (1982)

Roepstorff, A., Frith, C.: What's at the top in the top-down control of action? Script-sharing and 'top-top' control of action in cognitive experiments. Psychology Research 68, 189–198 (2004)

Sanfey, A.G., Chang, L.J.: Multiple systems in decision making. Annals of the New York Academy of Sciences 1128, 53–62 (2008)

Semendeferi, K., Armstrong, E., Schleicher, A., Zilles, K., Van Hoesen, G.W.: Prefrontal cortex in humans and apes: A comparative study of area 10. American Journal of Physical Anthropology 114, 224–241 (2001)

Tibbetts, P.E.: The anterior cingulate cortex, akinetic mutism, and human volition. Brain and Mind 2, 323–341 (2001)

Valdesolo, P., DeSteno, D.: The duality of virtue: Deconstructing the moral hypocrite. Journal of Experimental Social Psychology 44, 1334–1338 (2008)

13

Thinking beyond the *Bereitschaftspotential*: Consciousness of Self and Others as a Necessary Condition for Change

Sean A. Spence

University of Sheffield
Academic Clinical Psychiatry
The Longley Centre
Norwood Grange Drive
Sheffield S5 7JT
United Kingdom
s.a.spence@sheffield.ac.uk

Summary. The electroencephalographic (EEG) studies of Benjamin Libet and colleagues, published in the early 1980s, served to focus scientific and philosophical attention upon the processing constraints of the human brain, with respect to the question of how much (or how little) a human actor could be said to know about the genesis of their own acts, in real time. If taken at face value, Libet's findings (and those of others) seem to radically constrain the extent to which any actor may be said to be the "author" of his or her own voluntary acts, in the short-term present. Hence, there is a potential problem for traditional accounts of human agency and moral responsibility (i.e., if we learn of our intentions-to-act only after action-initiation has commenced, can we really be held responsible for what we have done?). However, such a problem is susceptible to solution if we adopt a longer-term perspective, one that focuses upon the "meanings" of acts for their agents and the latter's pursuit of certain states of consciousness. For although consciousness does not initiate action, conscious states nevertheless provide the motive for much of what it is that humans "do" (for good or ill). An organism lacking consciousness would fail to constitute a moral agent; an unconscious being would be incapable of "sin." Furthermore, an overly simplistic interpretation of Libet's findings faces stern tests in certain areas of psychiatric and forensic practice. While we continue to uphold a legal distinction between murder and manslaughter it is highly likely that we shall imbue consciousness with some (long-term) influence over voluntary acts. Finally, conscious awareness of our "selves," our patterns of behavior (e.g., our habits), and our effects upon others provides us with necessary data, should we wish to change our behaviors, our characters, in the future.

Keywords: *Bereitschaftspotential,* electroencephalography, EEG, neural adequacy, obsessive compulsive disorder, hysterical conversion disorder, sin, consciousness, meta-responsibility.

> "… everything we do belongs to a world that we have not created."
> Thomas Nagel (1982, first published 1979)

There is a problem at the interface of human volition and moral responsibility that is perhaps most tellingly exposed in some areas of psychiatric practice. Attempting to articulate this problem and to seek its resolution, may help us to relocate the altruistic act at the heart of psychological medicine. That is the aim of this essay.

1 The Problem

On the one hand, since the pivotal findings of Benjamin Libet and colleagues (1983), demonstrating that electrical activity in the brain predictive of voluntary behavior precedes not only manifest movement but also the *conscious intention* to move, it has seemed that free will is radically subverted at just that moment when it might have been regarded as being most likely to effect change: in those hundreds of milliseconds immediately preceding spontaneous voluntary action. *Post*-Libet, how can we defend "free will" without resorting to speculation that freedom must be enacted *unconsciously,* that is, before its author is aware of her own authorship (Spence 1996)?

On the other hand, much of day-to-day life, social interaction, ideas of responsibility and culpability hinge upon our notions of what people can reasonably be said to have known that they were doing and, by inference, what they might reasonably have *prevented* themselves from doing. Indeed, while certain legal problems might simply "go away" if we wished solely to identify the organism that performed a certain behavior at a certain time in a certain place, usually this is not good enough: what we really wish to know is whether that organism was "acting" as an "agent," that is, whether he "intended" to do what he did (Macmurray 1991).

Our problem then, is to reconcile our current recognition of curtailed agency at the point of action (*post*-Libet) with our strong intuition and phenomenological experience that we (and others) are free.

2 Are We Responsible but Not in Control?

> You have heard that it was said, "You shall not commit adultery." But I say to you that every one who looks at a woman lustfully has already committed adultery with her in his heart. (Matthew 5:27–28)

This Biblical statement seems rather unfair if we understand Jesus to be critiquing automatic responses, thoughts, and temptations, ideas that appear in the mind unbidden. For, if (*post*-Libet) we cannot claim authorship of our physical acts, are we any more likely to "own" the thoughts arising in our heads? Can I help what I am thinking? One might compromise and suggest that the thoughts referred to by Jesus are more likely to be those that are "intentionally" ruminated upon, the fantasies we might occasionally entertain, so that the postulated inner conflict is more akin to that occurring within the patient who suffers from obsessive compulsive disorder: trying *not* to think of certain thoughts that strike him as morally unacceptable. Hence, in one sense, the obsessive identifies a property common to us all: we have to be *conscious* to be immoral. It seems clear that consciousness is instrumental to each of those states of mind associated with what the religious might call "sin"; indeed, it is hard to imagine that an unconscious automaton or zombie (so beloved of philosophers of mind) could be guilty of lust, pride, or avarice; we seem to "require" consciousness to be "able" to "sin." "Sinning" becomes nonsensical if the sinner herself is unaware, unconscious. So, without consciousness we cease to be moral agents.

However, my resorting to quotation marks so frequently in the preceding paragraph flags up the problems involved in trying to address moral matters while simultaneously attempting to keep in mind Libet's findings: if we accept his data and those that have followed with similar (though clarifying) results (e.g., the work of Patrick Haggard and others) must we not regard all ethical, intentional states of mind as somehow conditional, illusory, a pragmatic shorthand borrowed from "folk psychology," because we can no longer defend the *authentic* authorship of thoughts and deeds? Can authenticity only be salvaged if conscious intentions *precede* actions (rather than vice versa; Libet et al. 1983)? Now the problem is that we arrive at rather an incoherent, stymied view of life, hostage to an idea that all authorship is illusory (because authorship occurs prior to the author's conscious awareness). Hence, we risk rehearsing the problem of patients with schizophrenia who experience thought insertion and thus believe that their thoughts are not their own. As Chris Frith intimated in some of his earlier papers (e.g., 1987), in a sense, these patients are recognizing something that may (again) apply to all of us: we are not consciously generating thoughts but instead *experiencing* them. (And if Libet's earlier work on neuronal "adequacy" is to be accepted [see Libet 2004, for a review], then we experience thoughts and sensations only *after* a sufficient, finite period of neuronal activity has occurred [indeed, it is this finding that serves to undermine Libet's own proposition that consciousness acts as free will's power of "veto"; for a conscious veto would be, itself, the product of *preceding* unconscious activity, and therefore *caused* not causal; Spence 1996].)

So, our problem is not merely a scientific one: how to account for "voluntary" behavior that arises out of (prior to) awareness. Our problem is also a moral one: how to retain coherence in our day-to-day lives and to hold on to a sense of responsibility, bearing in mind the true extent of what one human being can do to another (e.g., Waller 2002).

3 Respecting Our Automatisms

If we cannot claim authorship in the short term, over the milliseconds preceding action, and in those situations where we respond instantaneously to our environment, might we nevertheless retain some other form of responsibility, some rather more custodial stewardship over those things that we *might* do in a given, future situation? If so, should we be anticipating future difficulties, rehearsing what we might do and thereby protecting ourselves from our own precipitated "automatisms" (please note, I use the term loosely here to describe reflexes, unthinking responses, habits, and also to infer what actions become if we apply Libet's findings universally)? Thus, our problem becomes one of "meta-responsibility": our moral accountability for our own future situations and behaviors (see Spence 2008).

Consider what happens if we ignore these responsibilities. The drunken driver may not have wished (consciously) to have driven across the pavement and killed the small child, but he might nevertheless have anticipated that his being inebriated at the wheel of a car would precipitate some form of mishap. Similarly, the youths who are intoxicated on a Saturday night might not have planned to beat up the boy they met in the alleyway as they emerged from the club, but they might have anticipated causing some sort of trouble through their use of disinhibiting substances and their choice of companions when they went out that night. (Actually, in forensic settings there tends to be a lesser penalty afforded such "reactive," "spontaneous," or even "provoked" violence, than that applied to violence that is "instrumental" or "premeditated": again society places a premium upon conscious awareness and planning, *ahead of the event*.)

Furthermore, don't we all, routinely, experience and exhibit some awareness of "what we are like"? The future is not totally unknown or unpredictable to us. Would we go to the bakery if we did not wish to buy bread? Would we go to the cinema if we thought we would not enjoy the main feature?

Nevertheless, there may be some occasions when we do, genuinely, surprise ourselves:

> "I did not know I loved you till I heard myself telling you so – for one instant I thought 'Good God, what have I said?' and then I knew it was the truth," Bertrand Russell to Lady Ottoline. (Cited in Dennett 1991, p. 246)

Such is his distance from his agency, that when Russell tells his friend that he loves her we might (again) resort to quotation marks: "he" said "he" loved her but "he" seems to be saying that "he" didn't realize "he" was about to say what "he" said; he (which we may interpret as meaning his conscious mind) identifies the utterance as some form of unconscious product but he also identifies *with* it ("I knew it was the truth"). Hence, at least in Russell, some form of conscious self identifies with and accepts what its own unconsciousness seems to be saying; then he abides by it.

The point is that although we may not "author" actions in the conscious way that we experience them, phenomenologically (because action initiation has occurred prior to our awareness), we nevertheless live in some longer-term relationship with our behaviors and propensities and, if we are mindful of ourselves, we recognize the patterns emerging.

However, it might be argued that there is still something rather "unfair" at work here: we are the authors of actions that contribute to our fates yet (*post*-Libet) "we" cannot claim to have consciously authored them (in the very short term, at least). "We" then have to live with the consequences of these unthinking (unthought, unplanned, unconsciously initiated) behaviors. So, are "we" really victims?

4 Blaming People and Passing Judgment

> It is concluded that cerebral initiation of a spontaneous, freely voluntary act can begin unconsciously, that is, before there is any (at least recallable) subjective awareness that a "decision" to act has already been initiated cerebrally. This introduces certain constraints on the potentiality for conscious initiation and control of voluntary acts. (Libet et al. 1983, p. 623, noting the emergence of the *Bereitschaftspotential* prior to conscious intention)

As we have seen, the impact of Libet's work has been to radically subvert the short-term authorship of voluntary action: the "agent" owns an act that began prior to "her" awareness of its authorship. Yet, the EEG signal that Libet identified as constitutive of that moment of generation of movement prior to awareness (the *Bereitschaftspotential*) has elsewhere been seen as a potentially incriminating factor in those patients who exhibit "functional," psychogenic movement disorders:

> Terada *et al.* demonstrated that in five out of six patients with [psychogenic/hysterical] myoclonus, a *Bereitschaftspotential,* indicative of voluntary causation, preceded abnormal movements. 'Therefore, it is most likely that the jerks in these patients were generated through the mechanisms common to those underlying voluntary movement.' (Spence 2006a, p. 227, citing Terada et al. 1995)

Such "hysterical" patients present the doctor with the quasi-psychic task of determining whether they are "unconsciously" generating "psychogenic" symptoms or "consciously" malingering – "acting" them (Spence 1999). It is hard to justify such a distinction on purely phenomenological grounds, but if one takes Libet's findings seriously then the task becomes even more problematic. For, if we accept that the "healthy" subject who moves her limb exhibits a *Bereitschaftspotential,* which emerges *before* her intention to act, then what significance should we attribute to the discovery that many patients performing psychogenic movements evince the same signal? What is the significance of this wave (the *Bereitschaftspotential*), which simultaneously serves to subvert the agency of the healthy subject (*post*-Libet) while implicating the voluntary involvement of the "psychogenic" in their symptom (in Terada et al. 1995)? How can "proof" of unconscious agency in one context constitute "proof" of voluntary agency in another? (If nothing else, this suggests that we should be very careful in drawing causal inferences, and attributing blame to others; clinicians might benefit from pausing for uncertainty; see table 13.1.)

Table 13.1. The significance afforded the *Bereitschaftspotential,* according to context

Context	Significance
The work of Libet et al. (1983)	The onset of the *Bereitschaftspotential* preceding the conscious intention to act is interpreted as evidence for *unconscious* initiation of activity.
Terada et al. (1995)	The appearance of *Bereitschaftspotentials* prior to psychogenic movements is interpreted as evidence of *voluntary* causation.
A mentally ill offender who stabs another	Would the presence of the *Bereitschaftspotential* affect the verdict?
A "saint" who walks into persecution, with negative consequences for herself	Would a *Bereitschaftspotential* make her actions "more" or "less" saintly?

What can any person be said to "know" of any action they perform? We cannot articulate an awareness of motor units in our motor cortices or neuronal signals traversing our spinal cords. We are always speaking "at a distance" from the raw mechanics of our actions. Yet, in the psychogenic patient, we seem to be looking for just such a form of awareness:

> [T]he difference between the malingerer and the [psychogenic] patient who sincerely acts the illness role may be less a matter of the latter's relative honesty than his relative lack of insight. (Wenegrat 2001, p. 226)

For Wenegrat, it is as if he would ideally like the psychogenic patients to "own" their symptoms, to "admit" to their authorship. It is (reported) awareness (or its lack) that implicates the patients, not their EEG signal.

So, now consider a problem that faces psychiatrists in the forensic sphere: an organism is known to have performed an awful behavior, but the question is, did it know what it was doing (i.e., was it an agent?) and did it know what it was doing was wrong (i.e., was it a *moral* agent?)?

> The McNaughton Rules: "To establish a defence on the grounds of insanity, it must be clearly proved that, at the time of the committing of the act, the party accused was labouring under such a defect of reason, from disease of the mind, as not to know the nature and quality of the act he was doing, or, if he did know it, he did not know he was doing what was wrong." (Cited in Gregory 2005, p. 254)

Hence, forensic psychiatrists are called upon to gauge the intentional stance of an agent, often at some remove from the events described. Did the deluded man believe the victim was the "devil" and, if he did, did he think it was all right to stab him? It feels to me that there are inevitably shades of grey here: we may intuit that the subject was mistaken and thereby absolved of guilt for "his" actions, but are we not (again) attempting to label or attribute significance to an action that might well have "behaved" normally, neurologically, rather like any of those described by Libet? To clarify: as the killer raised the knife did he exhibit a *Bereitschaftspotential,* and does it matter whether he did or not? (While a "truly" automatic, involuntary movement conveying the knife towards the victim might have been associated with an abnormal *Bereitschaftspotential,* thereby "exonerating" the perpetrator, on most occasions this will not be the case: the "act" will have been voluntary, it is the *reason* for the act with which we shall become concerned: "he did not know he was doing what was wrong.")

Hence, in both the psychogenic patient and the psychiatric "offender" what we seem to value is *awareness,* and it is an awareness of the *consequences of actions.* This suggests that the emphasis in our judgment and the location of what we value most in human responsibility is *not* that instant (those hundreds of milliseconds) immediately preceding "voluntary action" (a *Bereitschaftspotential* cannot be moral or immoral), it is instead more of a stance towards the world, an attitude which the subject adopts, and a consideration of consequences, for others, and for their own future "selves" that matters. Hence, the drunken driver and the brutal gang are guilty not only for what they *did* but for not *minding what they might do:* for being reckless with their own future selves.

As we navigate the world, "we" may not consciously initiate the electrical signals that trigger "our" actions, but there is a sense in which our very real phenomenological awareness, our consciousness, provides a field, a context, and a source for such action. If we do not care for others and we do not care for ourselves, then we will not care for those actions executed by components of ourselves (and our future selves): and we will constitute a particular kind of agent, an antisocial one at that.

Also, if and when we come to make reparation, there is another apparent disconnection: we are apologizing, or making amends, for self-states which preceded "us," for which we might retain little current sympathy; but nevertheless we must simply accept that "we" (as our conscious minds/awarenesses) are indeed answerable for all the rest (the habits, mannerisms, losses of control, distractions and failures of planning, the things we did in the past that we may now believe to be wrong). This might seem unfair, but then life's not fair!

> Saintliness presupposes free will – a conscious choice. (Yuli Schreider, cited in Luxmoore 2000, p. 709)

So then, what Schreider has to say about saintliness seems to constitute the mirror image of these states of responsibility (the good "field" displacing the bad), for again, the saint must walk into some very trying circumstance, fully aware that her current behaviors will have negative consequences for (her) future selves. However, unlike the drunk driver or the gang member, these consequences are acknowledged, and accepted, for some better, "higher" purpose, which while it might cost the self its life, is nevertheless "worth dying for." Again, we are not so much concerned with the *Bereitschaftspotential* of the "voluntary" act as the knowledge that the putative "saint" seems to have accepted its likely *consequence:* if she speaks truth unto power she will probably pay dearly for it, but she does so anyway.

5 Trying to Do the Right Thing

Hence, what we seem to be aiming at, in considering responsible volition, is the planning that arises over relatively long periods of time, during which the agent is fully conscious, aware of her situation in the world and of the likely outcomes of events in her future. In other words, it matters that a person can inhabit a mental, virtual space of potential outcomes and consequences (and that such a space is "conscious" because, again, a man who walks into the firing line without thinking is not saintly, merely reckless).

However, while brain centers that are superordinate in the executive hierarchy (e.g., dorsolateral prefrontal cortex) can be shown to project to "lower" centers and to modulate their outputs (e.g., Ganesan et al. 2005), the contents of phenomenological awareness exert an influence that seems less straightforward, and certainly rather more protracted in time. An agent may choose, over very long time scales indeed (hours, years), to rehearse certain behaviors in preference to others: "men become builders by building and lyre-players by playing the lyre; so too we become just by doing just acts, temperate by doing temperate acts, brave by doing brave acts" (Aristotle 1998, p. 29). Hence, these (rehearsed) behaviors can become the attribute that best defines their host. However, the contents of consciousness are themselves the products of neural activity, so while their role in our futures seems obvious (I go to the cinema to see the film of which I am thinking), nevertheless, my thinking is the product of an earlier (unconscious) stream of events. But there is something in my experience, the quality of what it feels like to have thoughts in my head, that then affects my future plans: if I don't like the thought of the film I won't go; if something better arises, I'll "change my mind." So consciousness does not cause action in the short term (milliseconds prior to action), but it certainly affects the course of my acts in the longer-term cycle of action (Spence 2006b; and see Burgess et al. 2007 and Koechlin & Hyafil 2007 for biological accounts of how anterior prefrontal cortex might support such virtual "futures"). We act *towards* conscious states (i.e., to produce them).

6 Becoming "Ourselves"

> What comes out of a man is what defiles a man. For from within, out of the heart of man, come evil thoughts, fornication, theft, murder, adultery, coveting, wickedness, deceit, licentiousness, envy, slander, pride, foolishness. All these evil things come from within, and they defile a man. (Mark 7:20–23)

On the one hand this statement suggests that it is our actions that will define us and, also, in a sense, "create us" (the circuits of our future behaviors are formed by our current habits; yes, we have wandered into the domain of brain imaging studies demonstrating the long-term changes arising in the brains of London taxi drivers and classical musicians; e.g., Maguire et al. 2000; Gaser & Schlaug 2003). However, consideration of this statement, in all its ramifications, also prompts another realization: that we are dependent for our character formation upon our early carers, those who were in a position to dictate or facilitate our early experiences and activities, those activities that later contributed towards and constitute our characters. In the brain of the musician, it wasn't solely the hours of practice and tuition that mattered, which changed his motor circuits; there was that person who sent him to study (rather than allowing him to play outside!). That person's influence is also evinced through those motor circuit changes. Similarly, in the life of a moral being there are beliefs acquired from those who were important to her, whose examples she witnessed, whose words she may remember. As the unfortunate examples of "wolf children" bear (counterfactual) witness, we are reliant upon other humans for our development as agents.

7 The Unknowing Altruism of Others

We don't choose our genetic endowment or our families and we only gradually learn about ourselves: certain things we might only learn in certain environments (I won't know how I interact with and treat people from other cultures unless I meet them); we are dependent upon others showing us what we are like. They also, perhaps unwittingly, hold us in check. When people, consciously or unconsciously, set limits upon our behavior they may be helping or hindering us. Consider some examples:

When I give a lecture, the motor behaviors I exhibit might easily be construed statistically: for instance, while it is highly probable that I shall stand near the podium, speak towards the microphone, point towards the screen, it is statistically highly improbable that I shall dance, sing, or recount my desire to ensure that yes, I definitely do own each of Jackie McLean's *Blue Note* recordings from the 1960s (the latter might appear highly inappropriate, even incomprehensible to some). Nevertheless, as if to prove that we have statistical presuppositions, one may attend certain lectures where the conduct of the speaker does make people uneasy. Is he wandering too near to the people in the front row? If he walks along an aisle is he going to make an example of someone? If he started to cry what would we all do? Our presence to each other holds us all in some kind of equilibrium, and that can be good for us as well as bad.

One emerging theme in the biology of psychopathy and antisocial personality disorder is the extent to which genetic endowment and early environment interact, how certain individuals may fail to learn from punishment or conditioning (because of a proposed absence of feeling, possibly implicating their amygdalae), and how the presence of at least one "good" relationship may be important in preventing an even worse future trajectory (I am drawing, here, on the work of Blair, Caspi, Moffitt, Pincus, Raine, Robbins, Rutter, and others.) One of the violent men I visit in the "community," who has suffered extreme things and done extreme things to others, remarked spontaneously to me one day: "I didn't choose to become what I've become." One way of understanding his problem, which invokes both biology (genetics) and early environment (abuse and genealogical uncertainty) is to say that through the absence of positive influences (that he was able to assimilate) he has arrived at the end of some very bad decisions, so that now it is difficult for him to go back (the product of former bad choices is a situation wherein it is very difficult to find a "good" choice now, in the present). Without a good-enough carer to "mirror" our behavior, and the good-enough neurology to assimilate feedback, we may be deprived of ethical development (a thought developed in Mullen 1992). To progress, we need help from others and we need help from within (do we have the "biological luck," the good fortune, to possess the requisite brain systems?).

8 Doing Things in Front of Others

> Strange things were said about [Charcot's] hold on the Salpetriere's hysterical young women and about happenings there.… [D]uring a patient's ball … a gong was inadvertently sounded, whereupon many hysterical women instantaneously fell into catalepsy and kept [the] plastic poses in which they found themselves when the gong was sounded. (Ellenberger 1994, p. 95)

If we return to the subject of psychogenic movement disorders, then part of the problem in interpreting Charcot's historical contribution is our uncertainty regarding the influence that this man's personality exerted upon others; did he "train" his patients to be dysfunctional; were patients (and indeed, their physicians) playing roles? This is Wenegrat's (2001) contention (above). However, this may be too harsh; it may be quite unfair to focus upon Charcot in this way. In the daily assessment of people exhibiting "functional" and "personality" disorders, much is made of whether their behaviors change in certain contexts: social influence is ubiquitous (and not confined to the Salpetriere). Does a tremor get worse when the doctor is passing by? Is an "overdose" consumed in full view of the nursing station? Does the patient anticipate his spouse's arrival as he ties the noose? Whether or not we describe events in these terms, much of neurological, psychiatric, and forensic practice involves implicit judgments concerning the contextual modulation of conduct by the presence of other human beings.

Of course, context may inhibit bad behavior as much as good: loss of "normal" feedback (and restraint) can isolate leaders and allow them to become tyrants (and see Owen 2006); footballers and rock stars can become grotesques in the absence of someone who will say "enough."

Just as we might conceive of our own behaviors through the representation of statistical probabilities, so we might conceive of others' influence upon us as capable of statistical impact: this group of friends will encourage me towards finer acts; those might facilitate lesser parts of me.

9 A Very Physical Redemption

So, what may we conclude from all that has gone before?

First, we might accept that the agency exerted within the short term over individual physical acts is indeed rather restricted following the discovery that our actions are initiated prior to our awareness of our own (immediate) intentions. So, this is a problem if we had really wished to "own" the typing movements that we make at a keyboard or the strokes of a pen when we write our signature. Much of what we do is (thankfully) automated, allowing us to be free from the constant monitoring of procedural events (movements). However, we have also seen that when we are really required to consider what it is that is important to us, when we have to answer for what we have done, it is not the muscular deformations of hands or tongues that constitute guilt. We are much more concerned with what we knew, what we believed, and what we cared about over relatively long periods of time.

Hence, we see that the contents of consciousness *influence* our behaviors, even if they do not initiate them. Consciousness is required to sin and to imitate saints. It is awareness and feelings that matter.

Furthermore, besides the limits constraining our short-term agency, we have other reasons to be humble, for not only is our agency less than it might have seemed (in the milliseconds preceding voluntary action), but any sincere consideration of our characters and those of others reveals multiple constraints exerted by our psychobiological environment: we are formed as a consequence of what others do for us (for good or ill) and what we can appreciate them doing for us (depending upon our biological "luck"). We are social beings. We exist at the interface of nature and nurture.

So it is that in psychiatry, we encounter the human consequences of a range of social, psychological, and biological perturbations which have impacted upon the humans whom we meet. As with the notion of capacity, there is likely to be a spectrum of "freedom," variously constrained in different contexts and company. The man with debilitating negative symptoms of schizophrenia may have very limited freedom to express; the manic woman may find her behavioral parameters episodically extended in uncharacteristic ways. And while we cannot predict whether the violent man will be provoked on a Saturday or a Sunday morning by a chance encounter on the street, we nevertheless intermittently share a space, a conscious present with the people that we see as patients: a moment for understanding, elucidation of past and future selves, a time to know what they (and we ourselves) care about. These are moments when we contribute to the other's "subject-in-process" (to borrow Julia Kristeva's phrase), to what it is that they might become (as they contribute to ours). We only gather this if we are fully in the room and if we are fully present "for the patient."

> Because the doctors cared, and because one of them still believed in me when I believed in nothing, I have survived to tell the tale. It is not only the doctors who perform hazardous operations or give life-saving drugs in obvious emergencies who hold the scales at times between life and death. To sit quietly in a consulting room and talk to

> someone would not appear to the general public as a heroic or dramatic thing to do. In medicine there are many different ways of saving lives. This is one of them. (Coate 1964)

In the right circumstances, and in the right company, conscious awareness is potentially redemptive; it tells us about ourselves, and it may tell us where we are going. It is not the instant of the act but its context that seems to matter. We have to take care of our automatisms.

References

Aristotle: The Nichomachean ethics. Oxford University Press, Oxford (1998)

Burgess, P.W., Gilbert, S.J., Dumontheil, I.: Function and localization within rostral prefrontal cortex (area 10). Philosophical Transactions of the Royal Society of London: Series B, Biological Sciences 362, 887–899 (2007)

Coate, M.: Beyond all reason. Constable, London (1964)

Dennett, D.C.: Consciousness explained. Little, Brown and Company, Boston (1991)

Ellenberger, H.F.: The discovery of the unconscious. Fontana, London (1994)

Frith, C.D.: The positive and negative symptoms of schizophrenia reflect impairment in the perception and initiation of action. Psychological Medicine 17, 631–648 (1987)

Ganesan, V., Green, R.D., Hunter, M.D., Wilkinson, I.D., Spence, S.A.: Expanding the response space in chronic schizophrenia: The relevance of left prefrontal cortex. NeuroImage 25, 952–957 (2005)

Gaser, C., Schlaug, G.: Brain structures differ between musicians and non-musicians. Journal of Neuroscience 23, 9240–9245 (2003)

Gregory, R.L. (ed.): Oxford companion to the mind, 2nd edn. Oxford University Press, Oxford (2005)

Koechlin, E., Hyafil, A.: Anterior prefrontal function and the limits of human decision-making. Science 318, 594–598 (2007)

Libet, B.: Mind time: The temporal factor in consciousness. Harvard University Press, Cambridge (2004)

Libet, B., Gleason, C.A., Wright, E.W., Pearl, D.K.: Time of conscious intention to act in relation to onset of cerebral activities (readiness-potential): The unconscious initiation of a freely voluntary act. Brain: A Journal of Neurology 106, 623–642 (1983)

Luxmoore, J.: The quiet saints of the gulag. The Tablet, 708–709 (May 27, 2000)

Macmurray, J.: The self as agent. Faber & Faber, London (1991)

Maguire, E.A., Gadian, D.G., Johnsrude, I.S., Good, C.D., Ashburner, J., Frackowiak, R.S.J., Frith, C.D.: Navigation-related structural change in the hippocampi of taxi drivers. PNAS 97, 4398–4403 (2000)

Mullen, P.E.: Psychopathy: A developmental disorder of ethical action. Criminal Behaviour and Mental Heath 2, 234–244 (1992)

Nagel, T.: Moral luck. In: Watson, G. (ed.) Free will, pp. 174–186. Oxford University Press, Oxford (1982) (first published 1979)

Owen, D.: Hubris and nemesis in heads of government. Journal of the Royal Society of Medicine 99, 548–551 (2006)

Spence, S.A.: Free will in the light of neuropsychiatry. Philosophy, Psychiatry & Psychology 3, 75–90 (1996)

Spence, S.A.: Hysterical paralyses as disorders of action. Cognitive Neuropsychiatry 4, 203–226 (1999)

Spence, S.A.: The cognitive executive is implicated in the maintenance of psychogenic movement disorders. In: Hallett, M., et al. (eds.) Psychogenic movement disorders, pp. 222–229. American Academy of Neurology, Philadelphia (2006a)

Spence, S.A.: The cycle of action. Journal of Consciousness Studies 13, 69–72 (2006b)

Spence, S.A.: Can pharmacology help enhance human morality? British Journal of Psychiatry 193, 179–180 (2008)

Spence, S.A.: The actor's brain: Exploring the cognitive neuroscience of free will. Oxford University Press, Oxford (in press) (forthcoming, June 2009)

Terada, K., Ikeda, A., Van Ness, P.C., Nagamine, T., Kaji, R., Kimura, J., Shibasaki, H.: Presence of Bereitschaftspotential preceding psychogenic myoclonus: Clinical application of jerk-locked back averaging. Journal of Neurology, Neurosurgery and Psychiatry 58, 745–747 (1995)

Waller, J.: Becoming evil: How ordinary people commit genocide and mass killing. Oxford University Press, Oxford (2002)

Wenegrat, B.: Theatre of disorder: Patients, doctors, and the construction of illness. Oxford University Press, Oxford (2001)

Part IV: Human Implications of the Debate

14

Criminal Responsibility, Free Will, and Neuroscience

David Hodgson

Supreme Court of New South Wales
Queens Square, Sydney
NSW 2119
Australia
raeda@tpg.com.au

Summary. This chapter identifies retributive and consequentialist purposes of the criminal law, and it outlines arguments that retribution should be abandoned, including arguments, based on philosophy and neuroscience, that free will and responsibility are illusions. The author suggests that there are good reasons to retain retribution, and identifies ways in which this might be supported, including compatibilist and libertarian views of free will. The author gives reasons for preferring libertarian views, and concludes by considering the role that neuroscience may be expected to play in the future development of the law.

Keywords: free will, responsibility, retribution, criminal law, neuroscience, justice, consequentialism, guilty mind, punishment, compatibilism, determinism, reasons, gestalt experiences, consciousness, folk psychology.

In recent time, a great deal has been written concerning the possible impact of neuroscience on the operation of the criminal law, particularly having regard to the implications of neuroscience for notions of free will and personal responsibility for conduct.

There are some who see the ongoing development of neuroscience as sounding a death knell for notions of free will and responsibility, and who welcome this (e.g., Green & Cohen 2004). They see it as promoting an approach to criminal behavior that is not distorted by outmoded views about the causes of human conduct, and that dispenses with primitive and inhumane ideas of retribution and vengeance.

There are others who see no conflict between the findings of neuroscience and our ideas about criminal responsibility (e.g., Morse 2000); and others again who do see a conflict and consider that erosion of belief in free will and responsibility would be highly damaging, and that these notions should at least be maintained as convenient fictions (e.g., Smilansky 2002).

A central issue in this debate is the question of *retribution*: do we punish people because they *deserve* it, or merely because to do so has the best consequences for the community? And if we do punish people because they deserve it, is this justified? Do offenders really deserve punishment, or is retribution inhumane because their criminal conduct is ultimately down to things wholly outside their control, particularly their genes and their upbringing and circumstances?

And this gives rise to further questions. Is it desirable or undesirable that ideas of desert and retribution operate in the criminal law? Do such ideas require and depend on a belief that people are truly responsible for what they do, and does this in turn depend on a belief that they have free will? Is it reasonable to continue to believe in responsibility and free will, or at least in some kind of "downward causation" operating in the brain? In the light of these considerations, how should we view what neuroscience is telling and will tell us about the operation of the brain and its relationship to conduct? And what role do we see for neuroscience in the ongoing development and application of the criminal law?

These are the questions I will be addressing in this chapter.

1 Retributive and Consequentialist Purposes

The criminal justice system in my country Australia, and also in other countries with similar legal systems, including the United Kingdom and the United States of America, serves two broad types of purposes, which are together considered as justifying the imposition of restraints or other detriments on offenders: retributive, backward-looking purposes, and consequentialist, forward-looking purposes. These purposes guide the development of the criminal law, and inform decisions as to when to impose such restraints or detriments and what they should be.

The former (retributive) purposes are based on the idea that a person who has acted criminally *deserves* to be punished for this conduct, and that it is *just* that appropriate detriment be inflicted on that person. And the idea that a person deserves punishment for criminal conduct presupposes that the person is truly *responsible* for it, and is not deprived of that responsibility because the conduct was the inevitable outcome of things outside the person's control, notably genes and environment (nature and nurture).

The latter (consequentialist) purposes involve no such ideas. They simply look to the *good consequences* that imposition of detriment on offenders may be expected to have, notably:

1. Demonstration that certain types of behavior are unacceptable, and deterrence of the criminal and of others from engaging in that behavior;
2. Restraint of the criminal from further crime during incarceration;

3. Reform of the criminal;
4. Placating victims and perhaps compensating them (although the latter may be considered a matter for civil rather than criminal law); and
5. Reassurance of the community that they are protected and that criminals will be punished (promoting confidence in security and in the rule of law, and discouraging self-help).

The interplay of these two types of purposes can be seen in two areas of criminal law. First, in the general principle that for a person to be convicted of a criminal offence, the prosecution must prove not merely a guilty act, but also a *guilty mind*; and in exceptions to that principle. And second, in the principles applied in determining what *punishment* is appropriate.

As regards the requirement of a guilty mind, criminal liability generally requires that the action in breach of the law be willed or done voluntarily, in circumstances where the offender was responsible for this. Generally this will be presumed, when it is shown that the act was done by a person over a certain age (in Australia, 14) who was apparently conscious, and when there is no evidence that what was done was done under duress or in self-defense or under a mistaken belief that facts existed that would have made the act innocent. If there is evidence of any of these things, then generally the prosecution must exclude them beyond reasonable doubt. If the person is between the ages of 10 and 14 (in Australia), the prosecution has to prove the person knew the act was seriously wrong, not merely mischievous; and if the person is under the age of 10, then there is (in Australia) no criminal liability.

Responsibility can also be challenged by evidence of mental abnormality. There may be evidence of insanity within the rules established by *M'Naghten's Case*, namely, that by reason of "a defect of reason, from disease of the mind" the person did not know what he or she was doing or did not know it was wrong. Or there may be evidence that for some other reason the action was not conscious and voluntary, the so-called defense of sane automatism (discussed by the High Court of Australia in *R v Falconer* 1990). Evidence of mental abnormality that does not *exclude* responsibility in either of these ways, but is considered merely to *diminish* responsibility, is generally not relevant to whether a person is guilty of a crime, although in Australia a finding of substantially diminished responsibility can reduce murder to manslaughter. Otherwise, it can be relevant only to the amount of punishment that is imposed, to which I will come.

There are also a limited number of offences of strict or absolute liability, where the requirement of a guilty mind is reduced or absent, because consequentialist considerations are considered sufficient to justify placing an onus on citizens to make quite sure that some adverse event does not occur. Examples in Australian law are offences associated with polluting the environment, and offences associated with failure by employers to provide a safe system of work for their employees.

As regards punishment, in Australian law there is an overriding principle stated by the High Court of Australia in the case of *Veen v The Queen (No. 2)* (1988), to the effect that in no case should the punishment exceed what is proportionate to the criminality involved in the offence. This is an important application of the retributive purpose of punishment, excluding the possibility that consequentialist considerations, such as deterrence or protection of the community by longer detention of the offender, be permitted to justify more being imposed on an offender than the offending conduct deserves.

Retributive considerations may operate in addition to *reduce* the penalty from the *Veen* maximum, for example if there is evidence of mental abnormality that contributed to the offending conduct. However, if this abnormality means that the offender is more *dangerous* than persons without it, then consequentialist considerations may mean that no reduction in penalty should be given. Thus the court does not look for a perfect match between desert and punishment: it is not possible to achieve perfection; and in any event, so long as an offender was responsible to some extent for his or her conduct, and had a real choice as to whether or not to commit an offence, it is reasonable that protection of society be considered as justifying a penalty limited only by what is proportionate to the criminality of the offending conduct.

2 Arguments against Retributivism

There is a school of thought that retributive purposes of the criminal law should be abandoned or at least deemphasized, which has recently been given some impetus by developments in neuroscience (see Greene & Cohen 2004; and material on philosopher Thomas Clark's website http://www.naturalism.org/).

It is argued that crime is an illness to be treated rather than wrongdoing deserving punishment, and that retribution is inhumane and harsh, and based on primitive impulses for vengeance. It is argued in particular that retribution is not justified, because the real responsibility for criminal conduct lies in the genes and circumstances of the criminal, and free will and responsibility are illusions. I will look further at that argument in the next section.

It is also argued that everything that is reasonable about the criminal justice system can be supported by its consequentialist purposes. It is those people who commit crimes unaffected by insanity or other disabling mental conditions who may be deterred by punishment and threats of punishment, and whose rehabilitation may be assisted by appropriate punishment; while those who act dangerously because of insanity or other disabling mental conditions are most usefully treated not as criminals but as persons who may need to be restrained to protect the community. Further, it is said, abandonment of ideas of retribution would eliminate harsh treatment that is not justified by any good consequences and would enable proper attention to be given to things that really matter, namely what are the genetic or environmental causes of crime and how are criminals best treated to free them from those causes and to ensure they do not commit crimes in the future.

Advocates of this view note arguments to the effect that general deterrence and reassurance of the community are achieved so long as the person punished *appears to be guilty*, so that punishment of the innocent is not ruled out; but they point to powerful consequentialist reasons for having a system that so far as possible ensures that only those who are truly guilty are punished, proportionately to their guilt, namely: (1) to promote confidence in the justice system and the assurance that law-abiding citizens are not punished; and (2) to limit punishment to those who need and can benefit from deterrence and reform.

However, I suggest it is still the case that, on a purely consequentialist approach, if a mistake is made and an innocent person is punished, this can be considered a bad thing only if the overall consequences are worse – *injustice* as such does not count for anything. Thus, if the accidental and undetected punishing of an innocent person happens to have better consequences than the exoneration of that person, because the harm to the individual is outweighed by the general deterrent effect and the placating of the community, the consequentialist purposes of the criminal law would be served.

3 Challenges to Free Will and Responsibility

The denial of free will and responsibility referred to above is not based on new ideas, but rather on old ideas that have been given fresh impetus by recent neuroscience.

In the eighteenth century Pierre Laplace argued that everything must happen as determined by Newton's (wholly deterministic) laws of motion, leaving no room for any contribution from free choice. The Scottish philosopher David Hume argued that what we do is determined by the preponderance of our desires: we always do what we most want to do, and what we most want to do is in turn determined by our characters and our circumstances.

In the nineteenth century, Darwin's theory of evolution offered an explanation of how random mutations and natural selection could lead to the emergence of organisms that give the appearance of making decisions and pursuing goals, while in fact being controlled by the operations of physical brains operating wholly in accordance with laws of nature.

At the end of the nineteenth century and beginning of the twentieth century, the work of Sigmund Freud drew attention to the extent to which our behavior is affected by unconscious processes, thus further challenging the idea that what we do is a matter of conscious choice. At around the same time, the deterministic physics of Newton was displaced by the indeterminism of quantum mechanics; but this has not generally been regarded as supporting free will and responsibility, because it did no more than introduce the possibility of *randomness* into physical processes, and randomness may be considered the antithesis of efficacious decision-making and responsibility for conduct.

Arguments against free will and responsibility advanced by philosophers in the 20th century have included the argument that, for anything that happens, there must be sufficient prior causes, and there must in turn be sufficient prior causes for those causes, and so on; so that there is no room for any contribution from a human decision-maker that is not itself determined by prior causes outside the decision-maker's control. Thus the Australian philosopher J.J.C. Smart (1961) argued that every event is either causally determined or due to chance, and that there is no logical room between determinism and chance.

A basic dilemma about free will and responsibility has been well expressed by contemporary British philosopher Galen Strawson (1998, 2002), in the following four propositions:

1. We do what we do because of the way we are, in terms of character and motivation.
2. So we cannot be responsible for what we do unless we are responsible for the way we are.
3. We cannot be responsible for the way we are when we first make decisions in life (that must be all down to genes and environment).
4. So we can never by earlier decisions become responsible for the way we are or for what we do.

These philosophical arguments have been reinforced in various ways by recent work in neuroscience. As more becomes known about how the brain works, in terms of processes that can be understood in terms of physics, chemistry and biology, the less room there may appear to be for free will. Neuroscience approaches the brain as a kind of machine, and seeks to explain its operations as being the operations of a machine.

One particular area of scientific research that is sometimes seen as excluding the possibility of free will is that undertaken by neuroscientist Benjamin Libet and his colleagues (Libet et al. 1983). These experiments suggest that consciousness comes too late to initiate certain actions that the agent considers to be voluntary. In particular, where subjects were asked to make a movement at any time they decided to, neurological preparation for the movement was detected about half a second *before* the time identified by the subject as the time of the decision to move, suggesting that the subject's feeling that the decision was freely chosen must be an illusion.

Developments in neuroscience have led to suggestions from neuroscientists such as Colin Blakemore and Francis Crick that our feeling that we have free will is an illusion that arises because we are unaware of the physical brain processes that actually cause our actions (Blakemore 1988, pp. 257, 269–71; Crick 1994, p. 266); and psychologist Daniel Wegner has discussed experiments in which people claim to have made free choices in circumstances where they could not possibly have done so (Wegner 2002, pp. 74–78).

4 Virtues of Retribution

Notwithstanding the above arguments, I contend there are good reasons for retaining retribution as a guiding purpose of criminal law, both as a general basis for determining when detriment is to be imposed on citizens and as an important factor in determining what detriment is to be imposed (see Hodgson 2000). These reasons include the following.

1. If one has regard only to the *consequences* of what the State does to its citizens, they are being treated as *objects* to be manipulated for the general good. That is not appropriate in relation to people who are capable of acting rationally; and it does not encourage people to take responsibility for their conduct.

2. Far from being inhumane, it is humane to refrain from imposing detriments on persons who have not been proved to have voluntarily breached a public law, and to impose on persons who have been proved to have breached a public law detriments that are no greater than are proportionate to the criminality of the offences committed. In Western societies, it is widely regarded as *legitimate* for governments to impose sanctions on people who have acted voluntarily in breach of public laws and, with limited exceptions, *not legitimate* for governments to impose sanctions or other loss of liberty on people who are innocent of doing this: to impose sanctions or loss of liberty on people who are not at fault in this regard, without powerful justification, is considered a gross violation of human rights.

3. Retribution that is proportionate to the criminality of the offence does not justify treatment that is harsh in a way that does not contribute to deterrence, and it is consistent with pursuing rehabilitation. Where there is harsh treatment that does not contribute to deterrence and/or interferes with rehabilitation, this is due to defects in the system rather than correct application of retribution.

4. This retributive approach has further indirect advantages, in particular:

(a) if punishment is not generally limited to those who deserve it, no one could feel secure that compliance with the law would generally ensure they were not subjected to loss of liberty or confiscation of property by the State; and

(b) if punishment is not made dependent on wrongdoing and proportional to the seriousness of the wrongdoing, the law would be less respected and resort to self-help and vengeance may not be kept in check.

5. The need to prove voluntary conduct in breach of a public law before imposing detriment on persons is a necessary restraint on the conduct of the State and its officials. If the imposition of detriment is considered as justified whenever doing so has good consequences, whether or not the person concerned has done something to deserve it, there is no clear basis in principle for inhibiting the State and its officials from arbitrary arrest and detention.

It may be argued that these (or at least most of them) are consequentialist considerations that can be taken into account in a broader consequentialist theory, and that they do not in any event bear on the question whether or not there is any *truth* in the idea that people are really responsible for their conduct.

However, I am not here seeking to make out a general case against such a broad consequentialist approach, but rather to show there are good reasons for retaining retribution as a guiding purpose of the criminal law; and I suggest that the advantages of doing so cannot be maintained if ideas of retribution and desert are given up in favor of a purely consequentialist approach.[1]

[1] I contend there is no paradox here. Not only can there be mistakes made in applying consequentialism and misuse of appeals to consequentialism, but also direct application of consequentialism can preclude honest application of beneficial nonconsequentialist principles: see generally my book, Hodgson 1967, and Regan 1980.

In particular, I contend that, unless one regards people as truly responsible for their conduct, there is no moral reason requiring different treatment for a person who *has had the bad luck* to be regarded by a State official to be a danger to the State, from that given to a person who *has had the bad luck* to be caused by genes and environment to breach a public law. Both would be equally undeserving of what happens to them. The principle of human rights referred to earlier would thus be in jeopardy if ideas of retribution are abandoned or weakened. This is particularly so, having regard to the fact that consequentialist arguments are so often indecisive. What is required to protect human rights is a requirement of justice limiting interference with liberty to those cases where citizens have voluntarily breached a public law, except where there is clear and powerful justification for overriding it. If we do not punish people *because* they are guilty, there is less reason to refrain from punishing them if *and because* they are innocent.

And while it is true that these considerations do not bear on the *truth* of free will and responsibility, they give us good reason to consider very critically arguments that may be advanced against retribution and against free will and responsibility.

5 Three Ways to Maintain Retribution

There are three broad ways in which various writers have sought to maintain retribution despite challenges to free will and responsibility.

One approach is to say that ideas of responsibility and retribution should be maintained even though there is no such thing as free will. Some argue, on the basis of considerations of the type mentioned in the previous section, that we should maintain the illusion that there is free will even though there is not (e.g., Smilansky 2002). One prominent legal theorist who might possibly be included in this category, or possibly in the next, is Michael Moore (1997), who argues that the persistence and ubiquity of retributive impulses are signs of the moral reality that retribution is a *good thing* and a primary aim of the law, even though there is no free will. However, he also supports the view that it is sufficient for culpability and responsibility that a person has the capacity to act rationally and to respond to reasons; and that a person is excused if sufficiently compromised in his or her rational capacity or coerced to act against his or her wishes; and in this respect his views are similar to those in the next category.

The second approach, adopted by philosophers such as Daniel Dennett (2003), is to argue that, although the world is deterministic for all practical purposes, nevertheless free will and responsibility exist and are *compatible* with this determinism. Human beings have free will and responsibility just because they are free to act in accordance with their own choices and to do whatever it is they most want to do; and it does not matter that their choices and their wants may themselves be determined by prior circumstances and impersonal laws. A prominent legal theorist Stephen J. Morse (2000) argues that responsibility is explained by our capacity to grasp and be guided by good reasons, and that this is so despite the truth of determinism.

The third approach is to say that we have free will and responsibility in a sense that is incompatible with determinism. This view is called *libertarianism*, and it is a minority view in philosophy and science.

For my part, I do not think that retribution could be justified if there is no free will or responsibility for conduct, and I do not think it is a good idea to act on the basis of a pretence. It is best always to try to ascertain the truth, and to guide our conduct and our practices on the basis of what we reasonably believe. Accordingly I do not support the first of the above approaches, to the extent that it advocates maintenance of an illusion.

The second approach, compatibilism, is a defensible view (it is probably the majority view among philosophers). Certainly, even if the world is deterministic for all practical purposes, there still would be an important distinction between those persons who are rational in their behavior, and those who by reason of mental abnormality are not rational or not fully rational; and this distinction supports retributive ideas, in that those who are not rational are less likely to be deterred by the threat of punishment, and are more appropriately dealt with by treatment, and if necessary confinement, than by punishment.

And I would accept Stephen Morse's advocacy of the capacity to grasp and be guided by good reasons as supporting our responsibility concepts and practices, in a way that can be adopted without necessarily having to deny determinism. A person who has that capacity may be considered to have a fair opportunity to conform his or her behavior to the requirements of the criminal law (an approach advocated by the jurist Herbert Hart; see Lacey 2007, p. 237), and thus fairly to be subjected to punishment if he or she does not do so. This approach is also consistent with that adopted by Nancey Murphy and Warren Brown (2007), suggesting that the "downward causation" of a person acting for reasons can support attribution of responsibility, whether or not determinism is true.

However, I have concerns about this approach. I think the commonsense notion of capacity to grasp and be guided by reasons depends heavily on the assumption that this capacity is exercised by *conscious decision–making* in which conclusions are reached on the basis of consciously grasped reasons that may be incommensurable and inconclusive, and which thus call for conscious resolution (see Hodgson 2005). Determinism, and particularly a deterministic understanding of neuroscience, strongly challenges this assumption, suggesting the real "decisions" are by rule-governed processes that do not require consciousness, thus tending to undermine this approach. Further, this compatibilism remains vulnerable to Strawson's argument set out above: if we and all our actions are ultimately and completely the products of our genes and environment, how can anything we do be other than the inevitable result of things outside our control?

I myself believe there are good reasons for questioning determinism and for preferring the third, libertarian, approach; and I think there are good arguments for it that are not widely appreciated (see my 2008 and 2007b articles and other articles at http://users.tpg.com.au/raeda). It is not possible in this chapter adequately to present these arguments, but I will very briefly outline my position and how it deals with Strawson's argument.

I accept that the capacity of human beings to grasp and be guided by good reasons varies enormously, so that what we do is at least greatly influenced by the way we are. However, what I say is that, given our circumstances, the way we are plus laws of nature provide available alternatives, inconclusive reasons (and how they appeal), and unconscious tendencies, and also the capacity to decide between the alternatives on the basis of the reasons; and what we do is what we decide in exercise of that capacity. Thus the constraining effect of the way we are, although very substantial, nevertheless is limited to determining alternatives, reasons and unconscious tendencies. Our decisions are not *otherwise* constrained by any distinguishing features of the way we are, and to this extent (and contrary to Strawson) we are truly responsible for them.

My position can be illustrated by an analogy. A prominent compatibilist philosopher John Fischer has written that "our behaviour may well be 'in the cards' in the sense that we simply have to play the cards that are dealt us" (Fischer 2005, p. 129). This has drawn the apt comment from philosopher Kip Werking (http://people.wm.edu/~ktwerk/cards.doc) that it misleadingly suggests there is a *player of the cards* distinct from the hand that is dealt, whereas in truth human beings simply *are* the cards that are dealt them by genes and environment. My view can be understood as accepting this, but suggesting that each of us includes, in the hand of cards that is dealt us and that constitutes us, along with particular cards like aces, tens, jacks and so on, one powerful and flexible general-purpose card, like a joker. The particular cards engage with circumstances and laws of nature to limit our conduct to a spectrum of possibilities, while the joker, our capacity for conscious decision-making, can combine with our other cards to steer a course within this spectrum of possibilities.

I am not suggesting that this joker is a self or soul that itself makes decisions, or that it corresponds to any particular region of our brains. Rather, it is a capacity that operates only in conjunction with our other cards. It is however powerful and flexible: so long as our other cards are not seriously deficient, for example because of mental abnormality or senility or immaturity, the joker enables us to make reasonable albeit fallible decisions about what to believe (including what to believe about right and wrong) and about what to do, for good or ill. And these decisions in turn can affect what particular cards we come to hold for the future, for better or worse. Since we all have this joker, we all have some ultimate responsibility for our conduct, again so long as our other cards are not seriously deficient.

Why then do I think our cards include this joker, this capacity to decide? In brief, I say there are very strong reasons to accept that conscious experiences make a substantial positive contribution to our decision-making, and that this contribution is *not* one wholly constituted by rule-determined processes: if it were, as Alan Turing's arguments show, consciousness would be a superfluity. There is however a plausible account of how conscious experiences can make a contribution *that is not rule-determined*, namely by providing feature-rich gestalt experiences that do not engage as wholes with laws or rules of any kind but to which, as wholes, we can respond reasonably.

To take a simple example from a different area of discourse: in making a judgment about the aesthetic merits of a work of art such as Picasso's *Les Demoiselles d'Avignon*, I contend that our all-at-once grasp of gestalts, in which many features of the work are combined, plays a role that is not rule-determined, because rules do not engage with gestalts of that kind. (This general argument, and this and other examples, are discussed at some length in Hodgson 2002, 2007a).

On the view I am suggesting, the role of consciousness is to make contributions of this kind to our decision-making, giving us a capacity to make decisions that are not wholly determined by the engagement of preexisting circumstances (including our characters) with laws of nature. That is, it gives us our joker.

6 The Role of Neuroscience

So far I have set out the general nature of the challenge that some see neuroscience making to notions of free will and responsibility, and in particular to the retributive approach to criminal behavior; and I have set out my views on why it is desirable to maintain ideas of criminal responsibility and retribution, and on various approaches to how this might be done.

I now turn, finally, to consider the role that I see neuroscience playing in the future operation and development of the law as it impinges on criminal conduct.

There is no doubt that neuroscience will continue to develop and tell us more and more about how brains operate. There is every reason to expect that neuroscience will play an increasing role in the criminal law; and so long as its claims are critically appraised, this should be welcomed. Neuroscience may contribute in at least the following ways:

1. Evidence of neuroscientists will increasingly be used in determining questions about criminal responsibility, in accordance with the categories the law prescribes.

2. Neuroscience may be expected to influence the development of the law concerning the attribution of criminal responsibility, particularly in the case of those affected by mental abnormalities.

3. Neuroscience is likely to assist in identifying brain conditions that involve particular risks of criminal behavior, and in devising methods to minimize these risks.

4. More generally, increasing knowledge of how the brain works may be expected to guide programs for addressing environmental factors contributing to criminal conduct and for rehabilitation of offenders, and to contribute to the development and implementation of educational strategies to discourage criminal conduct.

5. Neuroscience may also contribute to the operation of the criminal law by enabling more reliable evaluation of evidence, for example through more reliable lie-detection technology; but since this is remote from the question of responsibility for conduct, I will not consider it further.

In relation to the first two areas I have mentioned, there is an underlying difficulty. The categories used by the criminal law are not those of neuroscience, but rather are pre-scientific commonsense categories, which have their source in what has been disparagingly called folk psychology. Categories such as willed or voluntary action, belief, and intention (of alleged offenders), and consent (of alleged victims), are often central to determining criminal liability; and they are not matters that are susceptible to scientific proof or even description. Even when what is in issue is a question of mental abnormality, the categories used by the law are nonscientific categories such as disease of the mind, knowing what one is doing, knowing that what one is doing is wrong, and substantially diminished responsibility.

These categories used by the law presuppose an active conscious agent who is generally responsible for conduct; whereas neuroscience focuses on brain functions in terms of physical causes and effects, and seeks to provide explanations of things and processes that contribute to the functioning or malfunctioning of the brain. And there is not at present any accepted theoretical framework that links and reconciles these two approaches.

Further, if I am right in my contention that it is desirable that notions of personal responsibility and retribution be maintained by the criminal law, then it is to be hoped that the criminal law will continue to use categories that give effect to these notions; that is, categories which, like those presently used, presuppose active conscious agents generally responsible for their actions.

Work of the nature discussed in other chapters of this book may promote reconciliation of these folk-psychological categories with the categories of neuroscience; but I think this is unlikely to be achieved any time soon. Thus I would expect that the contribution of neuroscience in these first two areas will for some time be limited to giving more detailed and accurate accounts of relevant aspects of brain functioning, which can then be used in commonsense reasoning to arrive at conclusions that engage with existing legal categories, or else guide future developments of the law. In particular, I would hope and expect that the law will retain the presumption that persons of sufficient maturity are generally responsible for their conduct; and I would also expect that, if this presumption is to be rebutted by neuroscientific evidence in particular cases, then this evidence will need to identify some brain abnormality and the effects of that abnormality, on the basis of which a commonsense conclusion can be reached as to whether and if so to what extent responsibility is excluded.

However, the combination of neuroscience and commonsense may enable the law to provide a more systematic approach to questions of this kind.

The present category of insanity is presumably intended to capture those cases where responsibility for conduct is effectively absent due to a brain abnormality that is of indefinite duration and/or difficult to treat; so that, if that brain abnormality is such as to make the person dangerous, there will be the option available of indefinite detention in order to protect the public. Neuroscience may well assist in achieving a better definition of these cases than is provided by the *M'Naghten* rules.

The present category of sane automatism is presumably intended to capture those cases where responsibility for conduct is absent either without there being any brain abnormality, or else because of a brain abnormality that is temporary and/or easy to treat and/or not such as to make the person a danger; so that it is reasonable to excuse the person altogether without there being any need for extensive future restraint or even monitoring. Again, neuroscience may assist in better defining these cases.

Where the law recognizes that murder should be reduced to manslaughter in cases where there is substantially diminished responsibility, neuroscience may contribute to a better definition of when such diminished responsibility is to be recognized. Again, in this area also it might be appropriate to distinguish between those cases where the abnormality that causes this diminished responsibility is of indefinite duration and/or difficult to treat, and those cases where it is temporary and/or easy to treat and/or not such as to make the person a danger; and neuroscience may assist in providing rules to distinguish these cases. This approach may have some application to those cases where addiction contributes to criminal conduct.

Another area where neuroscience may contribute to better legal rules concerns the criminal liability of children and young people. As mentioned earlier, in Australia a child under 10 cannot be criminally liable, while a child between 10 and 14 can be criminally liable if the prosecution proves that he or she knew that the conduct in question was seriously wrong. Children over the age of 14 are subject to the same presumption of responsibility as adults, although until they are 18 they are dealt with by separate courts and, if custody is considered appropriate, imprisoned in different institutions from adults. Neuroscience may contribute to a more systematic approach to the criminal responsibility of children and young people.

As regards the third area mentioned above, it is highly likely that neuroscience will be increasingly be capable of identifying brain conditions that involve particular risks of criminal behavior, and of devising methods to minimize these risks. The question is what should be done with this capability.

I would myself strongly oppose its being used to justify detention and/or compulsory treatment of persons who have not yet committed any criminal offence. That course might be advocated by those who wish to abandon retribution and be guided purely by the consequentialist purposes of criminal law; but I say it would be contrary to the principle of human rights referred to earlier, which proscribes adverse treatment by the State of persons who have not committed any criminal offence, unless there is powerful justification. Mere identification of risk factors would not provide that justification; although I accept justification could be provided where what is identified is a substantial mental illness that makes the person a clear danger to himself or herself or to others, which could then support the detention of the person for no longer than is necessary for medical treatment of the illness so as to minimize the danger.

Neuroscientific identification of risk factors could properly be taken into account in determining what sentence to pass (up to the limit considered proportionate to the criminality of the offence) on a person who has committed a criminal offence, and in guiding the treatment of that person during the period of that sentence. And it could properly provide a basis for offering voluntary treatment to persons who have not committed an offence.

One other area where identification of risk factors is now being used in some Australian jurisdictions is to justify extended detention of persons convicted of serious sex offences, who have served their sentences for their crimes but, having not successfully completed sex offender programs during their sentences, are considered at high risk of re-offending. I would argue that this procedure should be kept within strict limits, being applied only to serious sex offenders, and where the risk of re-offending is clearly proved to be high. Even then, it be could be argued to be a denial of human rights, amounting to punishment for an offence going beyond that considered proportionate to the criminality of that offence; although against this, it could be argued that extended detention is not unfair in the case of someone who has already committed a serious crime and has not taken the opportunity, provided by the sentence for that crime, to address the risk factors that contributed to its commission.

Turning to the fourth area, it is to be expected that neuroscience can and will make substantial contributions towards the addressing of environmental factors that contribute to criminal conduct and towards educational strategies to discourage criminal conduct, as well as towards rehabilitation of offenders. However, I would argue that central to both education and rehabilitation are belief in and acceptance of personal responsibility for conduct, and recognition and acceptance of appropriate moral standards of right and wrong. If neuroscience is permitted to undermine belief in and acceptance of personal responsibility for conduct and/or belief in the importance of moral considerations, then this would be highly counterproductive.

References

Blakemore, C.: The mind machine. BBC, London (1988)

Crick, F.: The astonishing hypothesis. Simon & Schuster, London (1994)

Dennett, D.: Freedom evolves. Allen Lane, London (2003)

Fischer, J.: The cards that are dealt you. Journal of Ethics 10, 107–129 (2005)

Green, J., Cohen, J.: For the law, neuroscience changes nothing and everything. Philosophical Transactions of the Royal Society of London: Series B, Biological Sciences 359, 1775–1785 (2004)

Hodgson, D.: Consequences of utilitarianism. Oxford University Press, Oxford (1967)

Hodgson, D.: Guilty mind or guilty brain: Criminal responsibility in the age of neuroscience. The Australian Law Journal 74, 661–680 (2000)

Hodgson, D.: Three tricks of consciousness. Journal of Consciousness Studies 9, 65–88 (2002)

Hodgson, D.: Responsibility and good reasons. Ohio State Journal of Criminal Law 2(2), 471–483 (2005)

Hodgson, D.: Making our own luck. Ratio 20, 278–292 (2007a)

Hodgson, D.: Partly free. The Times Literary Supplement, 15–16, July 6 (2007b)

Hodgson, D.: A role for consciousness. Philosophy Now 65, 22–24 (2008)

Lacey, N.: Space, time and function: Intersecting principles of responsibility across the terrain of criminal justice. Criminal Law and Philosophy 1, 233–250 (2007)

Libet, B., Gleason, C., Wright, W., Pearl, D.: Time of conscious intention to act in relation to onset of cerebral activities (readiness potential): The unconscious initiation of a freely voluntary act. Brain 106, 623–642 (1983)

M'Naghten's Case, 10 Clark & Finnelly, 200–214 (1843) (8 English Reports 718-24)

Moore, M.: Placing blame: A general theory of criminal law. Oxford University Press, Oxford (1997)

Morse, S.J.: Rationality and responsibility. Southern California Law Review 74, 251–268 (2000)

Murphy, N., Brown, W.: Did my neurons make me do it? Philosophical and neurobiological perspectives on moral responsibility and free will. Oxford University Press, Oxford (2007)

Falconer, R.v.: 171 Commonwealth Law Reports, pp. 30–87 (1990)

Regan, D.: Utilitarianism and co-operation. Oxford University Press, Oxford (1980)

Smart, J.J.C.: Free-will, praise and blame. Mind 70, 483–494 (1961)

Smilansky, S.: Free will and illusion. Oxford University Press, New York (2002)

Strawson, G.: Luck swallows everything. Times Literary Supplement, 8–10, June 26 (1998)

Strawson, G.: The bounds of freedom. In: Kane, R. (ed.) Oxford handbook of free will. Oxford University Press, New York (2002)

Veen, v.: The Queen (No. 2), 143 Commonwealth Law Reports, 458–498 (1988)

Wegner, D.: The illusion of conscious will. MIT Press, Cambridge (2002)

15

Law, Responsibility, and the Brain

Dean Mobbs[1,2,*], Hakwan C. Lau[2,3], Owen D. Jones[4], and Chris D. Frith[2]

[1] MRC-Cognition and Brain Sciences Unit
Cambridge, CB2 2EF
United Kingdom
dmobbs@gmail.com
[2] Wellcome Trust Centre for Neuroimaging
University College London, London
[3] Department of Experimental Psychology
University of Oxford, Oxford
[4] School of Law and Department of Biological Sciences
Vanderbilt University
Nashville, Tennessee

Summary. In perhaps the first attempt to link the brain to mental illness, Hippocrates elegantly wrote that it is the brain that makes us mad or delirious. Epitomizing one of the fundamental assumptions of contemporary neuroscience, Hippocrates' words resonate far beyond the classic philosophical puzzle of mind and body and posit that our behavior, no matter how monstrous, lies at the mercy of our brain's integrity. While clinicopathological observations have long pointed to several putative neurobiological systems as important in antisocial and violent criminal behavior, recent advances in brain-imaging have the potential to provide unparalleled insight. Consequently, brain-imaging studies have reinvigorated the neurophilosophical and legal debate of whether we are free agents in control of our own actions or mere prisoners of a biologically determined brain. In this chapter, we review studies pointing to brain dysfunction in criminally violent individuals and address a range of philosophical and practical issues concerning the use of brainimaging in court. We finally lay out several guidelines for its use in the legal system.

Key words: brainimaging, prefrontal cortex, freewill, responsibility, violence.

Abbreviations: APD, antisocial personality disorder; fMRI, functional magnetic resonance imaging; PFC, prefrontal cortex.

[*] Corresponding author.

Archaeological discoveries of traumatic injuries in primitive hominid skulls strongly hint that our species has a long history of violence (Walker 2001). Despite repeated attempts throughout history, including efforts to eliminate violence through the imposition of criminal sanctions, we have yet to dispel our violent nature. Consequently, criminal violence remains a common feature of most societies. As policy-makers seek deeper understandings of criminally violent and antisocial behavior, many contemporary neuroscientists assume that the essential ingredients of the human condition, including free will, empathy, and morality, are the calculable consequences of an immense assembly of neurons firing. Intuitively, this view opposes Cartesian dualism (i.e., the brain and mind are separate, but interacting, entities) and assumes that violence and antisocial behavior emanate from a mechanistically determined brain (see appendix 1).

From this standpoint, the exciting discoveries of neuroscience resonate far beyond mere philosophical banter and may have important implications for the way government institutions, including education and legal systems, operate. For example, to the extent that legal systems attempt both to move behavior in socially desirable directions and also to adjudicate transgressions fairly, the legal system's effectiveness can be improved by deepening our understandings about why people behave as they do and both how and why people respond to various changes in legal incentives. Specifically, neuroscience may have important implications for both how we understand the multiple influences on violent behavior and how the legal system may better engage with violent criminals.

1 Studies of the Prefrontal Cortex in Antisocial and Violent Populations

The birth of what may be coined modern "forensic neurology" lies in John Harlow's nineteenth-century observations of Phineas P. Gage (Harlow 1848). Gage, a railroad worker, suffered the unfortunate experience of having an iron bar blasted through the front of his brain, which resulted in extensive damage to the prefrontal cortex (PFC). Despite Gage's miraculous physical and intellectual recovery, conspicuous changes in his personality were reported. Briefly, the once courteous and diligent man became explicitly antisocial. As Gage's friends famously articulated, "Gage is no longer Gage." Since Harlow's lurid description, computerized reconstructions based on Gage's skull fractures have determined more precisely the damaged PFC regions, which current evidence associates with autonomic, social, and affect regulation (Damasio 1994). The case of Phineas Gage is compelling to both neuroscientists and legal thinkers because it provided the first indication that reasoning and regard for others can be compromised by frontal lobe injury. Harlow's observations have led many experts to speculate that neurological insult may be a prominent factor in recidivistic and violent criminal transgressions.

Modern empirical endeavors support the claim that the human PFC, a latecomer in the brain's phylogenic history, is what makes us rational, intellectual, and moral entities (table 15.1). For example, several studies on patients with focal frontal lobe injuries have supported Harlow's case (Meyers et al. 1992; Blair & Cipolotti 2000). In one of the largest studies of patients with brain damage to date, Grafman and colleagues found that increased aggressive/violent scale scores were most strongly associated with similarly localized PFC lesions in a sample of 279 veterans of the Vietnam War (Grafman et al. 1996). Higher scores were, however, mostly associated with verbal aggression and less so with physical aggression, again fitting with Harlow's observations of Gage (Harlow 1848). These studies, along with clinical observations, have led many to suggest that damage to the PFC results in "acquired sociopathy" or "pseudopsychopathy."

Table 15.1. Examples of Prefrontal Brain Regions Associated with Pro-Social Behavior

Brain Region	Pro-Social Behavior
Anterior cingulate cortex	Empathy (Amodio & Frith 2006; Singer et al. 2004)
Orbital PFC	Regret (Coricelli et al. 2005)
Ventromedial PFC	Ethical decisions (Heekeren et al. 2003; King et al. 2006)
Ventrolateral PFC	Inhibition of behavior (Aron et al. 2003)
Dorsoolateral PFC	Reasoning (Baird & Fugelsang 2004; Bechara & Van Der Linden 2005)

Given the PFC's historical and theoretical relevance to adaptive social behavior, it is not surprising that this region was among the first to be examined in antisocial and violent populations. Raine and colleagues used noninvasive structural brain imaging to show an 11% reduction in PFC grey matter in patients with antisocial personality disorder (APD) (Raine et al. 2000). These decreases in grey matter were also associated with decreased autonomic arousal to a social stressor (i.e., videotaped speech about an individual's faults). Similar reductions have been observed in a study of aggressive patients (Woermann et al. 2000) and of pathological liars (Yang et al. 2005). Nonetheless, such morphological and volumetric abnormalities may not necessarily relate to behavior.

In principal, using brain imaging to look at function rather than structure should reveal stronger relationships between brain and behavior. Using positron emission tomography scanning, neuroscientists have found attenuated resting regional cerebral blood flow in the frontal lobes of violent individuals (Volkow & Tancredi 1987) and convicted criminals (Raine et al. 1994). In healthy volunteers, evoked anger and imagined aggressive transgressions are associated with reduced modulation of the orbital and medial PFC (Gordon et al. 2004). Collectively, these studies suggest that impulsive violent acts stem from diminished recruitment of the PFC's "inhibition" systems.

2 Beyond the PFC

The PFC is not, however, the only area where damage may increase propensity toward behaviors deemed criminal or antisocial. It has long been known that ablation of the monkey temporal lobe, including the amygdala, results in blunted emotional responses (Bucy & Kluver 1955) (fig. 15.1C). In humans, brain-imaging and lesion studies have suggested a role of the amygdala in theory of mind, aggression (van Elst et al. 2001), and the ability to register fear and sadness in faces (Blair et al. 1999). According to the violence inhibition model, both sad and fearful facial cues act as important inhibitors if we are violent towards others. In support of this model, recent investigations have shown that individuals with a history of aggressive behavior have poorer recognition of facial expressions (Weiss et al. 2006), which might be due to amygdala dysfunction (Adolphs et al. 2005). Others have recently demonstrated how the low expression of X-linked monoamine oxidase A (MAOA) – which is an important enzyme in the catabolism of monoamines, most notably serotonin (5-HT), and has been associated with an increased propensity towards reactive violence in abused children (Caspi et al. 2002) – is associated with volume changes and hyperactivity in the amygdala (Meyer-Lindenberg et al. 2006).

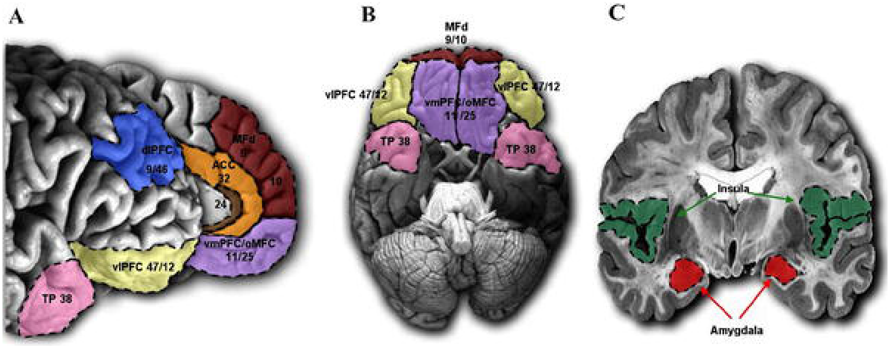

Fig. 15.1. Regions Associated with Normal and Atypical Social Behavior
A. Medial and lateral view of the PFC
B. View of the ventral surface of the PFC and temporal poles
C. Coronal slice illustrating the amygdalar and insular cortex.
See also Table 15.1.
ACC, anterior cingulate cortex; dlPFC, dorsolateral PFC; MFd, medial PFC; oMFC, orbitomedial PFC; TP, temporal pole; vlPFC, ventrolateral PFC; vmPFC, ventromedial PFC.

The amygdala has been a major focus of attempts to understand the poor empathy and fear responses observed in psychopathic criminals. Using functional magnetic resonance imaging (fMRI), Birbaumer and coworkers (Birbaumer et al. 2005) presented individuals with a paradigm in which the appearance of a face on a screen was followed by a painful shock in one condition but not in a second condition. Analysis showed normal volunteers to have increased activity in the amygdala (see fig. 15.1) in response to faces associated with shock, whereas psychopathic individuals showed no significant change in activity in this region. In addition, psychopaths also failed to show normal increases in skin conductance responses. Importantly, Birbaumer et al.'s findings are supported by studies showing that the limbic structures (i.e., amygdala and hippocampus) are functionally abnormal in psychopathic criminals during emotional memory (Kiehl et al. 2001) and by studies showing how activity in the amygdala decreases with increased scores on the Psychopathy Personality Inventory (Gordon et al. 2004; Tiihonen et al. 2000). A prevailing hypothesis is that in psychopathic criminals the prefrontal-amygdala connections are disrupted, leading to deficits in contextual fear conditioning (LeDoux 1996), regret (Coricelli et al. 2005), guilt (Takahashi et al. 2004), and affect regulation (Dolan 2002).

3 Does the Crime Fit the Brain?

While many behaviors can be unambiguously defined, labeling a behavior as "criminal" is to define how the behavior will be considered socially. That is, the very same behavior that might not be deemed criminal in one social context (say, shooting a gun at a target at a shooting range) may be deemed criminal in another (such as shooting a gun in the direction of a crowd of people). Such definitional ambiguities are at their least frequent, however, with respect to interpersonal violence, which is broadly proscribed.

It is clear in at least some contexts that different violent antisocial behaviors can arise from different etiologies. Animal studies have shown that distinct networks underlie different types of aggression (e.g., predatory attack and defensive rage [Gregg & Siegel 2001]). From these studies, one might expect that in humans, distinct neural topographies exist in, for example, the sexual criminal, the sadistic murderer, and the political terrorist. At first glance, such reasoning looks like phrenological folly; however, evidence does suggest that violent behavior can be placed into two broad, yet distinct categories: affective aggression (i.e., impulsive, autonomic arousing, and emotional) and predatory aggression (i.e., premeditated, goal-directed, and emotionless) (Vitiello et al. 1990).

With this dichotomy in mind, Raine and colleagues (Raine et al. 1998) reanalyzed positron emission tomography data to tease apart functional differences between premeditated psychopaths and impulsive affective murderers. Compared to controls, the impulsive murderers had reduced activation in the bilateral PFC, while activity in the limbic structures was enhanced. Conversely, the predatory psychopaths had relatively normal prefrontal functioning, but increased right subcortical activity, which included the amygdala and hippocampus. These results suggest that predatory psychopaths are able to regulate their impulses, in contrast to impulsive murderers, who lack the prefrontal "inhibitory" machinery that stop them from committing violent transgressions. Although more work is necessary, these studies strongly suggest that some kinds of criminal behavior are associated with dysfunction of different regions of the brain.

4 Does Some Criminal Behavior Result from Mental Disorder?

A great deal of empirical research demonstrates that mental illness is higher in incarcerated populations and estimates that as many as 25% of defendants evaluated for competency are medically and legally incompetent to stand trial (Golding et al.1984). Moreover, only 36% of the public perceive recidivistic crime as an organic disorder (Raine 1993). Consequently, weighing discrepancies between intuitions, expert views, and empirical findings is of fundamental importance to a legal system.

Both the *Diagnostic and Statistical Manual of Mental Disorders* and *International Classification of Diseases 10* classifications of mental and behavioral disorders include APD, which is defined respectively in the two classifications as a lack of regard for the feelings of others and a failure to abide by society's rules. While it can be said that any given population of incarcerated criminals may not be a representative sample of all criminals, or even of all criminals who pass through the prison system, a systematic review of studies examining mental illness in 23,000 prisoners showed that these prisoners were several times more likely to have some form of psychosis or major depression, and ten times more likely to exhibit APD than the general population (Fazel & Danesh 2002). The authors suggest that, worldwide, several million prisoners have serious mental illness (Fazel & Danesh 2002). Several studies also show levels of head injury to be higher in violent and death-row criminals (Volavka et al. 1995), while birth complications, which can often result in neurological damage (e.g., hypoxic-ischemic encephalopathy) and parental mental illness, are higher in antisocial populations (Raine 1993). More often than not, people with APD and violent behavior have a history of childhood maltreatment or trauma (Widom 1989); having such a history has been linked to anomalous development of regions associated with antisocial behavior, including the PFC, hippocampus, amygdala, corpus callosum, and hypothalamic-pituitary-adrenal axis (Bremner 2005). Early damage to the orbitofrontal cortex in particular appears to result in poor acquisition of moral and social rules (Anderson et al. 1999), thus showing the importance of the interaction between environment and brain development.

Discussing the possibility of meaningful links between some antisocial and violent behavior and various brain disorders can, however, enrage retributivists, who point out that moral responsibility lies in the social rules by which acts are judged – not in the brain itself (Abbott 2001). Nonetheless, there are many instances where brain disease can lead to antisocial behavior, and these inevitably pose important complications for moral and legal systems that tend to divide responsibility for actions into dichotomous alternatives – guilty versus not guilty – instead of seeing responsibility as existing along a continuum. For example, compared to the general population, individuals with frontotemporal dementia, Huntington disease, and attention deficit/hyperactive disorder have a higher prevalence of episodic aggression or antisocial conduct. One disturbing example cited by Goldberg (2001) is the case of a New York surgeon who, after finishing surgery, carved his signature in the patient's stomach. The surgeon was later diagnosed with Pick disease (a form of dementia associated with personality changes that presumably result from progressive degeneration of frontal and anterior temporal cortices). He was not considered responsible for his actions by experts, the jury, or even the victim. Beyond these examples lies the possibility that some forms of antisocial or violent behavior are of unspecified origin, which could place them in the same category as many other neuropsychiatric disorders. Presumably, such unrecognized brain abnormalities might cause acts of gratuitous violence, but the individuals concerned would be considered to be criminally responsible.

To be clear, there is at present no reason to believe that all criminal behaviors, or indeed even all violent criminal behaviors, are the result of organically dysfunctional brains. However, there is ample evidence to suggest that some kinds of dysfunction are likely to increase the probability of some kinds of behaviors that society labels as criminal. This suggests that research is urgently needed to elucidate the links between mental illness, neurological disorder, and criminal conduct. And modern and rapidly improving brain-imaging techniques may contribute significantly.

5 Possible Legal Implications

Advances in neuroscience could have several implications for the legal system. At the broadest level, these include (1) understanding how cognitive processes of key legal participants (such as judges and jurors) influence trial outcomes, (2) discovering whether various assumptions underlying the evidentiary rules (such as one suggesting that "excited utterances" are less likely than average to be falsehoods) have any basis in fact, (3) learning more about how people determine "just" punishments, (4) anticipating how jurors may overreact to certain kinds of character evidence, (5) determining the extent of injuries from accidents, (6) improving our abilities to detect mental biases and prejudices that may affect the proper function of legal fact-finding and decision-making, and (7) learning more about the limits of witness memories. Yet even against this broad background, few implications for the legal system are more important than trying to gain a better understanding of important influences on criminal behavior.

However, that very significance brings its own important challenges. On one hand, a better understanding may lead to more effective deterrence, to more effective treatment, and to more just and morally sound sentencing. On the other hand, determining criminal responsibility is a normative legal conclusion, not an empirical factual one, made in the context of a variety of often conflicting aspirations (Morse 2006). Therefore, even the best neuroscientific study can only afford factual evidence to be weighed alongside other behavioral evidence and normative considerations, rather than actually resolve the legal question as to which the factual evidence is relevant.

Generally speaking, in the Anglo-American criminal justice system, a person can be held criminally responsible if he performs a prohibited act intentionally and with a statutorily specified mental state (which may span such things as "purpose," "knowledge," "recklessness," or "negligence") (Morse 2006). Yet even if these criteria are satisfied, the defendant can be excused from liability if legally insane. That is, he may have intentionally and knowingly committed a proscribed act, but be found not blameworthy nonetheless, because a mental condition meeting a specified legal (as distinct from medical) threshold prevented him either from knowing the nature and quality of his act, or from understanding the wrongfulness of the act (LaFave 2003).

The possibility of being "not guilty by reason of insanity" can be traced back to the well-known *M'Naghten Case* in 1843. While attempting to kill the British prime minister, Daniel M'Naghten mistakenly killed the prime minister's secretary. Experts maintained that M'Naghten exhibited such a vast deterioration in his reasoning abilities (believing the prime minister to be heading a murderous conspiracy) that he had no comprehension of the act he committed. The modern standards for determining legal insanity, in the long wake of M'Naghten, vary markedly across jurisdictions, with results that have prompted many calls for reform. For example, psychiatrists have been plagued by the need to answer dichotomously whether a defendant is "mad" or "bad" or to opine that "it is not him, it is his disease" (Sapolsky 2004). Furthermore, medical research indicates that patients with selective damage to the PFC can often know right from wrong, but still be unable to act on such knowledge. This has naturally led defense attorneys and prosecutors to pursue more objective ways of determining whether a defendant is competent to stand trial, and if so, whether he can be held legally responsible for his actions. This, in turn, has generated significant interest in brain-imaging evidence concerning a defendant's mental functioning (appendix 2).

Several examples illustrate the kinds of contexts in which many believe that brain imaging may aid the law's ability to accurately assess a defendant's mental functioning. Consider the case of a 40-year-old man who inexplicably became a sexual impulsive with paedophilia. The patient had no prior history of sexual misconduct, but it was soon noted that he was frequenting prostitutes and that he attempted to molest his 12-year-old step-daughter. He was quickly reported to the local authorities, was found guilty of child molestation, and was sentenced to either attend a 12-step sexual addiction program or face jail. Despite a strong yearning not to go to prison, the patient could not inhibit his sexual impulses. It was soon discovered that the defendant had a large tumor pressing on his right orbitofrontal cortex (fig. 15.2). Upon the resection of the tumor, the patient's sexual impulsiveness diminished. When the sexual impulsiveness later reappeared, a brain scan revealed that the tumor had grown back. A second resection of tumor again diminished the patient's sexual impulsiveness (Burns & Swerdlow 2003). Another illustration is the 1998 case of 15-year-old Kip Kinkel, who shot and killed his parents and two high-school colleagues in the state of Oregon. Brain imaging was used as evidence in court to support Kinkel's "not guilty by reason of insanity" plea. The trial defense provided evidence of small cavities in Kinkel's frontal lobe. Although there was no evidence that this abnormality caused his behavior (Kinkel was ultimately convicted as an adult and sentenced to 111 years in prison [Frontline 2000]), future developments in neuroscience may again aid courts in these kinds of inquiries.

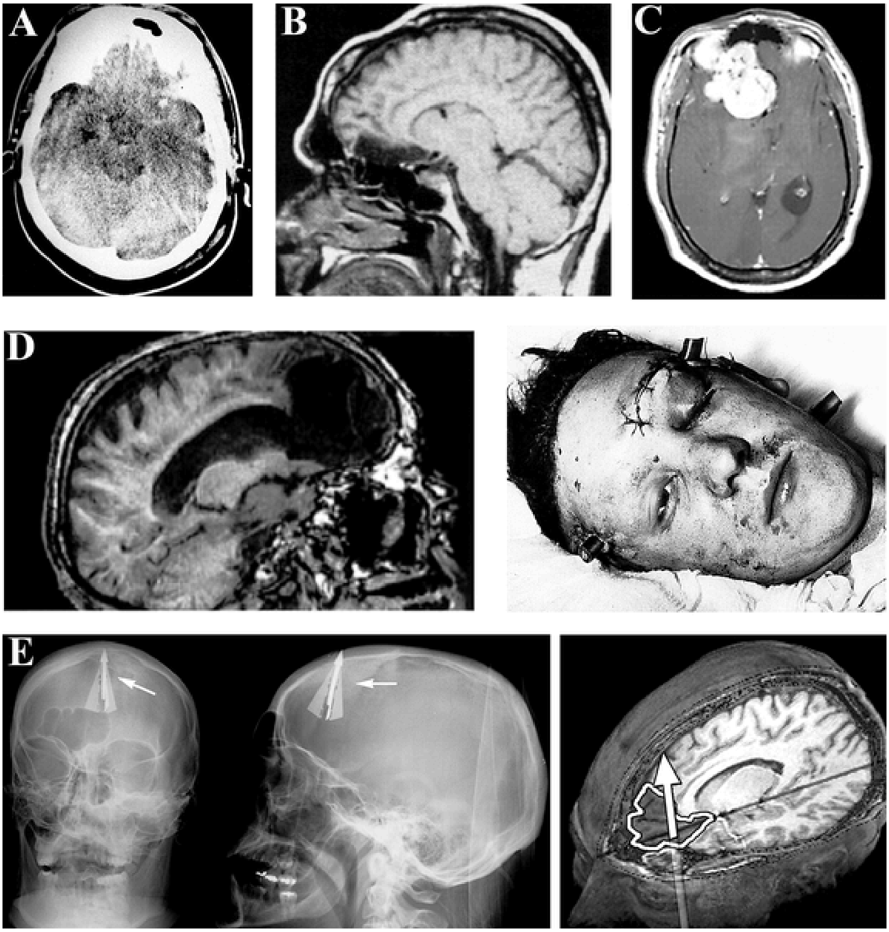

Fig. 15.2. Cases where brain anomalies have, or have not, been linked to antisocial behavior

A. Brain scan of patient J.S., who exhibited sociopathic behavior (Blair & Cipolotti 2000). The image shows a lesion in the orbital frontal cortex.

B. fMRI sagittal slice of the brain of patient J.Z., showing a lesion that was caused by the resection of pituitary tumor (Meyers et al. 1992). This lesion led to antisocial conduct, which was not exhibited before the surgery.

C. Orbitofrontal damage associated with symptoms of paedophilia and sexual misconduct in the case of a 40-year-old male patient.

D. Photograph of a patient after head injury (right) and fMRI scan 60 years later showing PFC damage (left) (Mataró et al. 2001). This patient showed personality changes, but no signs of antisocial conduct.

E. Cranial X-ray of a man who attempted suicide with a crossbow. Although the individual exhibited premorbid APD, the PFC damage caused by the crossbow arrow resulted in reversal of antisocial conduct (Ellenbogen et al. 2005).

These examples raise important questions not only about the extent to which neuroimaging may affect particular trial outcomes, but also about the ways in which the legal system can come to understand changing views of the brain, assess when those views are relevant, and determine how, in appropriate circumstances, to integrate that knowledge into legal decision-making (Feigenson 2006) (see appendices 2 and 3). For example, recent evidence suggests that the PFC continues to mature until the age of 25 (Giedd et al. 1999) and that this maturation correlates with ability in counterfactual (if-then) thinking (Baird & Fugelsang 2004). An underdeveloped ventrolateral PFC can be directly associated with poorer cognitive control (Bunge et al. 2002), which some consider a core variable in criminal activity (Gottfredson & Hirschi 1990). Such research and theory likely warrants serious consideration, given the robust relationship between age and violent criminal offences. For example, British Crime Survey statistics show that individuals between the ages of 16 and 24 commit more violent acts than all other age groups combined.

Such statistics have a special relevance in countries such as the United States where the death penalty is applied. For example, many lawyers who oppose capital punishment of juveniles hold the view that the legal system should take such neuroscientific evidence into account (e.g., the Justice for Children Project; http://moritzlaw.osu.edu/jfc) (Scott 2005). It is possible that the 2005 decision of the Supreme Court of the United States (Roper v. Simmons) that made it illegal to use capital punishment for any offender who was under the age of 18 when he committed his crime was influenced in part by evidence presented in amicus (so-called friends of the court) briefs, which included neuroscientific evidence (McLaughlin et al. 2004).

6 The Limits of Brain Imaging as Evidence

There are many exciting possibilities for how law and neuroscience may eventually partner – with neuroscientists discovering new things about the brain potentially relevant to law, and law asking questions that new neuroscientific research may help address. However, it is important to keep in mind a variety of limitations of brain-imaging technology. We highlight six.

First, functional brain imaging is not mind reading. Not only can it not tell us what or how a person was thinking at the time of a legally relevant act, it also cannot tell us with reliable accuracy what a person is thinking while being scanned. In this respect, brain imaging can only provide post hoc explanations (Raine 1993). The challenge of functional brain imaging has been likened to looking from an airplane window at night: when we look down from the plane we see complex patterns of lights, which we can demarcate into towns and cities and we can gaze at their connections through linking road lights. However, from the plane we achieve little understanding of the social, cultural, and political differences that exist in these blobs of light (Nichols & Newsome 1999). With respect to fMRI, this analogy is supported on a technical level, as the details of the relationships between metabolic demand and increased neuronal activity are poorly understood.

Second, as important as brain functioning is, brain imaging provides only one window of many into the multiple influences on behavior that can be relevant to understanding why a person acted in an antisocial manner. Such influences include the intricate interaction between genetic, prenatal, endocrinological, social, cultural, and economic factors: "No pixel in a brain will ever be able to show culpability or nonculpability" (Gazzaniga 2005, p. 100).

Third, despite showing remarkable consistency with lesion, animal model, and electrophysiological data, brain imaging is not yet in Kuhnian terms a "pure science." Interpretation of brain scans is admittedly somewhat subjective. Anatomical landmarks in the form of gyri and sulci differ very much from individual to individual, and even in adulthood the brain is not fixed, but shows plasticity and change in response to injury that also varies from individual to individual. Moreover, in the case of fMRI, differences in haemodynamic response may not necessarily relate to neuropathology, but to vascular and endocrinological pathology. Thus, even if brain abnormalities are found, individual differences in the extent and location of the injury, and in recovery and plasticity, present major problems for the interpretation of brain images in the legal setting.

While these problems can be reduced in research through averaging across many individuals, these are critical issues when examining a single individual. For example, all the brain-imaging studies conducted on violent and antisocial populations have studied group effects. Moreover, most studies have examined adult males, and the results cannot be generalized to females and children. Accordingly, if brain imaging is to be applied to the forensic evaluation of the single patient, a standardized set of tests, procedures, and imaging parameters are needed to achieve more valid conclusions (see appendix 3).

Fourth, correlations between brain function and criminal behavior are imperfect, calling into question both the diagnostic and predictive validity of brain-imaging evidence. That is, brain defects are not observed in all violent criminals, and conversely, not all people with PFC damage exhibit antisocial behavior. For example, one longitudinal case study showed PFC damage to result in personality changes, but without signs of antisocial behavior (Mataró et al. 2001). Some studies have shown how prefrontal damage can even decrease antisocial behavior (Ellenbogen et al. 2005). Differences in the PFC may also be caused by other variables, including levels of education and alcoholism (Laakso et al. 2002). A similar pattern emerges for the amygdala, where damage can result in increased or decreased aggression (LeDoux 1996; Blair 2004). Moreover, in court proceedings, many experts have argued against the use of ambitious speculations concerning the brain (e.g., State of Tennessee v. Paul Dennis Reid Jr., 2002, No. 38887), particularly where the link between the criminal act and the neurological damage is based solely on brain-imaging data.

Fifth, just as it would be inappropriate to expect full localization of criminality genetically (Jones 2006), it would be inappropriate to expect full localization of criminality neurologically (Abbott 2001). Indeed, sociologists have long provided explanations for crime and deviance without the slightest reference to the brain.

Sixth, brain images are not only powerful, they can potentially be too powerful – an effect we have referred to as the "Christmas tree phenomenon." For example, in much the same way that a prosecutor may sway jurors with sympathetic pictures of the innocent victim, the defense may show brightly colored images of the perpetrator's allegedly dysfunctional brain. The vividness and technological sophistication of the images may be overweighted by the jurors, which can warp justice just as surely as can underweighting of relevant evidence. Brain imaging can be admissible in courts of different jurisdictions (e.g., under the Federal Rules of Evidence in the United States). However, given the increasing public interest in brain imaging (Racine 2005) and the misinterpretations of what brain imaging is and can do (Illes et al. 2003), it is crucial for proper legal decision-making that judges and jurors understand the limitations of brain imaging.

7 Concluding Remarks

The goals of science and of law are different. However, important legal questions such as moral blameworthiness, culpability, responsibility, and the likelihood of recidivism depend to some degree on improved understandings of human behavior. Therefore, biological advances in understanding human brain architecture and function may overlap in important ways with legal inquiries. New studies of the criminal brain are likely to shape moral views on responsibility and free will, with possible impacts on how legal systems punish and treat criminals (Greene & Cohen 2004).

A growing body of research gives us good reason to believe that some kinds of brain dysfunction can affect the probability of different kinds of criminal behaviors. However, despite our growing knowledge of the brain abnormalities associated with antisocial and psychopathic behavior, there are as yet no concrete biological markers – genetic or physiological – that can predict such behaviors. Violent and antisocial behaviors undoubtedly arise from a symphony of factors. Optimal understanding will require cooperation among many disciplines such as economics, sociology, psychology, evolutionary biology, cellular physiology, and neuroscience (Jones & Goldsmith 2005).

Acknowledgments

We thank Cindy C. Hagan, John Monaghan, and Joshua Buckholtz for their helpful comments.

Appendix 1: Should We Rethink Free Will?

Research linking the brain to antisocial and criminal behavior also raises neurophilosophical questions concerning our liberty. Most neuroscientists hold that "minds are simply what brains do" (Minsky 1986). Indeed, with the omission of metaphysical constructs like the "mind," many take the view that we are tied to the physical brain and, as a consequence, have little personal choice. A series of classic yet controversial studies by Benjamin Libet and colleagues showed that brain activity associated with deliberate decisions can be detected shortly before we are conscious of making the decision (Libet et al. 1983). In these studies, participants reported when they first felt the intention to make a spontaneous movement by noting the position of a dot moving on computer screen. They apparently first became aware of their intentions about 200 milliseconds before action execution, which is later than the onset of the so-called readiness potential (or *Bereitschaftspotential*) recorded from the scalp prior to movement. Despite criticisms about the accuracy of this timing method, recent research (Lau et al. 2006, 2007) has shown that if anything, the actual onset of conscious intention is likely to be even later. Moreover, psychologists report that our attributions of agency to actions are often illusory (Wegner 2002).

Despite these claims, free will as a concept is still unlikely to be eliminated. Clearly free will is a prerequisite for moral agency, and for society to run smoothly, we all need to believe that we are in full control of our actions. Not surprisingly, some have tried to find a middle ground in this argument. For example, Raine has entertained the idea that free will should be viewed along a "dimension rather than a dichotomy" (Raine 1993, p. 320), while Gazzaniga has argued that "brains are automatic, but people are free" (Gazzaniga 2005, p. 98). Is it reasonable, however, to posit that some people are more free than others? For example, few can dispute the fact that brain diseases such as schizophrenia and Huntington disease reduce the ability to act freely. Nonetheless, most juries may never have explicitly discussed the concept of free will (Gazzaniga 2005). Neurophilosophy may play an important role in understanding and updating the intuitions concerning free will and responsibility that may implicitly underlie juror deliberations.

Appendix 2: Brain Fingerprinting and Lie Detection

Lie detection technology is one of the most obvious legal uses of brain imaging, and several new companies (e.g., No Lie MRI) are beginning to commercialize their services to lawyers and prosecutors. However, despite there being several published empirical studies on lie detection, results seem to be far from conclusive. Early brain-imaging studies of how the brain responds when we willfully lie showed that specific zones of the PFC increase in activity when individuals lie – the same regions known to come online when tasks become more difficult and when we need to control or inhibit responses (Spence et al. 2001). However, one problem with most studies of lie detection is that they use group averages, which make firm conclusions about individual cases impossible. Although more work is needed, recent studies on single individuals have shown promise, with lie detection accuracy in the range of 80%–90% (Kozel 2005). Proponents argue that the use of brain imaging to detect deception is less prone to countermeasures, making it more reliable than the polygraph test (Honts et al. 1994). Not surprisingly, government institutions have become increasingly interested (e.g., U.S. Department of Defense) and have been criticized as being "Orwellian." However, like the polygraph, brain imaging is unlikely to be universally admissible in court until it is shown to be valid, reliable, and relevant.

Another technique – brain fingerprinting – uses electroencephalography to examine the memory- and encoding-related multifaceted electroencephalographic response (MERMER). To measure this, an individual is shown crime scene pictures (i.e., the murder weapon), and changes in brain activity (specifically the P300 component) are monitored. The brain reacts differently to images it recognizes versus ones that it does not recognize, so, for example, if an individual did use a specific weapon to kill a person, the brain will react differently to images of the murder weapon than to images of other weapons not used in the crime. Brain fingerprinting evidence has been admitted in some cases, such as in the Iowa murder trial of Terry Harrington. However, despite its claimed potential, brain fingerprinting has been criticized for problems with developing adequate test stimuli, vulnerability to countermeasures, and – because it's patented – a failure to be appropriately verified by peer review (Rosenfeld 2005).

Appendix 3: Plausible Uses of Brain Imaging and Questions for Future Research

Questions for which brain imaging might provide useful answers:

- Does the defendant exhibit any neurological damage?
- Do the brain abnormalities fit with the nature of the crime?
- Is the defendant faking an illness?
- Is the defendant lying about the crime?
- What is the likelihood of future transgressions?

To begin to answer such questions, society needs the following:

- More neurobiological research on antisocial and criminal populations (e.g., postmortem histology, diffusion tensor imaging, and brain morphometry).
- A better classification of the neural activity associated with different types of criminal activity.
- A set of criteria and parameters for using imaging on single individuals with and without neurological abnormalities.

- Better understanding of the effects of intrinsic and extrinsic factors on the brain (e.g., interplay between environment, development, and genetics).
- Agreed criteria concerning validity and reliability of brain images.
- Agreed procedures for presenting imaging evidence in the courtroom.

References

Abbott, A.: Into the mind of a killer. Nature 410, 296–298 (2001)

Adolphs, R., Gosselin, F., Buchanan, T.W., Tranel, D., Schyns, P., et al.: A mechanism for impaired fear recognition after amygdala damage. Nature 433, 68–72 (2005)

Amodio, D.M., Frith, C.D.: Meeting of minds: The medial frontal cortex and social cognition. Nature Reviews: Neuroscience 7, 268–277 (2006)

Anderson, S.W., Bechara, A., Damasio, H., Tranel, D., Damasio, A.R.: Impairment of social and moral behavior related to early damage in human prefrontal cortex. Nature Neuroscience 2, 1032–1037 (1999)

Aron, A.R., Fletcher, P.C., Bullmore, E.T., Sahakian, B.J., Robbins, T.W.: Stop-signal inhibition disrupted by damage to right inferior frontal gyrus in humans. Nature Neuroscience 6, 115–116 (2003)

Baird, A.A., Fugelsang, J.A.: The emergence of consequential thought: Evidence from neuroscience. Philosophical Transactions of the Royal Society of London: Series B, Biological Sciences 359, 1797–1804 (2004)

Bechara, A., Van Der Linden, M.: Decision-making and impulse control after frontal lobe injuries. Current Opinion in Neurology 18, 734–739 (2005)

Birbaumer, N., Veit, R., Lotze, M., Erb, M., Hermann, C., et al.: Deficient fear conditioning in psychopathy: A functional magnetic resonance imaging study. Archives of General Psychiatry 62, 799–805 (2005)

Blair, R.J.: The roles of orbital frontal cortex in the modulation of antisocial behavior. Brain and Cognition 55, 198–208 (2004)

Blair, R.J., Cipolotti, L.: Impaired social response reversal. A case of acquired sociopathy. Brain: A Journal of Neurology 123, 1122–1141 (2000)

Blair, R.J., Morris, R.S., Frith, C.D., Perrett, D.I., Dolan, R.J.: Dissociable neural responses to facial expressions of sadness and anger. Brain: A Journal of Neurology 122, 883–893 (1999)

Bremner, J.D.: Effects of traumatic stress on brain structure and function: Relevance to early responses to trauma. Journal of Trauma & Dissociation 6, 51–68 (2005)

Bucy, P.C., Kluver, H.: An anatomical investigation of the temporal lobe in the monkey (Macaca mulatta). Journal of Comparative Neurology 103, 151–251 (1955)

Bunge, S.A., Dudukovic, N.M., Thomason, M.E., Vaidya, C.J., Gabrieli, J.D.: Immature frontal lobe contributions to cognitive control in children: Evidence from fMRI. Neuron 33, 301–311 (2002)

Burns, J.M., Swerdlow, R.H.: Right orbitofrontal tumor with pedophilia symptom and constructional apraxia sign. Archives of Neurology 60, 437–440 (2003)

Caspi, A., McClay, J., Moffitt, T.E., Mill, J., Martin, J., et al.: Role of genotype in the cycle of violence in maltreated children. Science 297, 851–854 (2002)

Coricelli, G., Critchley, H.D., Joffily, M., O'Doherty, J.P., Sirigu, A., et al.: Regret and its avoidance: A neuroimaging study of choice behavior. Nature Neuroscience 8, 1255–1262 (2005)

Damasio, A.R.: Descartes' error: Emotion, rationality and the human brain. Putnam, New York (1994)

Dolan, R.J.: Emotion, cognition and behavior. Science 298, 1191–1194 (2002)

Ellenbogen, J.M., Hurford, M.O., Liebeskind, D.S., Neimark, G.B., Weiss, D.: Ventromedial frontal lobe trauma. Neurology 64, 757 (2005)

Fazel, S., Danesh, J.: Serious mental disorder in 23000 prisoners: A systematic review of 62 surveys. Lancet 359, 545–550 (2002)

Feigenson, N.: Brain imaging and courtroom evidence: On the admissibility and persuasiveness of fMRI. International Journal of Law in Context 2, 233–255 (2006)

Frontline, The killer at Thurston High. Arlington, VA: Public Broadcasting Service (2000), http://www.pbs.org/wgbh/pages/frontline/shows/kinkel (accessed, 18 February 2007)

Gazzaniga, M.S.: The ethical brain. Dana Press, New York (2005)

Giedd, J., Blumenthal, J., Jeffries, N.O., Castellanos, F.X., Liu, H., et al.: Brain development during childhood and adolescence: A longitudinal MRI study. Nature Neuroscience 2, 861–863 (1999)

Goldberg, E.: The executive brain: Frontal lobes and the civilized mind. Oxford University Press, Oxford (2001)

Golding, S.L., Roesch, R., Schreiber, J.: Assessment and conceptualization of competency to stand trial: Preliminary data on the Interdisciplinary Fitness Interview. Law and Human Behavior 8, 321–334 (1984)

Gordon, H.L., Baird, A.A., End, A.E.: Functional differences among those high and low on a trait measure of psychopathy. Biological Psychiatry 56, 516–521 (2004)

Gottfredson, M.R., Hirschi, T.: A general theory of crime. Stanford University Press, Stanford (1990)

Grafman, J., Schwab, K., Warden, D., Pridgen, A., Brown, H.R., et al.: Frontal lobe injuries, violence, and aggression: A report of the Vietnam Head Injury Study. Neurology 46, 1231–1238 (1996)

Greene, J.D., Cohen, J.D.: For the law, neuroscience changes nothing and everything. Philosophical Transactions of the Royal Society of London: Series B, Biological Sciences 359, 1775–1785 (2004)

Gregg, T.R., Siegel, A.: Brain structures and neurotransmitters regulating aggression in cats: Implication for human aggression. Progress in Neuro-Psychopharmacology and Biological Psychiatry 25, 91–140 (2001)

Harlow, J.: Passage of an iron bar through the head. Boston Medical and Surgical Journal 13, 389–393 (1848)

Heekeren, H.R., Wartenburger, I., Schmidt, H., Schwintowski, H.P., Villringer, A.: An fMRI study of simple ethical decision-making. Neuroreport 14, 1215–1219 (2003)

Honts, C.R., Raskin, D.C., Kircher, J.C.: Mental and physical countermeasures reduce the accuracy of polygraph tests. Journal of Applied Psychology 79, 252–259 (1994)

Illes, J., Kirschen, M.P., Gabrieli, J.D.: From neuroimaging to neuroethics. Nature Neuroscience 6, 205 (2003)

Jones, O.D.: Behavioral genetics and crime, in context. Law and Contemporary Problems 69, 81–100 (2006)

Jones, O.D., Goldsmith, T.H.: Law and behavioral biology. Columbia Law Review 105, 405–502 (2005)

Kiehl, K.A., Smith, A.M., Hare, R.D., Mendrek, A., Forster, B.B., et al.: Limbic abnormalities in affective processing by criminal psychopaths as revealed by functional magnetic resonance imaging. Biological Psychiatry 50, 677–684 (2001)

King, J.A., Blair, R.J.R., Mitchell, D.G.V., Dolan, R.J., Burgess, N.: Doing the right thing: A common neural circuit for appropriate violent or compassionate behavior. NeuroImage 30, 1069–1076 (2006)

Kozel, F.A.: Detecting deception using functional magnetic resonance imaging. Biological Psychiatry 58, 605–613 (2005)

Laakso, M.P., Gunning-Dixon, F., Vaurio, O., Repo-Tiihonen, E., Soininen, H., et al.: Prefrontal volumes in habitually violent subjects with antisocial personality disorder and type 2 alcoholism. Psychiatry Research 114, 95–102 (2002)

LaFave, W.R.: Criminal law, 4th edn. Hornbook series student edition. West Group Publishing, St. Paul (2003)

Lau, H.C., Rogers, R.D., Passingham, R.E.: On measuring the perceived onsets of spontaneous actions. Journal of Neuroscience 26, 7265–7271 (2006)

Lau, H.C., Rogers, R.D., Passingham, R.E.: Manipulating the experienced onset of intention after action execution. Journal of Cognitive Neuroscience 19, 81–90 (2007)

LeDoux, J.E.: The emotional brain: The mysterious underpinnings of emotional life. Simon & Schuster, New York (1996)

Libet, B., Gleason, C.A., Wright, E.W., Pearl, D.K.: Time of conscious intention to act in relation to onset of cerebral activity (readiness potential): The unconscious initiation of a freely voluntary act. Brain: A Journal of Neurology 106, 623–642 (1983)

Mataró, M., Jurado, M.A., García-Sánchez, C., Barraquer, L., Costa-Jussà, F.R., et al.: Long-term effects of bilateral frontal brain lesion: 60 years after injury with an iron bar. Archives of Neurology 58, 1139–1142 (2001)

McLaughlin, J.T., Rosenkranz, E.J., Wei, T.P., Clare, S.M., Haider, A., et al.: Brief of the American Medical Association, American Psychiatric Association, American Society for Adolescent Psychiatry, American Academy of Child and Adolescent Psychiatry, American Academy of Psychiatry and the Law, National Association of Social Workers, and National Mental Health Association as amici curiae in support of the respondent. Roper v. Simmons, US Supreme Court, no. 03-633 (2004), `http://www.abanet.org/crimjust/juvjus/simmons/ama.pdf` (accessed, 22 February 2007)

Meyer-Lindenberg, A., Buckholtz, J.W., Kolachana, B., Hariri, A.R., Pezawas, L., et al.: Neural mechanisms of genetic risk for impulsivity and violence in humans. Proceedings of the National Academy of Sciences of the United States of America 103, 6269–6274 (2006)

Meyers, C.A., Berman, S.A., Scheibel, R.S., Hayman, A.: Case report: Acquired antisocial personality disorder associated with unilateral left orbital frontal lobe damage. Journal of Psychiatry & Neuroscience 3, 121–125 (1992)

Minsky, M.: The society of mind. Simon & Schuster, New York (1986)

Morse, S.: Brain overclaim syndrome and criminal responsibility: A diagnostic note. Ohio State Journal of Criminal Law 3, 397–412 (2006)

Nichols, M.J., Newsome, W.T.: The neurobiology of cognition. Nature 402, C35–C38 (1999)

Racine, E.: fMRI in the public eye. Nature Reviews: Neuroscience 6, 159–164 (2005)

Raine, A.: The psychopathology of crime: Criminal behavior as a clinical disorder. Academic Press, San Diego (1993)

Raine, A., Buchsbaum, M.S., Stanley, J., Lottenberg, S., Abel, L., et al.: Selective reductions in prefrontal glucose metabolism in murderers. Biological Psychiatry 36, 365–373 (1994)

Raine, A., Meloy, J.R., Bihrle, S., Stoddard, J., LaCasse, L., et al.: Reduced prefrontal and increased subcortical brain functioning assessed using positron emission tomography in predatory and affective murderers. Behavioral Sciences & the Law 16, 319–332 (1998)

Raine, A., Lencz, T., Bihrle, S., LaCasse, L., Colletti, P.: Reduced prefrontal gray matter volume and reduced autonomic activity in antisocial personality disorder. Archives of General Psychiatry 57, 119–127 (2000)

Rosenfeld, J.P.: Brain fingerprinting: A critical analysis. Scientific Review of Mental Health Practice 4, 20–37 (2005)

Sapolsky, R.M.: The frontal cortex and the criminal justice system. Philosophical Transactions of the Royal Society of London: Series B, Biological Sciences 359, 1787–1796 (2004)

Scott, C.L.: Roper v. Simmons: Can juvenile offenders be executed? Journal of the American Academy of Psychiatry and the Law 33, 547–552 (2005)

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R.J., et al.: Empathy for pain involves the affective but not sensory components of pain. Science 303, 1157–1162 (2004)

Spence, S.A., Farrow, T.F., Herford, A.E., Wilkinson, I.D., Zheng, Y., et al.: Behavioural and functional anatomical correlates of deception in humans. Neuroreport 12, 2849–2853 (2001)

Takahashi, H., Yahata, N., Koeda, M., Matsuda, T., Asai, K., et al.: Brain activation associated with evaluative processes of guilt and embarrassment: An fMRI study. NeuroImage 23, 967–974 (2004)

Tiihonen, J., Hodgins, S., Vaurio, O., Laakso, M., Repo, E., et al.: Amygdaloid volume loss in psychopathy. Abstracts – Society for Neuroscience 26, 2017 (2000)

van Elst, L.T., Trimble, M.R., Ebert, D., van Elst, L.T., Raine, A.: Dual brain pathology in patients with affective aggressive episodes. Archives of General Psychiatry 58, 1187–1188 (2001)

Vitiello, B., Behar, D., Hunt, J., Stoff, D., Ricciuti, A.: Subtype aggression in children and adolescents. Journal of Neuropsychiatry and Clinical Neurosciences 2, 189–192 (1990)

Volavka, J., Mohammad, Y., Vitrai, J., Connolly, M., Stefanovic, M., Ford, M.: Characteristics of state hospital patients arrested for offenses committed during hospitalization. Psychiatric Services 46, 796–800 (1995)

Volkow, N.D., Tancredi, L.: Neural substrates of violent behaviour. A preliminary study with positron emission tomography. British Journal of Psychiatry 151, 668–673 (1987)

Walker, P.L.: A bioarchaeological perspective on the history of violence. Annual Review of Anthropology 30, 573–596 (2001)

Wegner, D.M.: The illusion of conscious will. MIT Press, Cambridge (2002)

Weiss, E.M., Kohler, C.G., Nolan, K.A., Czobor, P., Volavka, J., et al.: The relationship between history of violent and criminal behavior and recognition of facial expression of emotions in men with schizophrenia and socioaffective disorder. Aggressive Behavior 32, 1–8 (2006)

Widom, C.S.: The cycle of violence. Science 244, 160–166 (1989)

Woermann, F.G., Tebartz van Elst, L., Keopp, M.J., Free, S.L., Thompson, P.J., et al.: Reduction of frontal neocortical grey matter associated with affective aggression in patients with temporal lobe epilepsy: An objective voxel by voxel analysis of automatically segmented MRI. Journal of Neurology, Neurosurgery, and Psychiatry 68, 162–169 (2000)

Yang, Y.L., Raine, A., Lencz, T., Bihrle, S., Lacasse, L., et al.: Prefrontal structural abnormalities in liars. British Journal of Psychiatry 187, 320–325 (2005)

16

The Controversy over Brain Research

Hans Küng

Global Ethic Foundation
Waldhäuser Strasse 23
D-72076 Tübingen
Germany
office@weltethos.org

Summary. All mental processes are closely connected with the electro-chemical processes between the nerve cells in the brain, and these function in accordance with the natural laws of physics. But is free will therefore an illusion? The more precisely the neuroscientists can describe the ways in which our brain functions, the clearer it becomes that none of their measurements and models embraces the central aspect of consciousness: how we become subjectively aware of qualities such as color or smell, a reflection or an emotion. The discussion between the "scientists" and the "philosophers" in our symposium has shown that at present brain research seems not to have an empirically demonstrated theory to offer about the connection between brain and mind, between consciousness and the nervous system. In any case, chemistry and physics seem not to explain the experience of freedom of choice which is however universal and undeniable.

Keywords: Brain, determinism, free will, responsibility, self.

In this chapter I presuppose that without the brain there is no human intellect and without the activities of particular brain centers there can be no intellectual achievement. However, that raises the question which is my first train of thought.

1 Determined by Physical-Chemical Brain Processes?

The fascination of brain physiologists by the object of their research is understandable: with the human brain, evolution has beyond question produced its *natural top product*. The brain is by far the most complex structure in the whole universe, by comparison with which even a complicated computer works in a very simple way. This "grey mass" (large only by comparison with the smaller brain of the anthropoids) with its hollows and ridges displays structural levels and spheres of functions in which more than ten billion brain cells are at work with the help of thousands of billions of links and guiding connections, extending over more than 100,000 kilometers. The processes of the brain are the result of both genetic disposition and social learning. In her pioneering investigations Herta Flor has confirmed that the brain is not a mass fixed from early childhood, as was long believed. Rather, it shows an amazing plasticity, power of regeneration, and capacity for change, and at the same time proves to be amazingly stable for our perception of ourselves and the world.

Neurophysiologial brain research has given us great insights in recent years. With the help of new imaging procedures it has produced impressive new discoveries. After all this research it has become clear beyond dispute that all mental processes are closely connected with the electro-chemical processes between the nerve cells in the brain, and these function in accordance with the natural laws of physics. Whatever conclusions one draws from this, no philosopher or theologian should enter into discussion with a neurobiologist without taking these physical and biological presuppositions seriously and recognizing the human potential of brain physiology. Those who, like a Roman Catholic moral theologian in a recent press interview, seek to support the freedom of the will by over-hastily and dogmatically introducing into the debate a God who wants to be freely loved by human beings in return, will lose scientists from the start. They will forfeit the opportunity, having fully recognized his scientific achievement, of then asking the brain researcher bluntly whether he might not himself have his own dogmatic prejudice. For just as philosophers and theologians should consider biological brain research, so brain researchers should consider questions of philosophy and theology.

Instead of getting bogged down in trench warfare, here too I would like to build bridges. Hence now my second train of thought, another question:

2 Is Free Will an Illusion?

Brain researchers have established more and more correspondence between the appearance of a particular process or state of consciousness and the activity of a particular region of the brain (identifiable macroscopically) or the (microscopic) circuits of neurons from which the region of the brain is constructed. These insights are indubitable and welcome. However, neurophysiologists have now begun to draw momentous conclusions about the self or the self-consciousness of the human being from this evidence: they argue that while we certainly experience that we are free in our will, decision, and action, science shows us that we are deceiving ourselves. The brain with its unconscious neuronal processes constantly precedes our will.

For example, the Bremen researcher Gerhard Roth attributes to the limbic system with the basal ganglia hidden deep in our brain the "ultimate decisions of human beings": the conscious I, he says, is "not the real master of our actions" and "*freedom of will in the strict sense is a delusion.*"

> In our thinking, feeling, willing, our planning for action and the execution of our actions we human beings feel that we are free. Here our self feels that it is the cause of these states and actions. But this is evidently an illusion. Rather, psychological and neuroscientific experiments and observations show that thoughts and intentions which come to our mind are largely caused and guided by the limbic system, which has a particularly strong effect on the frontal brain. (Roth 2004; cf. Roth 2003)

So the feeling of being the author of our actions is as stubborn a delusion as the early notion that we human beings are at the center of the universe. In fact, all our purposes and decisions, ideas and wishes, are determined by physiological processes. Everything is guided by the unconscious, *by the limbic system*, where for example even in childhood the decision is made as to whether or not one will become a sexual offender. Note what consequences such an application of neurophysiological insights would have for law and ethics.

So are all our everyday experiences of freedom deceptive? Or, to put it the other way round, are such conclusions from neurological experiments perhaps colored by conscious or unconscious philosophical assumptions? Wolf Singer of the Max Planck Institute in Frankfurt also claims that our intuition errs "dramatically" in thinking that an I authority is responsible for decisions (Singer 2004). Singer does not see any essential differences between conscious processes of the brain guided by us, and unconscious, automatic processes. Singer's view wants to take account of the "trivial insight that a person did what he or she did because he or she could not do anything else at the moment in question, else he or she would have acted otherwise" (Singer 2004). So? In logic this is known as begging the question, a circular argument which presupposes what it wants to prove: "It could not be otherwise because it could not be otherwise." A circular argument easily arises when a brain researcher empirically establishes only what is structured by his consciousness and is attested with its help. This leads to a third train of thought:

3 The Trivialization of Responsibility and Guilt by the Neurosciences

What amazes me in such arguments is with what nonchalance, on the basis of very short-winded experiments like moving a finger or an arm, a brain researcher such as Roth foists his neurological hypothesis of the illusion of the freedom of the will as a "deep-rooted foundational problem" on the criminal law. This, he says, wrongly maintains a "principle of guilt and responsibility" which presupposes "all capacities of human beings to decide freely and rightly between right and wrong" (Roth 2003).

Of course the criminal law recognizes a limited capacity for guilt. But is the mental in principle merely an epiphenomenon of the neuronal? What "relief" does such a neurological hypothesis bring the criminal: no guilt feelings – everything is illusion! I don't want to discuss the horrific Nazi crimes against humanity. But at the same time as Roth's report there was a terrible account in the German press of a clique of adult men and women in the Saarland who gang-raped a five-year-old boy and finally killed him. So are such monsters and all the adults who in Germany abuse at least 15,000 children every year unfree because of the mechanisms of the limbic system and therefore relieved from guilt and responsibility by a perfect scientific excuse? The victims and their parents will have little time for such a neurological trivialization of the guilt of child abusers. Instead of reflecting in a differentiated way on personal responsibility and guilt (and of course also positively on merits), to appeal only to a "violation of social norms," as Roth does, seems empty in the face of almost total indifference to such social norms.

Authorities on forensic psychiatry such as Hans-Ludwig Kröber of Berlin find "suspicious the tendency of some brain researchers also to come forward as interpreters of the brain and to proclaim to an audience of lay people and amazed journalists with the aid of many colorful pictures that the freedom of the will is refuted and that criminal responsibility is a fiction.… In reality it is a very long way from the images from the PET, the functional positron emissions tomograph, to the question of criminal responsibility" (Kröber 2004, pp. 107–8). So when are we criminally responsible? "When we are in a position to make our decisions dependent on rational considerations, in other words when we are in a position to evaluate our wishes critically" (Kröber 2004, p. 109).[1]

Today medicine has at its disposal the most modern pieces of diagnostic equipment, a combination of a computer tomograph (CT) and a positron emissions tomograph (PET), which allow the smallest clusters of cancer cells to be recognized at an early stage. But unfortunately neurological hypotheses which declare that our understanding that we are free human beings is a delusion are partly to blame for the fact that the brain research which is making fantastic progress with the help of such instruments is today not just evoking hopes in the fight against serious diseases such as Alzheimer's, Parkinson's, schizophrenia, depression, and the regaining of autonomy and freedom of decision. It is also encouraging anxieties that we human beings will become cold bio-automatons; guided by neurons we could be exposed to every possible intervention to manipulate the consciousness and thus lose our identity and autonomy.

Happily, however, brain researchers too are now becoming increasingly aware of the problems of such reductionist procedures. So now, after assessing the progress of brain research, in a fourth train of thought it is time to show its equally clear limits.

[1] Cf. also the warning by the Frankfurt criminal lawyer K. Lüderssen (2004), about brain research "which (probably innocently) has succumbed to the danger of a self-suggestive metaphysic" (p. 102).

4 The Limits of Brain Research

Functional magnetic resonance imagery gives us information – often very crude – about "where" things are in the brain, but not "how" cognitive achievements by neuronal mechanisms are to be described. It is never possible to read the feelings and thoughts of a person simply from the colorful patterns which the tomographs produce from his or her brain activity. Of course there are countless reflections on the biological foundations of an I-consciousness, but can these interesting speculations really overcome the gap in explanation between physical processes and consciousness? No, the more precisely the neuroscientists can describe the ways in which our brain functions, the clearer it becomes that none of their measurements and models embraces the central aspect of consciousness: by becoming subjectively aware of qualities such as color or smell, a reflection or an emotion.

This is not just a careless remark by a theologian, but rather a surprising concession by advanced neurophysiologists. A few months after Gerhard Roth's striking publications, in 2004, eleven leading German neuroscientists – remarkably also including Roth and Singer, who have already been cited – published a "Manifesto on the Present and Future of Brain Research" (Das Manifest).[2] They said that the impression had been given that brain research was "on the threshold of wresting its last secrets from the brain." To reassure an alarmed public, they now drew up a sober and balanced assessment of their young science, which was storming boldly ahead.

They said that significant *progress* was being aimed at with the help of new methods:

- on the one hand on the *uppermost* level: research was being done into the functions and interplay of larger areas of the brain and thus a thematic division of the brain according to complexes of functions;

- on the other hand at the *lowest* level: today we largely understand the processes at the level of individual cells and molecules: the origin and further communication of neuronal stimulation;

- but not on the *middle* level: we know "terrifyingly little" of what goes on within hundreds or thousands of groups of cells: "We are completely ignorant about what takes place when hundreds of millions or even a billion nerve cells 'talk' to one another." (Das Manifest, pp. 30–33)

[2] Das Manifest: Über Gegenwart und Zukunft der Hirnforschung, in *Gehirn und Geist: Das Magazin für Psychologie und Hirnforschung,* 6 (2004), 30–37. The manifesto is signed by professors Christian Elger (Bonn), Angela Friederici (Leipzig), Christof Koch (Pasadena), Heiko Luhmann (Mainz), Christoph von der Malsburg (Bochum/Los Angeles), Randolf Menzel (Berlin), Hannah Monyer (Heidelberg), Frank Rösler (Marburg), Gerhard Roth (Bremen), Henning Scheich (Magdeburg), and Wolf Singer (Frankfurt am Main).

All this amounts to an *ignorance precisely at the decisive level of brain activity*. For this is where thoughts and feelings, purposes and effects, consciousness and self-consciousness are made possible:

> We still do not understand even the beginning of what rules the brain works by; how it depicts the world in such a way that direct perception and earlier experience fuse; how the inner action is experienced as "its" activity and how it plans future actions. Furthermore, it is not at all clear how we can investigate it with present possibilities. In this respect to some degree we are still at the stage of hunters and gatherers. (Das Manifest, p. 33)

Praiseworthy academic modesty! So to a fifth train of thought:

5 The Big Questions of the Neurosciences

Fortunately, the neuroscientists who subscribed to the "Manifesto on the Present and Future of Brain Research" show themselves restrained about the "big questions":

> How do consciousness and the experience of being a self arise, how are rational and emotional action linked, what about the notion of "free will"? Today it is already permissible to ask the big questions of the neurosciences – however, it is unrealistic to think that they will be answered in the next ten years. It even remains questionable whether we can approach them meaningfully by then. For that we would need to know essentially more about the way in which the brain functions. (p. 34)

One can only agree. Even such a refined picturing procedure as a "cyber-phrenology" cannot in fact fulfill the dream of an embodiment of the mind. Some hope that a theoretical neurobiology will one distant day supplement classical brain research, as quantum physics supplemented classical mechanics, then making it possible "so to speak to understand the small uniqueness of the brain." That may be, but it means that at present brain research has no empirically demonstrated theory to offer about the connection between brain and mind, between consciousness and the nervous system. To this degree one may hope that in future all brain researchers will refrain from reductionist statements and keep to the closing sentences of their manifesto:

> But all the progress will not end in a triumph for neuronal reductionism. Even if at some point we have explained all the processes of the neuron which underlie human sympathy, being in love or moral responsibility, the distinctive feature of this "internal perspective" nevertheless remains. For even a Bach fugue does not lose its fascination when one has understood precisely how it is constructed. Brain research will have to distinguish clearly between what it can say and what lies outside its sphere of competence, just as musicology – to keep to this example – has something to say about Bach's fugue, but can have no explanation of its unique beauty. (p. 37)

There is a wealth of confirmation of this antireductionist view. My Tübingen colleague, the famous behavioral neurologist Niels Birbaumer, also recommends to his colleagues "modest restraint in the generalization and interpretation of neurobiological data." He remarks that he cannot say whether or not the will is free, as this cannot be measured:

> Neither free nor unfree will can be observed, as we have no neuronal correlate of freedom. Freedom is certainly also a construct of the brain like all other behaviour and thought that human beings produce, but it is also and primarily a phenomenon which has grown up historically, politically and socially, and cannot just be derived from processes of the brain. (Birbaumer 2004, p. 28)

The change of position of the American brain physiologist Benjamin Libet is interesting in this connection. As early as 1985 he was the first to carry out the much-cited behavioral physiological experiments which showed that the brain builds up a neuronal "readiness potential" (for example on raising the right or left finger or arm – a very small unit of the will), which is said to precede the subjectively experienced will to act by 350 to 400 milliseconds (cf. Libet 1999, 2004). But does this "readiness potential" bind the will? In 1999 Libet then explained that the consciousness that lags behind in time is in a position to stop what the brain suggests as an action. So in all the pressure to act, "free will" at least has the power of veto. Libet's conclusion is now that "the existence of a free will is at least as good a scientific option as denying it by the deterministic theory, if not a better one" (Libet 1999, p. 55.).

Besides, scientists have only recently begun to analyze these short-winded experiments. It is pointed out that the experimenter communicates impulses to the brain simply through the attempted experiment and this immediately prompts unconscious neuronal activity. An analysis of the prehistory of an individual which made it possible for the decision processes of our own brain not to succumb to the limbic reflex in a particular situation would certainly be more illuminating than the analysis of milliseconds before a programmed movement of the finger. It is precisely here that we have freedom of the will: in the capacity of human beings to set themselves values and goals and pursue them in action, independently of the external and internal control of others, but rather in self-control, in autonomy, in the self-legislation of the self. But in reality is there a self at all?

6 Chemistry and Physics Do Not Explain the Self

Unlike the authors of the brain researchers' manifesto, Wolfgang Prinz of the Max Planck Institute for Cognition and Neurosciences in Munich thinks it is far from being demonstrated that on the basis of brain research "'our' picture of the human being has been considerably shaken." Like the beauty of a Bach fugue, he argues, so too the picture of the human being can remain untouched by any reduction and deconstruction: however, what certainly has to be revised is the naturalism which shapes this image of the human being and also that of some brain researchers, but which is hardly reflected on. Human beings are what they are not just through their nature but above all through their culture, and are so through and through, to the deepest roots of their cognitive achievements and the innermost corners and convolutions of their brains. "Therefore brain research can certainly do a lot here, but not everything. At all events it is of no use as the new leading discipline of the humanities, which it would very much like to be" (Prinz 2004b, p. 35). In a conversation Prinz is even clearer: “Biologists can explain how the chemistry and physics of the brain function. But so far no one knows how the experience of these comes about and how the brain produces meanings” (Prinz 2004a, p. 26).

The Berlin philosopher Peter Bieri regards the alleged empirical refutation of the freedom of the will as “a bit of adventurous metaphysics”:

> People look in vain in the material composition of a painting for the depiction or the beauty, and in the same sense one looks in vain in the neurobiological mechanics of the brain for freedom or a lack of freedom. There is *neither* freedom *nor* unfreedom there. The brain is logically the wrong place for this idea.... Our will is free if it controls our judgment as to what is rightly to be wanted. It is unfree if judgment and will fall apart. (Bieri 2005; cf. Bieri 2001)

Taking up Peter Bieri, the philosopher Jürgen Habermas makes a sharp distinction between causes and grounds: “Anyone who is subject to the causal compulsion of an imposed limitation,” that is, to a compelling cause, is in fact unfree. But anyone who “is subject to the uncompelling compulsion of the better argument” and decides on an action for particular reasons is free. The bending of an arm or finger induced by an experimenter is not a free action in terms of moral responsibility. Moral responsibility is always the result of a complex interweaving of considerations about means and ends, resources and obstacles, which have to be weighed up. Interpersonal communication, which is at the center of interest for Habermas and his ethics of discourse, is not a “blind event of nature” that runs its course as it were behind the subject’s back. Already in the newborn child the human spirit develops only in social interaction, through cooperation and instruction. And to this degree the spirit by no means resides only in the brain, but is “embodied” in the whole human person. The self may be a social construction, but that does not mean that it is an illusion (Habermas 2004). Yet another aspect is important to me, my last train of thought:

7 Experience of Freedom

In their everyday self-understanding even brain researchers constantly presuppose responsible authorship in themselves, their colleagues, and the patients. Simply to explain this self-understanding as an epiphenomenon betrays a deterministic dogmatism which needs to be investigated. Here the laboratory perspective needs to be expanded by the perspective of the world in which we live, and external and internal views need to be dovetailed. Alongside the neurological method, introspection is by no means to be despised. After all, in practice it must also be used constantly by neurophysiologists if they want to interpret their images and the processes they have established. They must then also “look into themselves” instead of into the magnetic resonance imager: the self-observation which is possible for anyone, supported by the observation of the conduct of others, cannot only look back. It can even grasp psychological processes as they are happening.

Of course everyone has his or her own perspective on things, as the psychiatrist Manfred Spitzer of the University of Ulm observes:

> So for me things are again quite different from what they are for someone who so to speak looks at me from outside. For me the sky is blue. But anyone who roots around in my head, by whatever means, will not find anything blue. And just as I can always decide for myself here and now, so it can be that *someone who roots around in my head will never find this freedom.* Nevertheless: for myself I am always free, just as for me the sky is always blue. (Spitzer 2004 italics in original)

Spitzer, who is very concerned that this insight should be used in education, even thinks that "The better we get to know the machinery of our actions and decisions in a neurological way, the freer we become" (Ibid.).

In fact, everyone can establish for himself and herself that however much I am dependent and determined externally and internally in my whole being, I am still aware that this or that is ultimately up to me: whether I speak or keep silent, get up or remain sitting, prefer this or that drink or garment, this or that journey. However much my brain decides spontaneously that my eye will look at someone or my foot will evade an obstacle, as soon as this is not just a short physical movement (such as raising an arm or finger) as in those experiments but lengthy processes which require my reflection – for example, the choice of a profession, the acceptance of a job, the choice of a partner for life – I must grapple with different thoughts and alternative courses of action; I must decide, and in some circumstances correct my decision. Here the whole of my life comes into view.

The Tübingen developmental biologist Alfred Gierer is therefore right when alongside neurophysiology and introspection he emphasizes that our *deliberate actions* are an access to our consciousness and our freedom. This would also relate to the age-old fundamental problem of free will: "Presumably the will of others cannot be disclosed completely with objective means. We do not even know ourselves sufficiently – our look inwards is incomplete – and in many respects we experience ourselves first in our own actions" (Gierer 2005, p. 45).

So freedom is an experience not just of thought and feeling but of action. But it is also an experience not only of doing, but also, I would add, of not-doing, failing, and incurring guilt. For in my actions I can also directly experience this negative aspect: I have not done it but I should have done it; I have given a promise but not kept it; I am guilty, I acknowledge my guilt and ask for forgiveness; but I also require from others an acknowledgment of their guilt where I was not guilty; in the end, they were completely free to do this.

Indeed, what would morality be without responsibility, what would responsibility be without freedom, what would freedom be without a tie? And that brings us to the need for some elementary shared ethical values, criteria, and norms, in other words to the need for an ethic of humankind which shares basic elements, a global ethic. But that would be a great theme of its own, one which has occupied me daily for decades, and that does not need to be discussed here. I think I have already supplied sufficient material for discussion, and end with my thanks for your attentiveness.

Acknowledgment

The ideas presented here are developed at greater length in H. Küng, *The Beginning of All Things: Science and Religion* (Grand Rapids: Eerdmans, 2007); German: *Der Anfang aller Dinge: Naturwissenschaft und Religion* (Munich: Piper, 2005).

References

Bieri, P.: Das Handwerk der Freiheit: Über die Entdeckung des eigenen Willens. Hanser, Munich (2001)

Bieri, P.: Unser Wille ist frei. Der Spiegel 2, 124–125 (2005)

Birbaumer, N.: Hirnforscher als Psychoanalytiker. In: Geyer, G. (ed.) Hirnforschung und Willensfreiheit: Zur Deutung der neuesten Experimente, pp. 27–29. Suhrkamp, Frankfurt am Main (2004)

Gierer, A.: Biologie, Menschenbild und die knappe Ressource Gemeinsinn. Königshausen & Neumann, Würzburg (2005)

Habermas, J.: Um uns als Selbsttäuscher zu entlarven, bedarf es mehr. Frankfurter Allgemeine Zeitung, 15 November (2004)

Kröber, H.-L.: Die Hirnforschung bleibt hinter dem Begriff strafrechtlicher Verantwortlichkeit zurück. In: Geyer, G. (ed.) Hirnforschung und Willensfreiheit: Zur Deutung der neuesten Experimente, pp. 103–110. Suhrkamp, Frankfurt am Main (2004)

Libet, B.: Do we have a free will? Journal of Consciousness 6, 47–57 (1999)

Libet, B.: Mind time: The temporal factor in consciousness. Harvard University Press, Cambridge (2004)

Lüderssen, K.: Ändert die Hirnforschung das Strafrecht? In: Geyer, G. (ed.) Hirnforschung und Willensfreiheit: Zur Deutung der neuesten Experimente, pp. 98–102. Suhrkamp, Frankfurt am Main (2004)

Manifest, D.: Über Gegenwart und Zukunft der Hirnforschung. Gehirn und Geist: Das Magazin für Psychologie und Hirnforschung 6, 30–37 (2004)

Prinz, W.: Der Mensch ist nicht frei: Ein Gespräch. In: Geyer, G. (ed.) Hirnforschung und Willensfreiheit: Zur Deutung der neuesten Experimente, pp. 20–26. Suhrkamp, Frankfurt am Main (2004a)

Prinz, W.: Neue Ideen tun Not. Gehirn und Geist 6, 35, 37 (2004b)

Roth, G.: Aus Sicht des Gehirns. Suhrkamp, Frankfurt am Main (2003)

Roth, G.: Das Ich auf dem Prüfstand – Die Hirnforschung und ihre Sicht vom Menschen, radio broadcast on SWR2; id, Aus Sicht des Gehirns. Frankfurt am Main (June 10, 2004)

Singer, W.: Selbsterfahrung und neurobiologische Fremdbeschreibung: Zwei konfliktträchtige Erkenntnisquellen. Deutsche Zeitschrift für Philosophie 52(2), 235–255 (2004)

Spitzer, M.: Es gibt nicht Gutes, ausser man tut es – Die Hirnforschung und die Frage, was uns zum Handeln antreibt, radio broadcast on SWR 2 (June 13, 2004)

Author Index

Index