1

Philosophy Meets the Neurosciences

William Bechtel, Pete Mandik, and Jennifer Mundale

1 Cognitive Science, Neuroscience, and Cognitive Neuroscience: New Disciplines of the Twentieth Century

The Greek oracle admonished "Know thyself." But for more than two millennia, the only avenues to self-knowledge were to examine one's own thoughts or to review one's behavior. The idea of knowing oneself by knowing how one's brain worked was at best a philosopher's thought experiment. When, in the middle of the twentieth century, the philosopher Herbert Feigl (1958/1967) proposed the idea of an autocerebroscope through which people could examine the activities of their own brains, no one imagined that by century's end we would be close to realizing this fantasy. New tools for studying the brain, especially positron emission tomography (PET) and functional magnetic resonance imaging (fMRI), provide avenues for revealing which brain areas are unusually active when individuals perform specific tasks (see chapter 4, this volume). Knowing which brain areas are activated when a subject performs a given task helps us to better understand the mental processes involved in performing that task. While the false-color images produced by these techniques are captivating, they are only part of what has opened up the study of mental processes in the brain in the last half century. Over the last 150 years careful analyses of behavioral deficits resulting from brain damage have offered further clues about what different brain regions do. In addition, scalp recordings of electrical or magnetic activity (through the electroencephalograph or EEG machine, or the magnetoencephalograph or MEG machine) evoked by particular stimulus events have provided detailed information about the time course of brain processing.

The second half of the twentieth century has witnessed not just an explosion of research at the behavioral end of neuroscience, but also extraordinary advances in understanding the basic cellular, synaptic, and molecular processes in the brain. The idea that neurons constitute the basic functional, cellular units of the brain was not widely accepted until the beginning of the twentieth century, but this provided the foundation for subsequent micro-level research (see chapter 3, this volume). This research, in turn, has increasingly been integrated with research into higher brain

functions. For example, the introduction of electrodes that allow recording from awake, behaving primates has provided a means for linking local neural behavior to particular cognitive tasks. More recently new tools for manipulating genetic material have extended the inquiry even further. These advances have required the collaboration of scientists trained in a number of specialties, including neuroanatomy, neurochemistry, and neurophysiology. In the 1960s, the term *neuroscience* was introduced for this collaborative inquiry, which has since expanded rapidly; the Society for Neuroscience, founded in 1970, now has 25,000 members.

Simultaneous with the introduction of new techniques for studying the brain has been the development of sophisticated ways of analyzing behavior so as to determine the information-processing mechanisms that generate it. Until recently, psychologists did not have access to the tools for examining brain activity, and so had to rely on indirect measures. One was to measure the time it took for a subject to respond to a particular stimulus (known as *reaction time* or RT); another was to note the error patterns that could be induced by manipulating the conditions under which the stimulus was presented, from which researchers could hypothesize about what operations the brain must be performing. With these tools cognitive psychologists succeeded in developing detailed and well-supported models of the operations occurring as people carry out a variety of cognitive tasks including categorization, problem solving, planning, recognition, and recall. In these endeavors cognitive psychologists often collaborated with researchers in other professions, especially linguistics and computer science. The metaphor of the mind as an information-processing system united researchers from these disciplines in the 1950s into a common enterprise which has come to be called *cognitive science* (Bechtel et al., 1998). In the late 1970s these efforts became institutionalized with the creation of the journal *Cognitive Science* and the establishment of the Cognitive Science Society.

Although both neuroscience and cognitive science were robust interdisciplinary enterprises through the 1960s, 1970s, and 1980s, until the late 1980s there was little intellectual interaction between them. Investigations in the two fields were pursued independently from each other and there were prominent philosophical arguments on behalf of maintaining such autonomy. Hilary Putnam (1967), for example, argued that mental states were multiply realizable – they could be realized by different neural processes in different species, by different patterns of electrical activity in different computers, and by potentially very different kinds of processes in whatever extra-terrestrial life forms might exist. Similarly, Jerry Fodor (1975), argued that the taxonomies of cognitive science and neuroscience would cross-cut each other, spelling failure for any reductionist program. Fodor used as an example of such cross-cutting the way chemistry and economics may cross-cut each other. Different materials can constitute units of money (e.g. a silver dollar and a paper dollar bill) but even very similar objects made out of the same material (e.g. paper) would not count as genuine (but instead counterfeit) money. What makes a chunk of matter a genuine monetary unit and not a mere counterfeit is the role it plays in an economy of minters, bankers, and spenders. Analogously, according to Fodor, what makes something a

psychological state is its role in an economy of psychological states, not its intrinsic material (in this case, neural) properties. Consequently, laws in psychology would be independent of any laws characterizing brain processes.

Towards the end of the 1980s, however, spurred in part by exciting new results stemming from the analysis of various forms of brain deficits and the development of PET imaging, neuroscientists and cognitive scientists began to collaborate and integrate their methodologies in a sustained examination of how brain processes underlie cognitive processes. Psychologist George Miller and neurobiologist Michael Gazzaniga coined the term *cognitive neuroscience* to designate the collaborative inquiry that integrates the behavioral tools of the psychologist with the techniques for revealing brain function to determine how the brain carries out the information processing that generates behavior. Today, cognitive neuroscientists routinely study both psychological processes and neural activity, and since the 1990s cognitive neuroscience has taken off as one of the fastest-developing and most exciting areas of scientific study.

2 Why a Philosophy of Neuroscience?

Two hundred years ago, the world of scientific inquiry and academic scholarship had not yet been divided into specialized disciplines. Law, medicine, theology, literature, and history each had separate faculties, but most of the other scholarly pursuits were the province of philosophy. Those individuals now recognized as major figures in the development of philosophy, such as Descartes, Locke, and Kant, directed many of their inquiries at the natural sciences (e.g. by providing epistemological foundations for the new sciences) and drew upon the results of those sciences. Gradually, however, the various natural and social sciences developed their own techniques, modes of inquiry, and bodies of knowledge, and split off from philosophy into separate disciplines. Psychology, for example, is one of the most recent defectors from the philosophic fold, and was established as a distinct discipline by the end of the nineteenth century, largely through the efforts of the philosopher William James.

As a result of the diminished scientific content of the field, philosophy became identified primarily with inquiries into values (ethics) and attempts to address foundational and general questions about ways of knowing (epistemology) and conceptions of reality (metaphysics). Thus, epistemology became preoccupied with whether *justified true belief* sufficed for knowledge, and metaphysics addressed such questions as whether events or objects and properties are the basic constituents of reality. In the hands of some of its practitioners, philosophy became *purified*, relying only on what it took to be its own tools, such as logic, conceptual analysis, or analysis of ordinary language, to address its own specialized questions.

Not all philosophers accepted the divorce of philosophy from other disciplines. They attempted to maintain philosophy's links to the inquiries that separated from it while nonetheless addressing foundational epistemic, metaphysical, or value

questions. These philosophers have tried to relate their investigations to the ongoing inquiries in other fields, often focusing their philosophical analyses on foundational issues that arise within these fields. Thus, one finds subspecialties in philosophy for philosophy of art, philosophy of economics, philosophy of physics, and philosophy of biology. In particular, the emergence of psychology as an experimental discipline, and more recently of cognitive science has resulted in the increased popularity of philosophy of psychology and philosophy of cognitive science as focal areas within philosophy. Philosophy of neuroscience is a natural continuation of these efforts, comprising an inquiry into foundational questions (especially epistemic and metaphysical ones) that apply to neuroscience (a first step in developing this inquiry was Churchland, 1986).

Philosophy of neuroscience, like philosophy of psychology and philosophy of cognitive science, however, represents more than an attempt to address foundational issues in neuroscience. Insofar as psychology, cognitive science, and neuroscience all address the cognitive and intellectual capacities of humans (and other intelligent animals), their results can inform philosophical thinking about epistemology and metaphysics (especially metaphysical questions about human beings such as the relation of mind to body and the conditions for personal identity). Philosophers who think that results in the sciences themselves may provide material addressing philosophical questions often refer to themselves as *naturalized philosophers* (Callebaut, 1993). Naturalized philosophy involves a dialogue with the sciences, not just an analysis of the science. More generally, the *naturalized* approach to understanding the mind and brain involves seeing them as part of the natural world (rather than as miraculous or supernatural anomalies) and recognizing the biological, evolutionary, and environmental pressures which have helped to shape them. The approach to philosophy of neuroscience represented in this volume is naturalistic. The contributions represent either work in the neurosciences which is especially philosophically relevant or work by philosophers drawing upon or analyzing the scientific research. In part because the areas of neuroscience where discoveries and theories are of most consequence for philosophical issues have been primarily those that focus on neural systems and their relation to cognitive processes, rather than more basic processes such as the chemical events involved in neural transmission (but for an example of philosophy directed toward lower-level neuroscience, see Machamer et al., 2000), most of the focus in this volume will be on developments in systems and cognitive neuroscience.

The philosophical issues concerning neuroscience that are addressed in this book are characteristic of those that arise in philosophy of science and philosophy of mind. To appreciate these issues, some understanding of both areas is helpful. Although we cannot offer a detailed introduction to either in a short chapter (for such introductions, see Bechtel, 1988a, 1988b), the next two sections do provide a synopsis of the central issues in philosophy of science and philosophy of mind. We then focus briefly on four aspects of neuroscience that are especially interesting from a philosophical perspective and which will be examined further in other chapters.

3 Philosophy of Science

One of the main objectives of science is to provide explanation; accordingly, a major goal of philosophy of science is to specify what constitutes an explanation. We briefly review several of the most influential approaches to explanation that have been advanced in recent philosophy of science and point to how these approaches would apply to research in the neurosciences.

One of the most common views of explanation, which traces back to Aristotle and was developed in great detail earlier in this century (Hempel, 1965, 1966), holds that explanation of a phenomenon requires the logical deduction of the occurrence of the phenomenon from laws. In this approach, known as the *covering law model* or the *deductive-nomological* (D-N) model of explanation, laws specify relations between events. Laws are taken to specify general relations (as in Newton's law that force equals mass times acceleration ($f = ma$); to apply these general relations to particular events, one must specify conditions holding at a previous time, which are usually called *initial conditions*. Recognizing that multiple laws and initial conditions may be involved in a given explanation, such explanations can then be represented in the following canonical form (where L designates a law, C an initial condition, and E the event to be explained):

$L1, L2, L3, \ldots$
$C1, C2, C3, \ldots$
$\therefore$ Therefore E

Advocates of the D-N perspective generally assumed that the Cs and Es were sentences whose truth or falsity could be determined directly through observation. These *observation* sentences, accordingly, provided a grounding for the meaning of terms figuring in the laws. In addition to providing a basis of meaning, these observation sentences also provided the empirical support for the laws. In particular, just as one could derive a statement about an event already known to have happened so as to explain it, one could derive a statement about an event not yet known. In this way, the framework allowed for predictions, and the success of these predictions provided a basis for accepting or rejecting proposed laws. The D-N model was extremely influential in some areas of psychology earlier in the twentieth century. Many behaviorists, for example, sought to discover general laws of learning to characterize how various kinds of experiences (e.g. reinforcement) would change the behavior of organisms.

Recognizing that one might want to explain why laws held, the proponents also generalized this framework, allowing for the derivation of one or more laws from other laws. These other laws might be more general ones from which, under specific boundary conditions, the first set of laws might be derived. (Thus, the boundary conditions replace the initial conditions in the above formalism.) Proponents also suggested that this approach might be extended to relations between laws in one

science and those of a more basic science by providing bridge laws relating the vocabularies of the two sciences, giving rise to the following schema:

Laws of the lower-level science
Bridge laws
Boundary conditions
$\therefore$ Laws of the higher-level science

These derivations are known as *reductions*; they figure prominently in discussions about the relation between psychology and neuroscience in which some theorists propose that the laws of psychology ought to reduce to those of neuroscience (see the papers in Part VI of this volume).

In contexts of explanation, one already knows that the E events have occurred and one is trying to explain why they occurred. But the same formalism provided by the D-N model can be employed in cases where the E events are not yet known to have occurred; the formalism then provides for predictions. This is an extremely important aspect of the D-N framework. Finding a law-like statement under which one could subsume an event known to have occurred is extremely easy, but one has no check on whether the purported laws are true. By making predictions which turn out to be true, the logical positivists thought we could justify laws.

For this claim, however, they were criticized by Karl Popper (1935/1959), who noted that such arguments had the invalid form of affirming the consequent:

If L were true, then prediction P would be true
P is true
$\therefore$ L is true

This formalism is invalid since it is possible for both premises to be true, but the conclusion false. (To see this is so, consider the following case: if Lincoln were beheaded, then Lincoln would be dead. Lincoln is dead. Both of these statements are true. But the conclusion that Lincoln was beheaded is false.) Because it is invalid, the truth of the premises does not *guarantee* the truth of the conclusion. However, neither is this particular formalism completely without value, since a number of confirming instances do lend inductive (though inconclusive) support to the initial law or hypothesis. Popper argued that the only way evidence could bear with certainty on laws was through the use of *modus tollens* arguments in which failed predictions could be used to falsify a purported law:

If L were true, then prediction P would be true:
P is false
$\therefore$ L is false

Accordingly, Popper emphasized that the method of science was a method of conjectures and refutations in which one proposed explanatory laws and then sought

evidence showing that the hypothesized law was false. If a proposed law resisted all attempts at falsification, Popper would speak of it as *corroborated*, but not as shown true or confirmed, *recognizing that future evidence could always reveal it to be false.* Later, Hempel (1966) pointed out that even the logic of falsification is problematic, since the law or hypothesis that is ostensibly falsified may itself be a complex conjunction of several auxiliary hypotheses. In this case, a falsifying instance may be the result of a single, false auxiliary hypothesis, which, of course, will give the logical result of falsifying the larger hypothesis.

The D-N model of explanation and accompanying account of reduction coheres well with the general textbook account of the scientific method wherein a scientist is presented as first observing a range of phenomena, hypothesizing a law, testing it by deriving new predictions, and revising the law if the predictions are not borne out. Starting in the late 1950s, however, a number of philosophers and historians of science objected that this picture does not describe the usual practice of science. Thomas Kuhn (1962/1970), for example, argued that in the normal practice of science, researchers are not so much testing theoretical ideas as trying to make them fit nature. Much of the ongoing work of science involves developing and modifying experimental protocols to develop evidence that more and more phenomena fit already-accepted theoretical ideas, which he termed *paradigms*. Rather than testing whether or not $f = ma$ applied to a new range of phenomena, a scientist would be trying to devise ways of showing that it did apply. Only when these normal practices of science began to encounter repeated failures, would scientists explore alternative paradigms. Once a seemingly adequate alternative was found, they would abandon the pursuit of "normal science" and try to extend the range of application of a new, revolutionary framework. Although Kuhn's ideas of how science develops through normal science and revolutionary changes of paradigms have been adopted by many scientists and historians of science to characterize the development of science, they have also proven extremely controversial.

While the logical positivists were themselves very interested in the science of their time, their account was grounded primarily in logic, not in the details of scientific practice. (A consequence of this is that they viewed it as a normative model characterizing any possible science.) Kuhn's work drew philosophers' attention (as well as that of historians and sociologists of science) to the specific details of the process of scientific research. One consequence of this has been the recognition that there may be fundamental differences between scientific disciplines. Philosophers focusing on biology, for example, found that there are few laws associated with biology and that laws do not play a central role in biological explanations. While this might be evidence that biology is not a real science, another interpretation is that another explanatory framework is at play in biology. In particular, biological explanation typically makes references to goals, purposes, and functions. For example, an evolutionary account of an organism's features explains them in terms of what they are "for," in the sense of how they contributed to its predecessors' reproductive success. This kind of explanation is inherently *teleological*, and unlike explanations in physics (atoms are not "for" anything, they just are). Though the legitimacy of teleological

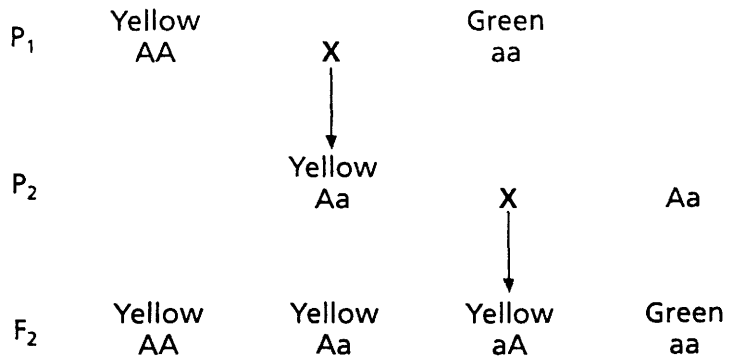

Figure 1.1 A typical diagrammatic representation of the recombination of Mendelian factors in inheritance. In the P_1 generation, a pure-breeding yellow pea is crossed with a pure-breeding green pea. The offspring in the F_1 generation will all be yellow hybrids. If two of them mate, the offspring in the F_2 generation will consist of one pure-breeding yellow pea, two hybrids, and one pure-breeding green pea.

explanation has been frequently challenged, most philosophers of science now accept it (see Mundale and Bechtel, 1996, for an example of how a teleological approach can serve to integrate neuroscience, psychology, and biology).

To further illustrate the importance of functional analysis in biological explanation, note the frequent use of figures and diagrams in biological texts, which often bear an uncanny resemblance to the flowcharts devised by computer scientists. Consider, for example, the familiar Mendelian diagrams of heredity in which the distribution of traits in successive generations are explained in terms of recombination of Mendelian factors (see figure 1.1). What these diagrams do is identify the operations that are being performed in a biological system as it carries out a given task. As such, they propose a *functional decomposition* of the system. At this stage, one does not have to specify what the components are that carry out the different tasks. While we now construe the units postulated as genes and link them with bits of DNA, Mendel merely referred to them as factors and offered no proposals as to how they were realized in the organism. But of course if the account is correct, then there ought to be components of the system that perform these tasks, and a major part of providing evidence that the proposed functional decomposition is correct is the *localization* of the functions in different components of the system.

Diagrams decomposing a system into its functional components (and generally identifying the physical components that perform the functions) provide an account of the mechanism operative in the system. Accordingly, to contrast this account with the D-N model, we will speak of such explanations as *mechanistic explanations*. The following are some of the major differences between mechanistic explanations and D-N explanations. First, as we note by the discussion of diagrams above, the explanations are not necessarily framed linguistically. While we can use language to

present such explanations, or use words in diagrams to identify what the features of the diagram represent, what is crucial to such explanations is the decomposition of the actual system into component functions and component parts. It is by identifying the parts of the system, what they do, and how they are organized to work together, that one explains how a mechanical system performs its operations. Second, as a consequence of the first point, logical deduction is not the *glue* that holds an explanation together. Rather, a diagram portrays a relation between operations (or components), and it is by *envisaging* what will happen when the component functions are performed in the manner portrayed that we appreciate the connection between the explanandum and the explanans. Third, laws do not have a central place in mechanistic explanations. Rather, it is the details of the particular organization of functions (or parts) that do the explanatory work. (This is not to deny any role for laws, or for the D-N framework. Sometimes one does appeal to laws to specify how a component will behave. But what is critical to a mechanistic explanation is the putting together of component functions.)

Here we cannot develop the mechanistic alternative to the D-N approach in detail (but see Bechtel and Richardson, 1993). However, we will see numerous examples through this volume of attempts to develop mechanistic explanations. The neuroscientific study of language, for example, has sought to identify the contributions of various brain areas, such as those identified by Broca (chapter 5, this volume) and Wernicke (chapter 6, this volume), to language processing. Likewise, the neuroscientific study of vision (Part III) has sought to identify the contributions of different brain areas to visual processing. But we will conclude this section by noting one difference in emphasis in developing mechanistic explanations in neuroscience that distinguishes it from some other efforts to develop mechanistic explanations. As we noted above, Mendel carried out his decomposition of the mechanism of heredity without knowledge of the physical components that were involved. Likewise, biochemists have often worked out models of chemical reactions underlying vital phenomena without having discovered the responsible enzymes. And psychologists and other cognitive scientists often developed functional decompositions of cognitive tasks without knowing the brain mechanism involved. But neuroscience inquiries often begin with information relating one or more brain areas with a given cognitive performance. The challenge is often then to develop a functional decomposition that identifies the particular functions performed by the different brain areas, thus insuring, in these cases, that information about implementation figures in the explanation of the higher function to be understood. This difference in approach may make neuroscience a useful place for discovering an important variation in the way scientists develop mechanistic explanations.

4 Philosophy of Mind

Philosophers since antiquity have been enticed by the distinctive character of the mental processes of which we are aware – our thoughts and feelings; our reason-

ing processes; and our affects and emotions. As they present themselves to our phenomenal consciousness, these events and states seem very different from the physical events and states in the world. When we are aware of these processes within us, we are not (at least, not obviously) aware of any physical processes. It seems as if we could have the thoughts and feelings we do even if we lacked a physical body. Consequently, the fundamental question in philosophy of mind has been to explain the relation mental states bear to physical states. This is known as the *mind–body problem*.

One venerable position on the mind–body problem is *dualism*, the view that minds are indeed distinct from physical bodies. Plato advanced one form of dualism as a result of his attempt to understand knowledge. He construed knowledge as involving contact between a mind and what he termed *Forms* or *Ideas*: eternal, non-physical entities which provided the patterns of which all physical entities are imperfect instantiations. Accordingly, in various myths he describes the mind as engaging in direct interaction with the Forms prior to incorporation within a physical body. The experience of being inserted in a body results in a profound loss of memory, and the epistemic challenge of life is to regain the pure knowledge of the Forms unimpeded by physical bodies.

In contrast to Plato, Aristotle rejected the idea of a separate, non-physical realm of Forms; he retained the notion of Forms, but construed them as realized in physical objects. They were what defined an object as of a particular kind. But they were not identical to the matter that comprised the object, and knowledge for him consisted of internalizing the Form of objects dissociated from the matter they possessed when they were realized in physical objects. Thus, even though Aristotle identified a central role for sense perception in learning about Forms, cognitive activity for him entailed a dissociation from the physical domain and he entertained the idea of pure intelligence as dissociated from anything physical.

Modern philosophy of mind originates with the seventeenth-century philosopher René Descartes, who maintained a division between mind and body as great as Plato had proposed. He identified mind as a thinking, unextended substance (occupying no space), and the physical body as non-thinking extended substance. While Descartes attributed many mental activities to non-physical minds, he did allow the brain a role in more basic activities that we now construe as cognitive, including perception and memory. For Descartes, thinking proper was reasoning, which is manifest more clearly in the use of language. Descartes was fascinated with the abilities of complex mechanisms to produce patterns of behavior much like those produced by non-human animals, but linguistic processes, he thought, exhibited a creativity that could not be produced by any machine (or by any non-human animals, which he viewed as mere machines). The creative use of language, exemplified by the ability to construct a sentence never uttered before, required, for him, the non-material mind. One of the challenges for Descartes and others who have maintained that the mind is totally different from physical objects was to explain how brains and minds could interact. The interaction could not be like ordinary physical interactions, since minds were not even located in space. His own proposal was that by effecting small

perturbations in the location of the pineal gland, the mind could alter the course of the "animal spirits" (fine fluids) flowing through the nerves and thus affect behavior. Even though such a proposal might minimize the physical work that the mind was required to do, the challenge remains to explain how the mind might perform any physical work (and be affected itself by physical work). Such difficulties have made dualism a rather unattractive framework for thinking about mind and brain, although even in neuroscience there have been some prominent dualists such as Sir John Eccles. The attraction of dualism is that it seems to enable the mind to be a creative agent outside the ordinary causal nexus and perhaps account for various religious convictions about the spiritual element of humans.

Philosophers of mind have advanced a variety of non-dualistic accounts. In the first half of the twentieth century theorists such as Gilbert Ryle (1949) attempted to diffuse the mind–body problem by arguing that separating mind from body involved a category mistake (comparable to the mistake involved in supposing that, after one has met each of the players on an athletic team, there is still some sense in which one has yet to meet the team). Mental states, Ryle proposed, are exhibited in the behavior of organisms such as humans. Thus, a belief might consist in the propensities to behave in ways we associate with such a belief. Because of its attempt to link mental states with behaviors, this view came to be known as *philosophical behaviorism*. While overcoming the need to account for mind–body interactions, behaviorism encountered its own problems, such as specifying the set of behaviors that were to be identified with particular mental states. Another worry was how it could account for affective states, such as pains or emotions, of which we have direct phenomenal awareness.

To overcome the problem posed by affective states, philosophers such as U. T. Place (1956) and J. C. Smart (1959) proposed in the 1950s that mental states were identical with brain states. For example, feeling pain consisted in being in a particular kind of brain state. Generalizing beyond sensations, what came to be known as the *identity theory* held that mental states in general were identical with brain states. A common objection to the identity theory was that we might characterize our mental states without knowing anything about the underlying brain states. But identity theorists argued that this failed to undercut the identity theory for the same reason that the FBI's initial ignorance that Kaczynski was the Unabomber failed to undercut their ability to characterize the Unabomber.

An objection that many philosophers took to be far more telling against the identity theory was the claim (discussed above) by Putnam (1967) and Fodor (1974) that the same mental states might be realized in very different kinds of systems, including alien life forms and computer systems (science fiction has long contemplated encounters with aliens and computers who have beliefs, desires, etc.). If the same the mental state could be realized in different ways, they contended, then the mental state could not be *identical* with the brain state. Rather, they proposed that the relation of mental states to brain states was comparable to that of software to hardware: a single piece of software characterizes processes that can be performed by many different kinds of hardware. Their view, which came to be known as *functionalism*, held

that mental states are defined by their relation to each other and to sensory inputs and motor outputs. As we noted above, functionalism has been taken as supporting the autonomy of cognitive inquiry from neuroscience.

Although the multiple realizability claim has been widely accepted on faith, its empirical warrant has been questioned (Bechtel and Mundale, 1999), especially with respect to terrestrial animals. The main thrust of the objection is that if we attend to the practice of neuroscientists, we discover that they use a coarse-grained criterion for identifying brain states and treat brain states in different species as the same (despite their more fine-grained differences) just as cognitivists use a coarse-grained criterion for identifying psychological states across species. When grain sizes are matched, researchers are able to advance identity claims between psychological and brain states. Moreover, even though it was a source of inspiration for functionalism, the main functionalist claim that mental states are defined in terms of their functional relations with other mental states can be maintained independently of the claim of multiple realizability. If one does dissociate the claims, then it is possible to adopt both functionalism and the identity theory, holding that mental processes are defined by their interactions, but allow that they are identical with neural processes. Such a stance on the mind–body problem seems most congenial to recent work in cognitive neuroscience and to the account of mechanistic explanation in terms of decomposition and localization we presented in the previous section. The emphasis on functional decomposition adopts functionalism's emphasis on characterizing mental processes in terms of their relations to other mental processes, while the concern for localization adopts the identity theory's claim that mental states are brain states.

5 Special Philosophical Issues for Understanding Neuroscience

The previous two sections provide an overview of some of the main philosophical issues that frame philosophers' consideration of the neurosciences, especially systems and cognitive neuroscience. But there are a number of more specific questions, which we briefly introduce here.

The indirectness of studies of mind and brain

One of the epistemic challenges confronting studies of both mind and brain is the indirectness of the inquiry. The concern with the inaccessible character of mental processes was one of the factors leading the behaviorist B. F. Skinner to attempt to explain behavior totally in terms of observable stimuli and responses. We do seem to be aware of our mental processes (e.g. we know the sequence of our internal thoughts), but each one of us is limited to such awareness of our own thoughts. Moreover, we have no awareness of many of the mental processes psychologists are interested in (e.g. *how* we remember things or recognize objects we see as opposed

to *what* we remember or see), and so any knowledge of these processes must be arrived at indirectly. Accordingly, over the last 150 years psychologists have developed a variety of indirect measures of the processes occurring in us. One measure is the patterns of errors we make on a variety of cognitive tasks (e.g. failing to remember some of the words on a list of words we are asked to remember, or falsely remembering that a word was on the list which was not). Perhaps the most powerful indirect measure psychologists have employed has been reaction time – the time it takes for subjects to perform particular tasks.

Compared to the indirect strategies on which psychologists must rely, it might seem that studying brains is rather direct. However, brains present their own set of problems. Merely looking at a brain is not informative; one must determine what processes occurring in brains are related to cognition. The brains we have been most preoccupied with are human brains, but ethical considerations prevent us from carrying out many invasive studies. Accordingly, until very recently the main source of information about human brains was derived from deficits manifested in patients with various forms of brain damage. Two problems with this source of evidence are that naturally occurring deficits are often quite diffuse in nature and there are challenges in inferring normal function from damaged systems. Brain imaging techniques such as PET and fMRI do provide a window into processes occurring in human brains, but these measures are themselves indirect. Thus, although difficult ethical issues are involved, much of what we know about brains stems from studies of other species in which more invasive approaches are used. From these studies researchers have been able to examine details of neuroanatomy, conduct electrophysiological studies recording from individual neurons, and surgically induce lesions to determine the effects of removing specific brain parts. But even these methods of studying the brain are indirect. For example, the details of neuroanatomy require the use of stains whose own mechanism of operation is often poorly understood.

Thus, rather than relying on some direct forms of observation for their data, psychologists and neuroscientists must rely on techniques and instruments, which may themselves generate artifacts and mislead scientific inquiry. On this topic, chapter 4 below examines the epistemic challenges faced, especially in neuroscience inquiries. While these challenges are not unique to studies of mind and brain (they arise in many other biological disciplines such as biochemistry and cell biology, as well as in the more basic sciences of physics and chemistry), they are not often attended to. They raise, however, significant questions for philosophers of science.

Relations between psychological and neuroscientific inquiries

The logical positivists put forward the model of theory reduction introduced above as a framework for relating different sciences, and this still guides much of the discussion about the relation between psychology and neuroscience. In chapter 22 below, for example, Paul and Patricia Churchland argue for the utility of the

reduction model in developing the relationship between psychology and neuroscience. They acknowledge, however, that not all aspects of psychology will be successfully reduced to neuroscience. For those aspects of psychology that resist reduction they propose elimination in much the same manner as theories of phlogiston chemistry were eliminated at the beginning of the nineteenth century. The Churchlands' claim that some areas of psychology, especially those invoking folk concepts such as belief and desire, should be eliminated has become the focus of much philosophical controversy. McCauley, for example, in chapter 23, this volume, argues that reductions usually emerge when a theory in one discipline is replaced by an improved theory in its own discipline, not as a result of theories in other disciplines.

There are other serious issues raised by the use of the reduction model as a way of relating disciplines. Reduction, as it is understood in most philosophical accounts, involves deriving one *theory* from another, where theories are construed as sets of laws. Within this framework the focal questions have been whether or not psychological theories can be derived from neuroscientific ones. If so, psychological theories seem to lose their autonomy. Accordingly, those arguing for the special status of psychology or other higher-level sciences have argued that such derivations are not possible. (It is, of course, precisely this failure of derivation that the Churchlands cite as the basis for elimination.) However, as we have argued above, most neuroscience explanations do not take the form of D-N explanations in which phenomena are derived from laws, but rather are models of mechanisms. This casts a different light on the issue of reduction. Models of mechanisms are inherently reductionist: each proposed mechanism is designed to show how a phenomenon ascribed to a system is due to its constituent parts and their interaction. On the other hand, reduction no longer threatens the autonomy of the higher-level science: the higher level characterizes the interaction of processes, the lower level accounts for the performance of individual processes. For example, the higher level may account for language processing in terms of the interactive performance of several functions, while the lower level explains how a particular brain part performs one of those functions. Both employ decomposition and localization to offer explanations, but each is explaining a different phenomenon in terms of a system located at one level in the natural hierarchy, its components and their functions, and the organization of the several components into a functioning system.

The perspective presented in the previous paragraph is one that incorporates both reduction and a form of autonomy of higher levels (they are concerned with the integration of components, something not addressed at the lower level) and one that provides a framework for understanding much of the research in contemporary c ognitive neuroscience which tries to link explanatory frameworks in psychology with information about neural mechanisms. But there are times when both disciplines are focused on essentially the same phenomena and are working at the same level of organization (the level of integrated neural systems). The reason this happens is that disciplines are not distinguished just by the particular levels of organization they consider. As Abrahamsen (1987) argues, they are also

differentiated by the manner in which they approach the phenomenon. Specifically, the behavioral sciences, including psychology, focus on the mental and behavioral aspects of the functioning of organisms, whereas the biological sciences, including most of the neurosciences, focus on the organic features of the physical world. Each discipline has developed special tools for conducting its inquiry, tools such as reaction time measures for diagnosing cognitive processes through their behavioral products, and tools such as electrophysiological recording for detecting the physiological processes occurring in organisms. Thus, even when the phenomena being examined overlap, the approach of practitioners from different disciplines is still different. But problems often require integrating the approach of different disciplines. The tools of cognitive neuroscience, such as neuroimaging and single-cell recording, are examples of this, since they require both a focus on the physiological processes and a focus on the behavioral activities these processes are subserving. In instances such as this, collaboration between disciplines does not involve reduction at all, but the integration of perspectives and experimental skills.

Modularity

Decomposition and localization inherently involve fractionating a system into components. In cognitive and neuroscience inquiries, these components are often referred to as *modules*; the most prominent example of a proposed module in the cognitive domain is that of a module for language. A critical issue is just what is intended in segregating a module. Is the module assumed to be totally responsible for the process? If so, what is the nature of its inputs and outputs? And how does it come to acquire such a dedicated capacity? And if there is no such dedicated capacity, what is the significance of assigning the process to a particular brain region?

In a 1983 book *Modularity of Mind*, the philosopher Jerry Fodor advanced a strong statement of what the commitments of such an explanation were. Fodor contended that the following properties were conjointly satisfied by what he termed a *module*: (1) domain specificity, (2) mandatory operation, (3) limited output to central processing, (4) rapidity, (5) information encapsulation, (6) shallow outputs, (7) fixed neural architecture, (8) characteristic and specific breakdown patterns, and (9) characteristic pace and sequencing of development. Of these, Fodor has placed the greatest emphasis on information encapsulation, which is the claim that processing within modules only has access to the limited information represented within the module, not to information stored elsewhere in the system. For Fodor, it is the fact that modules rely only on encapsulated information that allows them to be extremely fast in their processing, but limits them to specific domains of information, reduces their flexibility, and results in their outputs being shallow.

An important feature of Fodor's account is that he does not treat the whole cognitive system as modular; rather, he distinguishes central cognition from modules for sensory systems and language. Central cognition performs the general reasoning of which humans are capable. Fodor characterizes such reasoning as *isotropic* in that

any information a person knows can be invoked in reasoning about any subject and *Quinean* in that one's judgment about any proposition may depend upon its relation to all other propositions one believes. The modules provide input into this central system, but are not affected by its isotropy and Quinean character. One epistemic benefit of such an arrangement, according to Fodor, is that perception can provide an objective account of the world one is sensing, uncontaminated by one's beliefs and feelings; Fodor hopes thereby to avoid the epistemic relativism Kuhn and others have proposed.

While Fodor's account of modules has been very influential, few theorists who appeal to modules have actually adopted all of the characteristics Fodor associates with them. Neuroscientists emphasize that a characteristic feature of brain organization is backward projection. Areas in the temporal cortex involved in higher visual processing send backward projections to primary visual areas in the occipital cortex, and they in turn send projections back to the lateral geniculate nucleus of the thalamus and ultimately the retina. Although the function of these backward projections is not fully understood, they appear to allow downstream processing to modulate processing earlier in the system. If so, the earlier processing is not encapsulated. Basing arguments on behavioral data rather than neural data, Appelbaum (1998) argues that evidence from speech perception shows the effects of higher-level processing (e.g. lexical processing) on lower-level processing (e.g. phonetic processing) and that, as a result of the fact that these influences vary with context, Fodor's attempts to answer this objection by letting some apparently higher-level processing into the speech perception module cannot work.

What is left of modularity if one gives up informational encapsulation? At a minimum, modules would cease to be units which operate independently of the rest of the system except for inputs and outputs. But the consequences might be even more dire. Abandoning the requirement of information encapsulation might jeopardize decomposition and localization, especially if the alternative was to assume that the whole cognitive system was one integrated system. But these are not the only options – one can have a differentiated system in which different components specialize in performing particular tasks without encapsulation. Moreover, information flow between components can be limited without being encapsulated so that not all information in the system is available to every component. Information may reach a component, such as early visual processing, only through specific pathways and must be mediated by activity along that pathway in accordance with the functions of the intermediate processing areas. This provides for sufficient modularity for the strategy of decomposition and localization to be successful without the extreme consequences of Fodor-style modularity.

Computational or representational analysis of brain processing

Perhaps the feature that most clearly distinguished the cognitivism of cognitive psychology, cognitive science, and cognitive neuroscience from the behaviorism that dominated psychology and even neuroscience in the first half of the twentieth

century was the information-processing metaphor. The crucial idea was that various states within a system (computer or brain) would represent information about other things (e.g. objects in the world, events in the world, or even states of the system itself) and that these representations could be manipulated in fruitful ways such as those articulated in formal logic. What is crucial about representations in this account (see Newell, 1980), is that they are states within a system that stand in for that which they represent and enable the system that employs them to deal with that which is represented; in the language of Franz Brentano (1874), they exhibited *intentionality*. In our social world, representations take many forms, including pictures and diagrams, but many theorists found language-like representations to be especially suggestive for cognitive modeling. Fodor (1975), for example, defended the claim that cognition requires a *language of thought*. If cognitive representations were language-like, then the computations would consist of syntactical operations specified by formal rules (i.e. rules making reference only to syntactic structure, not the meaning or reference of the representations).

Although this conception of computation and representation was quite popular in artificial intelligence research of the 1980s, the language-like character of the representations made it seem very unpromising for characterizing brain-based cognitive processing (Churchland, 1986; Churchland, 1989). An alternative approach to computational modeling, known as *neural network modeling* or *connectionism*, on the other hand, has been seen as far more promising. In this approach, the computational system is construed as a network of very simple units partially analogous to neurons. Whereas neurons discharge or spike, these units become activated or deactivated and, depending on their activation, excite and inhibit other units to which they are connected. To model cognitive processing, some of these units are designated as inputs and others as outputs; cognitive tasks are supplied to a network by activating some of its input units and allowing activation to spread through the network until the network stabilizes or a pattern is produced on the output units (Bechtel and Abrahamsen, in press; Clark, 1993). Although patterns of activations in networks are very different from language-like representations in traditional artificial intelligence programs, many theorists construe them also as representations (often referring to them as *distributed* representations – see van Gelder, 1990). In particular, researchers often try to analyze the activation patterns on hidden units (units that are neither input nor output units) as constituting intermediate representations which the network employs in the course of trying to perform a cognitive task (Elman, 1991).

Neuroscientists often speak of representations in the brain when, for example, they are able to show that particular neurons fire most actively in response to a specific stimulus (see chapter 18, this volume), and their usage appears to be rather similar to that of neural network modelers. Moreover, when they speak of computation, they tend to focus on changes in neural processes that can be modeled mathematically, not on formal operations (as in traditional artificial intelligence models). But there are increasingly vocal critics of the whole concept of representation. Some object to the rather minimal notion of representation invoked, arguing

that the standing-in-for relation is insufficient to render something a representation; these critics offer proposals as to what more is needed to turn something (including a brain state) into a representation (see chapter 19, this volume). Others question the utility of analyzing neural systems in terms of representations altogether (chapters 20 and 21, this volume). Some of those raising questions are advocates of dynamical systems theory who emphasize the interdependent relationship of elements in the brain and the interactive relations of these with parts of the body and features of the world. They propose that it is often holistic, emergent features of such systems (such as the system itself settling into different attractor states in response to different environmental circumstances) that provide the key to understanding the behavior of these systems, and they advocate the tools of dynamics as the means of developing such explanations (Port and van Gelder, 1995).

As we noted above, it was the idea of information processing in which internal states were construed as representations that characterized the cognitivists' challenge to behaviorism. It is also the appeal to representations and to computational analyses of the processing of such representations in the brain that has helped spawn the collaboration of cognitive scientists and neuroscientists in the enterprise of cognitive neuroscience. If the appeals to representation and computation in analyses of the brain turn out to be viable, then this integration may have a secure foundation. If not, then alternative bases may need to be sought if the integration is to be successful. In any case, the analysis of representations, or any replacement notion, is a key issue in the foundation of neuroscience and cognitive neuroscience.

6 Summary

Our goal in this chapter has been to identify some of the key issues that arise as philosophy confronts the neurosciences. In particular, we have introduced some of the key issues in philosophy of science and philosophy of mind that are pertinent to the neurosciences, and identified four specific philosophical issues of particular relevance to the neurosciences.

References

Abrahamsen, A. A. 1987: Bridging boundaries versus breaking boundaries: Psycholinguistics in perspective. *Synthese*, 72, 355–88.

Appelbaum, I. 1998: Fodor, modularity, and speech perception. *Philosophical Psychology*, 11, 317–30.

Bechtel, W. 1988a: *Philosophy of Mind: An Overview for Cognitive Science*. Hillsdale, NJ: Erlbaum.

Bechtel, W. 1988b: *Philosophy of Science: An Overview for Cognitive Science*. Hillsdale, NJ: Erlbaum.

Bechtel, W., Abrahamsen, A., and Graham, G. 1998: The life of cognitive science. In W. Bechtel and G. Graham (eds), *A Companion to Cognitive Science*, Oxford: Blackwell, 1–104.

Bechtel, W., and Abrahamsen, A. A. in press: *Connectionism and the Mind.* (2nd edn). Oxford: Basil Blackwell.

Bechtel, W., and Mundale, J. 1999: Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of Science*, 66, 175–207.

Bechtel, W., and Richardson, R. C. 1993: *Discovering Complexity: Decomposition and Localization as Scientific Research Strategies.* Princeton, NJ: Princeton University Press.

Brentano, F. 1874: *Psychology from an Empirical Standpoint* (A. C. Pancurello, D. B. Terrell, L. L. McAlister, trans.). New York: Humanities.

Callebaut, W. 1993: *Taking the Naturalistic Turn, or, How Real Philosophy of Science is Done.* Chicago: University of Chicago Press.

Churchland, P. S. 1986: *Neurophilosophy.* Cambridge, MA: MIT Press.

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

Clark, A. 1993: *Associative Engines.* Cambridge, MA: MIT Press.

Elman, J. L. 1991: Finding structure in time. *Cognitive Science*, 14, 179–211.

Feigl, H. 1958/1967: *The "Mental" and the "Physical": The Essay and a Postscript.* Minneapolis: University of Minnesota Press.

Fodor, J. A. 1974: Special sciences (or: the disunity of science as a working hypothesis). *Synthese*, 28, 97–115.

Fodor, J. A. 1975: *The Language of Thought.* New York: Crowell.

Fodor, J. A. 1983: *The Modularity of Mind.* Cambridge, MA: MIT Press/Bradford Books.

Hempel, C. G. 1965: Aspects of scientific explanation. In C. G. Hempel (ed.), *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science.* New York: Macmillan.

Hempel, C. G. 1966: *Philosophy of Natural Science.* Englewood Cliffs, NJ: Prentice-Hall.

Kuhn, T. S. 1962/1970: *The Structure of Scientific Revolutions* (2nd edn). Chicago: University of Chicago Press.

Machamer, P., Darden, L., and Craver, C. 2000: Thinking about mechanisms. *Philosophy of Science*, 67, 1–25.

Mundale, J., and Bechtel, W. 1996: Integrating neuroscience, psychology, and evolutionary biology through a teleological conception of function. *Minds and Machines*, 6, 481–505.

Newell, A. 1980: Physical symbol systems. *Cognitive Science*, 4, 135–83.

Place, U. T. 1956: Is consciousness a brain process? *British Journal of Psychology*, 47, 44–50.

Popper, K. 1935/1959: *The Logic of Discovery.* London: Hutchinson.

Port, R., and van Gelder, T. 1995: *It's About Time.* Cambridge, MA: MIT Press.

Putnam, H. 1967: Psychological predicates. In W. H. Capitan and D. D. Merrill (eds), *Art, Mind and Religion.* Pittsburgh: University of Pittsburgh Press.

Ryle, G. 1949: *The Concept of Mind.* New York: Barnes and Noble.

Smart, J. J. C. 1959: Sensations and brain processes. *Philosophical Review*, 68, 141–56.

van Gelder, T. 1990: What is the "D" in "PDP"? An overview of the concept of distribution. In S. Stich, D. Rumelhart, and W. Ramsey (eds), *Philosophy and Connectionist Theory.* Hillsdale, NJ: Lawrence Erlbaum Associates.