Introduction

Andrew Brook and Pete Mandik

A small movement dedicated to applying neuroscience to traditional philosophical problems and using philosophical methods to illuminate issues in neuroscience began 20–25 years ago and has been gaining momentum ever since. The central thought behind it is that certain basic questions about human cognition, questions that have been studied in many cases for millennia, will be answered only by a philosophically sophisticated grasp of what contemporary neuroscience is teaching us about how the human brain processes information.

The evidence for this proposition is now overwhelming. The philosophical problem of perception has been transformed by new knowledge about the vision systems in the brain. Our understanding of memory has been deepened by knowing that two quite different systems in the brain are involved in short- and long-term memory. Knowing something about how language is implemented in the brain has transformed our understanding of the structure of language, especially the structure of many breakdowns in language. And so on. On the other hand, a great deal is still unclear about the implications of this new knowledge of the brain. Are cognitive functions localized in the brain in the way assumed by most recent work on brain imaging? Does it even make sense to think of cognitive activity being localized in such a way? Does knowing about the areas active in the brain when we are conscious of something hold any promise for helping with long-standing puzzles about the nature and role of consciousness? And so on.

A group of philosophers and neuroscientists dedicated to informing each other's work has grown up. It is a good time to take stock of where this movement now is and what it is accomplishing. The Cognitive

Science Programme at Carleton University and the McDonnell Centre for Philosophy and the Neurosciences at Simon Fraser University organized a conference, the McDonnell/Carleton Conference on Philosophy and the Neurosciences, at Carleton University, Ottawa, Canada, October 17–20, 2002, around this theme. Many of the essays in the current volume are derived from work presented at that conference, though all of them go well beyond what was presented there. The aim of the volume is to achieve a comprehensive 'snapshot' of the current state of the art in the project to relate philosophy and neuroscience.

One of the special features of the authors in this volume is that, with one exception, they all have at least PhD-level training or the equivalent in both neuroscience and philosophy. (The exception is the first editor.) The chapters are clustered around five themes:

- data and theory in neuroscience
- neural representation and computation
- visuomotor transformation
- color vision
- consciousness

History of Research Connecting Philosophy and Neuroscience

Prior to the 1980s, very little philosophical work drew seriously on scientific work concerning the nervous system or vice versa. Descartes speculated in (1649) that the pineal gland constituted the interface between the unextended mind and the extended body and did some anatomy in laboratories (including on live, unanaesthetized animals; in his view, animals do not have the capacity to feel pain), but he is at most a modest exception.

Coming to the 20th century, even when the identity of mind with brain was promoted in the mid-20th century by the identity theorists, also called state materialists, they drew upon very little actual brain science. Instead, the philosophy was speculative, even somewhat fanciful. Some examples: Herbert Feigl (1958/1967) proposed an autocerebroscope whereby people could directly observe their own mental/neural processes. This was science fiction, not science fact or even realistic scientific speculation. Much discussion of identity theory involved the question of the identification of pain with C-fibre firings (U. T. Place 1956 and J. J. C. Smart 1959). But it has been known for a very long time that the neural basis of pain is much more complicated than that (see V. G. Hardcastle 1997 for a recent review).

There were a few exceptions to the general ignorance about neuroscience among philosophers prior to the 1980s. Thomas Nagel (1971) is an example. This paper discusses the implications of experiments with commissurotomy (brain bisection) patients for the unity of consciousness and the person. D. C. Dennett (1978) discusses the question of whether a computer could be built to feel pain according to a thorough and still interesting summary of what was known about pain neurophysiology at the time. Barbara von Eckardt Klein (1975) discussed the identity theory of sensations in terms of then-current work on neural coding by Vernon Mountcastle, Benjamin Libet, and Herbert Jasper. But these exceptions were very much the exception.

The failure of philosophers of the era to draw on actual neuroscientific work concerning psychoneural identities could not be blamed on any lack of relevant work in neuroscience. David Hubel and Torsten Wiesel's (1962) Nobel Prize–winning work on the receptive fields of visual neurons held great promise for the identification of the perception of various visual properties with various neural processes. A decade earlier, Donald Hebb (1949) had tried to explain cognitive phenomena like perception, learning, memory, and emotional disorders in terms of neural mechanisms.

In the 1960s, the term 'neuroscience' emerged as a label for the interdisciplinary study of nervous systems. The Society for Neuroscience was founded in 1970. (It now has 25,000 members.) In the 1970s, the term 'cognitive science' was adopted as the label for interdisciplinary studies of 'cognition' – the mind as a set of functions for processing information. The idea of information processing might not have been much more than a uniting metaphor, but real effort was put into implementing the relevant functions in computational systems (artificial intelligence). Cognitive Science became institutionalized with the creation of the Cognitive Science Society and the journal *Cognitive Science* in the late 1970s. However, it has not grown the way neuroscience has. After 30 years, the Cognitive Science Society has about 1,500 members.

Until the 1980s, there was very little interaction between neuroscience and cognitive science. On the philosophical front, this lack of interaction was principled (if wrong-headed). It was based on a claim, owing to functionalists such as Jerry Fodor (1974) and Hilary Putnam (1967), that since cognition could be multiply realized in many different neural as well as non-neural substrates, nothing essential to cognition could be learned by studying neural (or any other) implementation. It is the cognitive functions that matter, not how they are implemented in this, that, or the other bit of silicon or goopy wetware.

The 1980s witnessed a rebellion against this piece of dogma. Partly this was because of the development of new and much more powerful tools for studying brain activity, fMRI (functional magnetic resonance imaging; the 'f' is usually lowercase for some reason) brain scans in particular. In the sciences, psychologist George Miller and neurobiologist Michael Gazzaniga coined the term 'cognitive neuroscience' for the study of brain implementation of cognitive functioning. Cognitive neuroscience studies cognition in the brain through such techniques as PET (positron emission tomography) and fMRI that allow us to see how behaviour and cognition, as studied by cognitive scientists, are expressed in functions in the brain, as studied by neuroscientists. The idea of relating cognitive processes to neurophysiological processes was not invented in the 1980s, however. For example, in the 1970s, Eric Kandel (1976) proposed explaining simple forms of associative learning in terms of presynaptic mechanisms governing transmitter release. T. V. P. Bliss and T. Lomo (1973) related memory to the cellular mechanisms of long-term potentiation (LTP).

In philosophy, an assault on the functionalist separation of brain and mind was launched with the publication of Patricia (P. S.) Churchland's *Neurophilosophy* in 1986 (a book still in print). Churchland's book has three main aims:

1. to develop an account of intertheoretic reduction as an alternative to the account from logical positivist philosophy of science;
2. to show that consciousness-based objections to psychoneural reduction don't work; and
3. to show that functionalist/multiple realizability objections to psychoneural reduction don't work.

A later neurophilosophical rebellion against multiple realizability was led by W. Bechtel and J. Mundale (1997). Their argument was based on the way in which neuroscientists use psychological criteria in determining what counts as a brain area.

With this sketch of the history of how the philosophy and neuroscience movement emerged, let us now look at particular topic areas in order to lay out some of the relevant history, examine what is going on currently, and connect the area to the contributions in this volume. By and large, the topics of primary interest in the philosophy of neuroscience are those that relate the mind/brain issue to concerns from the philosophy of science and the philosophy of mind. In fact, it is not always easy to distinguish philosophy of mind from philosophy of science in the philosophy and neuroscience movement. For example, the philosophy of mind question 'Are cognitive processes brain processes?' is closely related to the philosophy of science question 'Are psychological theories reducible to neurophysiological theories?' Either way, neurophilosophical interest is mostly concerned with research on the brain that is relevant to the mind (I. Gold and D. Stoljar, 1999, explore the relationship of neuroscience and the cognitive sciences in detail). There are a few exceptions, however. An important philosophical study of areas of neuroscience not directly relevant to cognition is found in P. Machamer et al. (2000), who discuss philosophically individual neurons, how neurons work, and so on.

First we will examine two big background topics: (1) neuroscience and the philosophy of science; and (2) reductionism versus eliminativism in neuroscience and cognitive science. Then we will turn to some of the areas in which philosophy and neuroscience are interacting.

Neuroscience and the Philosophy of Science

Much early philosophy of science held a central place for the notion of law, as in the Deductive-Nomological theory of scientific explanation or the Hypothetico-Deductive theory of scientific theory development or discussions of intertheoretic reduction. While the nomological view of science seems entirely applicable to sciences such as physics, there is a real question as to whether it is appropriate for life sciences such as biology and neuroscience. One challenge is based on the seeming teleological character of biological systems. Mundale and Bechtel (1996) argue that a teleological approach can integrate neuroscience, psychology, and biology.

Another challenge to the hegemony of nomological explanation comes from philosophers of neuroscience, who argue that explanations in terms of laws at the very least need to be supplemented by explanations in terms of mechanisms (Bechtel and R. C. Richardson 1993; Machamer et al. 2000). Here is how their story goes. Nomological explanations, as conceived by the Deductive-Nomological model, involve showing that a description of the target phenomenon is logically deducible from a statement of general law. Advocates of the mechanistic model of explanation claim that adequate explanations of certain target phenomena can be given by describing how the phenomena result from various processes and subprocesses. For example, cellular respiration is explained by appeal to various chemical reactions and the areas in the cell where these reactions take place. Laws are not completely abandoned but they are supplemented (P. Mandik and Bechtel 2002).

A related challenge to logical positivist philosophy of science questions whether scientific theories are best considered as sets of sentences. Paul (P. M.) Churchland (1989), for example, suggests that the vector space model of neural representation should replace the view of representations as sentences (more on vector spaces later in this section). This would completely recast our view of the enterprise of scientific theorizing, hypothesis testing, and explanation. The issue is directly connected to the next one.

Reductionism Versus Eliminativism

There are three general views concerning the relation between the psychological states posited by psychology and the neurophysiological processes studied in the neurosciences:

1. The autonomy thesis: While every psychological state may be (be implemented by, be supervenient on) a brain state, types of psychological states will never be mapped onto types of brain states. Thus, each domain needs to be investigated by distinct means (see Fodor 1974).

Analogy: Every occurrence of red is a shape of some kind, but the color-type, redness, does not map onto any shape-type. Colors can come in all shapes and shapes can be any color (see A. Brook and R. Stainton 2000, chapter 4, for background on the issue under discussion here).

2. Reductionism: Types of psychological states will ultimately be found to be types of neurophysiological states.

The history of science has been in no small part a history of reduction: Chemistry has been shown to be a branch of physics, large parts of biology have been shown to be a branch of chemistry. Reductivists about cognition and psychology generally believe that cognition and psychology, or much of them, will turn out to be a branch of biology.

3. Eliminativism (also called eliminative materialism): Psychological theories are so riddled with error and psychological concepts are so weak when it comes to building a science out of them (for example, phenomena identified using psychological concepts are difficult if not impossible to quantify precisely) that psychological states are best regarded as talking about nothing that actually exists.

Eliminativist arguments are antireductivist in the following way: They argue that there is no way to reduce psychological theories to neural theories or at best no point in doing so.

Philosophers of neuroscience generally fall into either the reductionist or the eliminativist camps. Most are mainly reductionists – most, for example, take the phenomena talked about in the 'cognitive' part of cognitive neuroscience to be both perfectly real and perfectly well described using psychological concepts – but few are dogmatic about the matter. If some psychological concepts turn out to be so confused or vague as to be useless for science, or to carve things up in ways that do not correspond to what neuroscience discovers about what structures and functions in the brain are actually like, then most people in the philosophy and neuroscience movement would cheerfully eliminate rather than try to reduce these concepts. Few are total eliminativists. Even the most radical people in the philosophy and neuroscience movement accept that *some* of the work of cognitive science will turn out to have enduring value.

Some philosophers of neuroscience explicitly advocate a mixture of the two. For instance, Paul and Patricia Churchland seem to hold that 'folk psychology' (our everyday ways of thinking and talking about ourselves as psychological beings) will mostly be eliminated, but many concepts of scientific psychology will be mapped onto, or 'reduced' to, concepts of neuroscience. For example, the Churchlands seem to hold that 'folk concepts' such as belief and desire don't name anything real but that scientific psychological concepts such as representation do (as long as we keep our notion of representation neutral with respect to various theories of what representations are), and that many kinds of representation will ultimately be found to be identical to some particular kind of neural state or process (P. S. Churchland 1986).

We cannot go into the merits of reductivist versus eliminativist claims, but notice that the truth of eliminativism will rest on at least two things:

1. The first is what the current candidates for elimination actually turn out to be like when we understand them better. For example, eliminativists about folk psychology often assume that folk psychology views representations as structured something like sentences and computations over representations as something very similar to logical inference (P. M. Churchland 1981; S. Stich 1983; P. S. Churchland 1986). Now, there are *explicit* theories that representation is like that. Fodor (1975), for example, defends the ideas that all thought is structured in a language – a language of thought.

But it is not clear that any notion of what representations are like is built into the *very folk concept* of representation. The picture of representation and computation held by most neuroscientists is very different from the notion that representations are structured like sentences, as we will see when we get to computation and representation, and so *if* the sententialist idea is built into folk psychology, then folk psychology is probably in trouble. But it is not clear that any such idea is built into folk psychology.

2. The second thing on which the truth of eliminativism will depend is what exactly reduction is like. This is a matter of some controversy (C. Hooker 1981; P. S. Churchland 1986). For example, can reductions be less than smooth, with some bits reduced, some bits eliminated, and still count as reductions? Or what if the theory to be reduced must first undergo some rejigging before it can be reduced? Can we expect theories dealing with units of very different size and complexity (as in representations in cognitive science, neurons in neuroscience) to be reduced to one another at all? And how much revision is tolerable before reduction fails and we have outright elimination and replacement on our hands? J. Bickle (1998) argues for a revisionary account of reduction. R. McCauley (2001) argues that reductions are usually between theories at roughly the same level (intratheoretic), not between theories dealing with radically different basic units (intertheoretic).

These big issues in the philosophy of neuroscience have been hashed and rehashed in the past 25 years. As we have seen, most people in the philosophy and neuroscience movement have arrived at roughly the same position on them. Thus, while they certainly form the background to current work, none of the contributions to this volume takes them up.

On many other topics, we are far from having a settled position. The chapters in this volume focus on these topics. Specifically, they contribute to the issues of

1. localization and modularity
2. role of introspection
3. three specific issues in the area of neural computation and representation:
   - the architecture, syntax, and semantics of neural representation
   - visuomotor transformation
   - color vision

and

4. consciousness.

We have grouped the contributions to the first two topics under the heading 'Data and Theory in Neuroscience'. Otherwise, the book examines these topics in the order just given.

Data and Theory: Localization, Modularity, and Introspection

Localization

A question with a long history in the study of the brain concerns how localized cognitive function is. Early localization theorists included the phrenologists Franz Gall and Johann Spurzheim. Pierre Flourens was a severe contemporary critic of the idea in the early 1800s.

Localizationism reemerged in the study of the linguistic deficits of aphasic patients of J. B. Bouillaud, Ernest Auburtin, Paul Broca, and Carl Wernicke in the mid-1800s. Broca noted a relation between speech production deficits and damage to the left cortical hemisphere, especially in the second and third frontal convolutions. Thus was 'Broca's area' born. It is considered to be a speech production locus in the brain. Less than two decades after Broca's work, Wernicke linked linguistic comprehension deficits with areas in the first and second convolutions in the temporal cortex now called 'Wernicke's area'.

The lesion/deficit method of inferring functional localization raises several questions of its own, especially for functions such as language for which there are no animal models (von Eckardt Klein 1978). Imaging technologies help alleviate some of the problems encountered by lesion/deficit methodology (for instance, the patient doesn't need to die before the data can be collected!). We mentioned two prominent imaging techniques earlier: positron emission tomography, or PET, and functional magnetic resonance imaging, or fMRI. Both have limitations, however. The best spatial resolution they can achieve is around 1 mm. A lot of neurons can reside in a 1 mm by 1 mm space! And there are real limitations on how short a time span they can measure, though these latter limitations vary from area to area and function to function. Especially in fMRI, resolution improves every year, however.

In PET, radionuclides possessing excessive protons are used to label water or sugar molecules that are then injected into the patient's bloodstream. Detectors arranged around the patient's head detect particles emitted in the process of the radioactive decay of the injected nuclides. PET thus allows the identification of areas high in blood flow and glucose utilization, which is believed to be correlated with the level of neural and glial cell activity (a crucial and largely untested, maybe untestable, assumption). PET has been used to obtain evidence of activity in the anterior cingulate cortex correlated with the executive control of attention, for example, and to measure activity in neural areas during linguistic tasks like reading and writing (D. Caplan et al. 1999). For a philosophical treatment of issues concerning PET, see R. Stufflebeam and Bechtel (1997).

fMRI measures the amount of oxygenation or phosphorylation in specific regions of neural tissue. Amounts of cell respiration and cell ATP utilization are taken to indicate the amount of neural activity. fMRI has been used to study the localization of linguistic functions, memory, executive and planning functions, consciousness, memory, and many, many other cognitive functions. Bechtel and Richardson (1993) and Bechtel and Mundale (1997) discuss some of the philosophical issues to do with localization.

In this volume, **Valerie Hardcastle** and **Matthew Stewart** present compelling evidence in 'Localization and the Brain and Other Illusions' that even a system as simple and biologically basic as oculomotor control is the very reverse of localized. To the contrary, it involves contributions from units dispersed widely across the cortex. They also show that a given nucleus can be involved in many different information-processing and control activities. They point out that the brain's plasticity – its capacity to recover function by using new areas when damage to an area affects function – holds the same implication. (They also make the point that these assays into how the brain actually does something undermine the claim that we can study cognitive function without studying the brain.)

Modularity

The question of localization connects to another big question in cognitive neuroscience, namely, modularity. Fodor (1983) advanced a strong modularity thesis concerning cognitive architecture. According to Fodor, a module is defined in terms of the following properties: (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. He then argues that most of the brain's peripheral systems are modular, sometimes multimodular, while the big central system in which the thinking, remembering, and so on is done is emphatically not.