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

Downward Causation and the Neurobiology of Free Will

Springer

Contents

1 Introduction and Overview
Nancey Murphy 1

Part I: Physics, Emergence, and Complex Systems

2 Free Will, Physics, Biology, and the Brain
Christof Koch 31

3 Human Freedom and "Emergence"
William T. Newsome 53

4 Top-Down Causation and the Human Brain
George F.R. Ellis 63

5 Top-Down Causation and Autonomy in Complex Systems
Alicia Juarrero 83

6 Toward a Complementary Neuroscience: Metastable Coordination Dynamics of the Brain
J.A. Scott Kelso and Emmanuelle Tognoli 103

Part II: Volition and Consciousness: Are They Illusions?

7 Physiology of Volition
Mark Hallett 127

8 How We Recognize Our Own Actions
Sarah-Jayne Blakemore 145

9 Volition and the Function of Consciousness
Hakwan C. Lau 153

Part III: Broader Understandings of Volition and Consciousness

10 Conscious Willing and the Emerging Sciences of Brain and Behavior
Timothy O'Connor 173

11 Contemplative Neuroscience as an Approach to Volitional Consciousness
Evan Thompson 187

12 Free Will and Top-Down Control in the Brain
Chris D. Frith 199

13 Thinking beyond the *Bereitschaftspotential*: Consciousness of Self and Others as a Necessary Condition for Change
Sean A. Spence 211

Part IV: Human Implications of the Debate

14 Criminal Responsibility, Free Will, and Neuroscience
David Hodgson 227

15 Law, Responsibility, and the Brain
Dean Mobbs, Hakwan C. Lau, Owen D. Jones, Chris D. Frith . . . 243

16 The Controversy over Brain Research
Hans Küng 261

Author Index 271

Index 273

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)