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 twenty to twenty-five 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 is at 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, Oct. 17 - 20, 2002 around this theme. Many of the papers in the current volume are derived from work presented at that conference, though all of them go well beyond what was presented at the conference. 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 Ph.D.-level training or the equivalent in both neuroscience and philosophy. (The exception is the first editor.) The papers in the volume are clustered around five themes,

- data and theory in neuroscience
- neural representation and computation
- visuomotor transformation

- colour vision
- consciousness

History of Research Connecting Philosophy and Neuroscience

Prior to the 1980's, 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 un-extended mind and the extended body and did some anatomy in laboratories (including on live, not anaesthetized 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 twentieth century by the identity theorists, aka central 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. 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 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. Dennett (1978) discusses the question of whether a computer could be built that felt pain in terms of 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 Mountcastle, Libet, and 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 thirty years, the Cognitive Science Society has about 1500 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 1980's 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 lower-case 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 techniques such as PET (positron emission tomography) and fMRI that allow us to see how behaviour and cognition, as studied by cognitive scientists, is 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. Bliss and 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 Bechtel and 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 some particular topic areas, 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 topics 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 (Gold and 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 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 vs eliminativism in neuroscience and cognitive science. Then we will turn to some of the areas in which philosophy and neuroscience are interacting.

N**euroscience 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 Richardson 1993, Machamer and Craver 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 results from various processes and sub-processes. 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 (Mandik and Bechtel 2002).

A related challenge to logical positivist philosophy of science questions whether scientific theories are best considered 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 below). 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.

Reduction Versus Elimination

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 (Fodor 1974)

Analogy: every occurrence of red is a shape of some kind, but the colour-type, redness, does not map onto any shape-type. Colours can come in all shapes and shapes can be any colour (see Brook and Stainton 2000, Ch. 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 (aka 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 anti-reductivist 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 in to 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 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, 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, '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 scientific psychological concepts such as representation do (so 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 vs. eliminativist claims, but notice that the truth of eliminativism will rest on at least two things:

(1) 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 to very similar to logical inference (P. M. Churchland 1981; 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 out *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, 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 (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? Bickle (1998) argues for a revisionary account of reduction. 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 twenty-five 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 papers in this volume focus on these topics. Specifically, they contribute to the issues of:

(3) localization and modularity

(4) role of introspection

(5) three specific issues in the area of neural computation and representation,

the architecture, syntax, and semantics of neural representation

visuomotor transformation,

colour vision

and,

(6) 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 Gall and Spurzheim. Flourens was a severe contemporary critic of the idea (early 1800s).

Localizationism reemerged in the study of the linguistic deficits of aphasic patients of Bouillaud, Auburtin, Broca, and 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 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 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 1mm. A lot of neurons can reside in a 1mm by 1mm 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 blood stream. 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 level of neural and glial cell activity (a crucial and largely untested, maybe untestable, assumption). PET has been used obtain evidence of activity in 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 (Caplan, Carr, Gould, and Martin 1999). For a philosophical treatment of issues concerning PET, see Stufflebeam and Bechtel 1997.

fMRI measures amount of oxygenation or phosphorylation in specific regions of neural tissue. Amounts of cell respiration and cell ATP utilization are taken to indicate 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 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. Fodor then argues that most of the brain's peripheral systems are modular, sometimes multi-modular, while the big central system in which the thinking, remembering, and so on is done, is emphatically not.

Fodor's account can be resisted in two ways. One is to argue that he has an overly restricted notion of what a module has to be like. The other is to argue that, no matter how characterized, there are precious few if any modules in the brain. This is the conclusion that Hardcastle's and Stewart's research supports. Another body of evidence supporting the same conclusion is back projection. For example, temporal cortical areas implicated in high levels of visual processing send back projections to lower level areas in primary visual cortex which in turn send back projections to even lower areas in the lateral geniculate nuclei and ultimately back to the retina. Applebaum (1998) argues for similar phenomena in speech perception: higher-level lexical processing affects lower-level phonetic processing. In fact, neuroscientific research shows that back projections are to be found everywhere. But where there are back projections, there cannot be encapsulated modules.

Introspection

In a variety of ways, the advent of sophisticated imaging of brain activity has created a new reliance on introspections – self-reports about what is going on in oneself consciously. Introspection has been in bad odour as a research tool for over 100 years. Researchers claim that introspective claims are unreliable in the sense that they are not regularly replicated in others. Subjects confabulate (make up stories) about what is going on in themselves when needed to make sense of behaviour. Introspection has access only to a tiny fraction of what is going on in oneself cognitively. And so on. Neuroscience researchers were forced back onto introspection because the only access as cognitive phenomena that they have to the things that they want to study in the activity of the brain is the access the subject him or herself has. Many find it ironic that this most scientific of approaches to human nature has been forced to fall back onto a technique rejected as unscientific 100 years ago!

In this volume, two contributions focus on the role of introspection in neuroscience. and **Evan Thompson, Antoine Lutz** and **Diego Cosmelli** argue in 'Neurophenomenology: An Introduction for Neurophilosophers' that first-person reports about real-time subjective experience made after training and practise can play and should play a vital role in neuroscience, especially the neuroscience of consciousness. They call their approach Neurophenomenology; their paper provides a comprehensive overview of the approach and the context out of which it arose, together with two extremely interesting examples of neurophenomenology actually at work experimentally. .

By way of contrast, in 'Out of the Mouth of Autistics: Subjective Report and Its Role in Cognitive Theorizing', **Victoria McGeer** urges that first-person utterances other than introspective reports can also be an important source of evidence. She studies autistics and their self-reports. Because of the extensive cognitive deficits found in autistic people, many hold that their first-person utterances are suspect. McGeer argues that behind this belief is an assumption that the job of first-person utterances is to describe what is going on in people's heads, If, instead, we treat self-reports as *expressions* of the underlying cognitive and affective system, then we can see that there is a great deal to be learned from them. She then draws two

implications from this analysis. At the philosophical level, she argues than once we abandon the model of first-person utterance as report, there will no longer be any conflict between the use of subjective report and the requirement that evidence be public. At the empirical level, she urges a radical rethinking of autism in terms of developmental connections between profound sensory abnormalities and the emergence of specific higher-order cognitive deficits.

Neural Computation and Representation

As we said, the topic of neural computation and representation is huge. Just the issue of neural representation alone is huge. Contributions to the latter topic can be thought of as falling into three groups, though the boundaries between them are far from crisp.

The neurophilosophical questions concerning computation and representation nearly all assume a definition of computation in terms of representation transformation. Thus, most questions concerning computation and representation are really questions concerning representation. There are three general kinds of question: questions to do with architecture, question to do with syntax, and questions to do with semantics. The question of architecture is the question of how a neural system having syntax and semantics might be structured. The question of syntax is the question of what the formats or permissible formats of the representations in such a system might be and how representations interact with each other based on their form alone. The questions of semantics is the question of how it is that such representations come to represent – how they come to have content, meaning.

Architecture of Neural Representation

In this volume, two papers are devoted to neural architecture, one on the issue in general, one on how time might be represented neurally.

In ‘Moving Beyond Metaphors: Understanding the Mind for What It Is’, **Chris Eliasmith** suggests that past approaches to understanding the mind, including symbolicism, connectionism, and dynamicism, fundamentally rely on metaphors for their underlying theory of mind. He presents a new position which is not metaphorical and synthesizes the strengths of these past approaches. The new view unifies representational and dynamical descriptions of the mind, so he calls it R&D Theory. In R&D Theory, representation is rigorously defined by encoding and decoding relations. The variables identified at higher levels can be considered state variables in control theoretical descriptions of neural dynamics. Given the generality of control theory and representation so defined, this approach is sufficiently powerful to unify descriptions of cognitive systems from the neural to the psychological levels. R&D Theory thus makes up for the absence of a proper dynamical characterization of cognitive systems characteristic of sententialism and shows how, contrary to dynamicist arguments (van Gelder, 1998), one can have both representation and dynamics in cognitive science.

Rick Grush, in ‘The Architecture of Temporal Representation’, focuses on the structures in a neural system that represent time. His target is not objective time, actual persistence, but rather the subjective time of behaviour: the temporal representation that is analogous to egocentric space (in contrast to objective or allocentric space). There are two parts to his theory. The first concerns the neural construction of states that are temporally indexed (see also Grush 2002); mechanisms of emulation can be augmented to maintain a temporally ‘thick’ representation of the body and environment. The second part of the story concerns how these temporally indexed states acquire specifically temporal phenomenal content. Building on Grush (1998, 2001) on spatial representation, Grush urges that these temporally indexed states are imbued with ‘egocentric’ temporal content for an organism to the extent that they cue and guide temporally relevant aspects of sensorimotor skills.

Neural syntax

The standard way of interpreting synaptic events and neural activity patterns as representations is to see them as constituting points and trajectories in vector spaces. The computations that operate on these representations will then be seen as vector transformations (P. M. Churchland 1989). This is thus the view adopted in much neural network modelling (connectionism, parallel distributed processing). The system is construed as having network nodes (neurons) as its basic elements and representations are states of activations in sets of one or more neurons. (Bechtel and Abrahamsen 2002; Clark 1993).

Recently, work in neural modelling has started to become even more fine-grained. This new work does not treat the neuron as the basic computational unit, but instead utilizes compartmental modelling whereby activity in and interactions between patches of the neuron’s membrane may be modelled (Bower and Beeman. 1995). Thus, not only are networks of neurons viewed as performing vector transformations, but so are individual neurons.

Neural syntax consists of the study of the information-processing relationships among neural units, whatever one takes the relevant unit to be. Any worked-out story about the architecture of neural representation will hold implications for neural syntax, for what kind of relationships neural representations will have to other neural representations such that they can be combined and transformed computationally. **Eliasmith** and **Grush** see their stories as holding such implications.

Neural semantics

Cognitive science and cognitive neuroscience are guided by the vision of information-processing systems. A crucial component of this vision is that states of the system carry information about or represent aspects of the external world. (see Newell 1980). Thus, a central role is posited for intentionality, a representation being about something, mirroring Franz Brentano’s (1874) insistence on its importance (he called it “The mark of the mental”, only slightly an exaggeration) a century before.

How do neural states come to have contents? There are two broad answers to this question that have been popular in philosophy. They both appear in the philosophy of neuroscience: the Functional Role approach and the Informational approach. Proponents of functional role semantics propose that the content of a representation, what it is about, is determined by the functional/causal relations it enters into with other representations (Block 1986). For informational approaches, a representation has content, is about something, in virtue of certain kinds of causal interactions with what it represents (Dretske 1981, 1988). In philosophy of neuroscience, Paul Churchland has subscribed to a functional roles semantics at least since 1979. His account is further fleshed out in terms of state-space semantics (P. M. Churchland 1989, 1995). However, certain aspects of Churchland's 1979 account of intentionality also mirror informational approaches.

The neurobiological paradigm for informational semantics is the feature detector, for example, the device in a frog that allows it to detect flies. Lettvin et al. (1959) identified cells in the frog retina that responded maximally to small shapes moving across the visual field. Establishing that something has the function of detecting something is difficult. Mere covariation is often insufficient. Hubel and Wiesel (1962) identified receptive fields of neurons in striate cortex that are sensitive to edges. Did they discover edge detectors? Lehky and Sejnowski (1988) challenge the idea that they had, showing that neurons with similar receptive fields emerge in connectionist models of shape-from-shading networks. (See P. S. Churchland and Sejnowski 1992 for a review.) Akins (1996) offers a different challenge to informational semantics and the feature detection view of sensory function through a careful analysis of thermoperception. She argues that such systems are not representational at all.

As was true of neural syntax, any worked-out story about the architecture of neural representation will hold implications for neural semantics, for the question of how neural representations can come to have content, meaning, be about states of affairs beyond themselves. **Eliasmith** and **Grush** see their stories as holding implications for this issue, too.

In addition to these papers, in 'The Puzzle of Temporal Experience' **Sean Kelly** tackles a very specific kind of representational content, the representation and conscious experience of temporality, and the neuroscience of same. He starts from Kant's famous distinction between a succession of independent representations and a representation of a single, unified, temporally extended object, the distinction between a succession of representations and a representation of succession. He shows how neither Specious Present Theory nor Kantian/Husserlian Retention Theory gives us a satisfying account of our experience of the passage of time, Since these are the only going accounts, how we manage to represent the passage of time is deeply puzzling. He then shows that there is a conceptual/phenomenological distinction to be made between the pure visual experience of motion and the visual experience of an object as moving, and that neuroscientists have tended to conflate the two. In response, he proposes experiments that could confirm a double dissociation between the two and recommends that we look for a certain kind of short-term visual storage that could serve as the basis for the experience of a single, unified, temporally extended, object moving through space. This short-term visual storage should have a certain phenomenology associated with it, but in isolation will be unlike any kind of memory,

including visual iconic memory, that we are familiar with. Since representation of objects as persisting through time is a completely general feature of representation, Kelly's paper identifies something that we have to understand if we are to understand neural representation. H concludes the paper with a revealing overview of the current state of neuroscience on the issue.

Visuomotor Transformation

Grush's paper and, less directly, Eliasmith's paper also connect to the first of the two more specific topics to do with neural representation examine in the book, visuomotor transformation, that is to say, the use of visual information to guide motor control.

In 'Grasping and Perceiving Objects', **Pierre Jacob** starts from the Milner & Goodale (1995) hypothesis that we have two complementary visual systems, vision-for-perception and vision-for-action, based on a double dissociation between two kinds of disorder found in brain-lesioned human patients. visual form agnosia and optic ataxia. Milner and Goodale claim that this functional distinction mirrors the anatomical distinction between the ventral pathway and the dorsal pathway in the visual system of primates. Using psychophysical experiments in normal human beings based on visual illusions, Jacob makes some new discoveries about how visuomotor representation and conscious visual perception relate. Jacob shows that these findings have major implications for theses advanced by others.

The paper by **Pete Mandik,** 'Action-oriented Representation', relates to Jacob's. Focussing on the claim that spatial perception and motor output are interdependent, Mandik asks how best to characterize this interdependence. There are two broad approaches. One favours the positing of mental representations mediating between perception and action, the other opposes the idea. Mandik favours the former proposal, urging that sensorimotor interdependence is best accounted for by a novel theory of representational content whereby the most primitive forms of representation are those that have the function of commanding actions. Other representations, including sensory representations of spatial properties, depend on these motor command representations. His argument draws heavily on both neurophysiology and computer simulations and hold striking implications for both antirepresentationalism and competing representational accounts.

Colour vision

The second of our two more specific topics has to do with neural representation is colour vision.

In 'Chimerical Colours', **Paul Churchland** presents a stunning example of neurophilosophy at work. He shows that by exploiting shifts in experienced colour due to tiredness and habituation, experiences of colour can be brought about where the colours do not exist in nature and, what is even more striking, could not exist in nature according to extremely long-held and well-confirmed colour theory. They are impossible. (Indeed, some of them cannot even be represented by a colour sample, which makes the argument for these cases a bit difficult

to present!) Churchland's paper is also a nice example of neurophenomenology at work. (Whether he would accept this description of his method is, of course, another question.)

Focussing on perceived colour similarity, in 'Opponent processing, linear models, and the veridicality of colour perception' **Zoltán Jakab** argues against views of colour experience that hold that the representational content of a colour experience is exhausted by the colour property in an object for which the experience stands. To the contrary, Jakab argues, colour experience arises from processing that distorts the stimulus features that are its canonical causes in numerous ways, thereby largely constructing our world of perceived colour. Perceived colour similarity is a systematic misrepresentation of the corresponding stimuli. From an evolutionary perspective, that distortion is either indifferent for the organism's fitness or may even serve its interests. However, in another, crucial respect, colour vision is veridical (or at least highly reliable): where it indicates a salient surface difference, there indeed exists a physical surface difference of value to the organism.

An earlier version of **Jakab's** paper won the prize for best paper by a graduate student at the 2002 Philosophy and Neuroscience Conference where many of the papers in this volume originated.

Consciousness

Most of the philosophical interest in consciousness started from the question of whether consciousness could possibly be a physical process, even a brain process. A common view in philosophy of neuroscience is that if there is anything appropriately given the name 'consciousness' (on this topic, however, there are few genuine eliminativists), it must be physical and, furthermore, explicable in terms of neurophysiology – no explanatory autonomy allowed.

Against this view, Nagel (1974) argued that because conscious experience is subjective, i.e., directly accessible by only the person who has it, we are barred from ever understanding it fully, including whether and if so how it could be physical. For example, even if we knew all there is to know about bat brains, we would not know what it is like to be a bat because bat conscious experience would be so different from human conscious experience. Later, Jackson 1986, McGinn 1991, Chalmers 1996, and others extended this line of thought with new arguments and more sharply delineated conclusions.

Like many neurophilosophers, **Jesse Prinz** simply sidesteps this debate. Using recent neurobiology and cognitive psychology, in 'A Neurofunctional Theory of Consciousness' Prinz argues that consciousness arises when mechanisms of attention allow intermediate-level perceptual systems to access working memory. He then supports this view by appeal to multiple other modalities, which suggests that consciousness has a uniform material basis. He then draws out the implications of his analysis for the traditional mind/body problem. Both leading current approaches, functionalism and identity theory or radical reductivism, fail to appreciate the extent, he argues, to which the solution to the mind-body problem may reply on multiple levels of

analysis, with constitutive contributions at relatively abstract psychological levels and levels that are often dismissed as merely implementational.

In a similar spirit, Akins (1993a, 1993b) and others urge that, for example, an investigation of bat neurophysiology can in fact tell us a lot about what bat 'subjectivity' would be like. It would be boring and myopic. Mandik (2001) actually tries to say what subjectivity would be like neurophysiologically. It would consist in or at least be built on what neuroscientists call egocentric representations. Paul Churchland (1995) had tried the same thing earlier, explaining consciousness in terms of recurrent connections between thalamic nuclei (particularly "diffusely projecting" nuclei like the intralaminar nuclei) and cortex. He showed how thalamocortical recurrency might account for certain features of consciousness such as the effects of short-term memory on consciousness, and dreaming during REM. However, his proposal was a bit short of detail.

Many sceptics about the very possibility of a complete neuroscience of consciousness are unmoved by these analyses. They suspect that either the researcher has changed the topic to something that can be understood neuroscientifically or at best is coming up with an account merely of neural correlates of consciousness (NCCs), not consciousness itself (for an introduction to the issues here, see **Thompson et al.,** this volume, section 8). One common way of arguing that consciousness cannot be anything neural or even physical is via the notion of what philosophers call *qualia:* the introspectible aspects of conscious experiences, what it is like to be conscious of something. Anti-physicalists argue that there could be beings who are behaviourally, cognitively, and even physically exactly like us, yet either have radically different conscious experience, or no conscious experience at all. For example, they might see green where we see red (inverted spectrum) but, because of their training, etc., they use colour words, react to coloured objects, and even process information about colour exactly as we do. Or, they have no conscious experience or colour or anything else – they are zombies – yet use words for conscious experiences, react to experienced objects, and even process information about represented things exactly as we do.

One way to argue that representations can have functionality as representations without consciousness is to appeal to cases of blind sight and inattentional blindness. Due to damage to the visual cortex, blindsight patients have a scotoma, a 'blind spot', in part of their visual field. Ask them what they are seeing there and they will say, "Nothing". However, if you ask them instead to *guess* what is there, they guess with far better than chance accuracy. If you ask them to reach out to touch whatever might be there, they reach out with their hands turned in the right way and fingers and thumb at the right distance apart to grasp anything that happens to be there. And so on (Weiskrantz, 1986).

Inattentional blindness and related phenomena come in many different forms. In one form of it, a subject fixates (concentrates) on a point and is asked to note some feature of an object introduced on or within a few degrees of fixation. After a few trials, a second object is introduced, in the same region but usually not in exactly the same place. Subjects are not told that a second object will appear. When the appearance of the two objects is followed by 1.5 seconds of masking, at least one-quarter of the subjects and sometimes almost all subjects have

no awareness of having seen the second object. (For more on this fascinating group of phenomena, see Mack, http://psyche.cs.monash.edu.au/v7/psyche-7-16-mack.htm or Mack and Rock 1998.)

There is a sense in which the inattentionally blind are not conscious of what they missed: they did not notice and cannot report on the item(s). However, their access to the missed items is extensive, much more extensive than the access that blindsight patients have to items represented in their scotoma. When the second object is a word, for example, subjects clearly encode it and process its meaning. Evidence? When asked shortly after to do, for example, a stem completion task (i.e., to complete a word of which they are given the first syllable or so), they complete the word in line with the word they claim not to have seen much more frequently than controls do. Thus, subjects' access to words that they miss involves the processing of semantic information. If so, their access to the missed words is much like our access to items in our world when we are conscious of them but not conscious of being thus conscious. Thus, far from inattentional blindness suggesting that representations can have full functionality without consciousness, the phenomenon tends to pull in the opposite direction. It is at least fully compatible with the idea that sufficient representational complexity *just is* a form of consciousness.

So, what should we say of the claim that at least some element of some kind of consciousness is not neural or even physical at all? Philosophers who care about neuroscience tend to have three kinds of reaction to this claim.

1. They just ignore the claim (most cognitive scientists),

or,

2. They throw science at it and attempt implicitly or explicitly to produce the kind of account that is supposed to be impossible, (Dennett 1978, Hardin 1988, Austen Clark 1993, Akins 1993a, 1993b, 1996, Hardcastle 1997),

or,

3. They try to show that the claim is wrong (or incoherent, or in some other way epistemically troubled) (Dennett 1991, 1995; Tye 1993; Brook and Raymont, forthcoming).

In 'Making Consciousness Safe for Neuroscience', **Andrew Brook** urges that neither (1) nor (2) is the best course of action. Ignoring the anti-physicalist claim or throwing science at it will leave many – and not just dyed-in-the-wool antiphysicalists – feeling that the real thing, consciousness itself, has been left out, that the researcher has covertly changed the subject and is talking about something else, not consciousness. This is how many react, for example, to suggestions that consciousness is some form of synchronized firing of neurons. 'Surely,' they react, 'you could have the synchronized firing without consciousness. If so, consciousness is not synchronized firing of neurons. Maybe this firing pattern is a *neural correlate* of consciousness, but it is not *what consciousness is.*' Ignoring this reaction, Brook argues, is a bad idea – it not going to fade away on its own.

Moreover, neuroscience is not going to be relevant. The most effective appeal to neuroscience in this context is probably the kind of appeal mounted in **Thompson et al.**, this volume. Because neurophenomenology puts first-person conscious experience front and centre, it does not even have the appearance of leaving consciousness out, changing the topic. However, even such consciousness-centred work can still be accused of studying mere correlates, of not telling us anything about the nature of *consciousness.* According to Brook, what we need to do instead is to tackle the urge to split consciousness off from cognition and the brain and the anti-physicalist arguments that aim to support the urge head-on, to show that attempts to split consciousness off from cognition and the brain do not succeed.

The aim of this volume is to provide a representative, fairly comprehensive snapshot of the work currently going on in the philosophy and neuroscience movement.

References

Akins, K. (1993a) 'What Is It Like to be Boring and Myopic? In B. Dahlbom, ed. *Dennett and His Critics.* New York: Basil Blackwell.

Akins, K. (1993b) A Bat Without Qualities. In M. Davies and G. Humphreys, eds. *Consciousness: Psychological and Philosophical Essays.* New York: Basil Blackwell.

Akins, K. (1996) Of Sensory Systems and The 'Aboutness' of Mental States. *Journal of Philosophy* 93(7), 337-372.

Appelbaum, I. (1998) Fodor, Modularity, and Speech Perception. *Philosophical Psychology* 11, 317–30.

Aston-Jones, G., Desimone, R., Driver, J., Luck, S., and Posner, M. (1999) Attention. In Zigmond, et al. 1999.

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. (2002) *Connectionism and the Mind :Parallel Processing, Dynamics, and Evolution in Networks.* Oxford: Blackwell.

Bechtel, W. and Mundale, J. (1997) 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.

Bickle, J. (1998) *Psychoneural Reduction: The New Wave.* Cambridge, MA: MIT Press.

Biro, J. (1991) Consciousness and Subjectivity in Consciousness. In E. Villaneuva, ed. *Philosophical Issues.* Atascadero, CA: Ridgeview.

Block, N. (1986) Advertisement for a Semantics for Psychology. In French, P. A., ed. *Midwest Studies in Philosophy*, 615-678.

Bliss, T.V.P. and Lomo, T. (1973) Long-Lasting Potentiation of Synaptic Transmission in the Dentate Area of the Anaesthetized Rabbit Following Stimulation of The Perforant Path. *Journal of Physiology (London)* 232, 331-56.

Bower, J. and Beeman, D. (1995) *The Book of GENESIS.* New York: Springer-Verlag.

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

Brook, A. and Stainton, R. (2000) *Knowledge and Mind*. Cambridge, MA: MIT Press/A Bradford Book.

Brook, A. and Raymont, P.(forthcoming) *A Unified Theory of Consciousness.* Cambridge, MA: MIT Press.

Caplan, D., Carr, T., Gould, J., and Martin, R. (1999) Language and Communication. In Zigmond et al. 1999.

Chalmers, D. (1996) *The Conscious Mind.* Oxford: Oxford University Press.

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

Churchland, P. S. and Sejnowski, T. (1992) *The Computational Brain.* Cambridge, MA: MIT Press.

Churchland, P. M. (1979) *Scientific Realism and the Plasticity of Mind.* Cambridge: Cambridge University Press.

Churchland, P. M. (1981) Eliminative Materialism and the Propositional Attitudes. *Journal of Philosophy* 78, 67-90.

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

Churchland, P. M. (1995) *The Engine of Reason, The Seat of the Soul.* Cambridge, MA: MIT Press.

Clark, Andy (1993) *Associative Engines.* Cambridge, MA: MIT Press.

Clark, Austen (1993) *Sensory Qualities.* Cambridge: Cambridge University Press.

Dennett, D. C. (1978) Why You Can't Make a Computer That Feels Pain. *Synthese* 38, 415-449.

Dennett, D. C. (1991) *Consciousness Explained.* New York: Little Brown.

Dennett, D. C. (1995) The Path Not Taken. *Behavioral and Brain Sciences* 18 (2), 252-253..

Descartes R (1649) *Les passions de l'âme*, Amsterdam, Lodewijk Elsevier, and Paris, Henry le Gras. In: Adam, C. & Tannery, P. *Œuvres de Descartes.* Paris: J. Vrin, 1964-74, vol. XI..

Dretske, F. (1981) *Knowledge and the Flow of Information.* Cambridge, MA: MIT Press.

Dretske, F. (1988) *Explaining Behavior.* Cambridge, MA: MIT Press.

Feigl, H. 1(958/1967) *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.

Gold, I. & Stoljar, D. (1999). A Neuron Doctrine in the Philosophy of Neuroscience. *Behavioral and Brain Sciences* 22 (5), 809-30.

Grush, R. (1998) Skill and Spatial Content. *Electronic Journal of Analytic Philosophy* 6(6). http://ejap.louisiana.edu/EJAP/1998/grusharticle98.html.

Grush, R. 2001. The Semantic Challenge to Computational Neuroscience. In P. Machamer, R. Grush and P. McLaughlin, eds. (2001) *Theory and method in the neurosciences* Pittsburgh, PA: University of Pittsburgh Press.

Grush, R. (2002). Cognitive Science. In P. Machamer and M. Silberstein, eds. *Guide to Philosophy of Science.* Basil Blackwell.

Hardcastle, V. G. (1997) When a Pain Is Not. *Journal of Philosophy* 94(8), 381-406.

Hardin, C.L. (1988) *Colour for Philosophers.* Indianapolis, ID: Hackett.

Hebb, D. (1949) *The Organization of Behavior.* New York: Wiley.

Hooker, C. (1981) Towards a General Theory of Reduction. Part I: Historical and Scientific Setting. Part II: Identity in Reduction. Part III: Cross-Categorial Reduction. *Dialogue* 20, 38-59.

Hubel, D. and Wiesel, T. (1962) Receptive Fields, Binocular Interaction and Functional Architecture In the Cat's Visual Cortex. *Journal of Physiology (London)* 195, 215-243.

Jackson, F. (1986) What Mary didn't know *Journal of Philosophy* 83(5), 291-5.

Kandel, E. (1976) *Cellular Basis of Behavior.* San Francisco: W.H. Freeman.

Lehky, S.R. and Sejnowski, T. (1988) Network Model of Shape-from-Shading: Neural Function Arises from Both Receptive and Projective Fields. *Nature* 333, 452-4.

Lettvin, J.Y., Maturana, H.R., McCulloch, W.S., and Pitts, W.H. (1959) What the Frog's Eye Tells the Frog's Brain. *Proceedings of the IRF* 47 (11), 1940-51.

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

Mack, A. Inattentional Blindness. http://psyche.cs.monash.edu.au/v7/psyche-7-16-mack.htm.

Mack, A. and Rock, I. 1998. *Inattentional Blindness.* Cambridge, MA: MIT Press.

Mandik, P. (2001) Mental Representation and the Subjectivity of Consciousness. *Philosophical Psychology* 14 (2), 179-202.

Mandik, P. and Bechtel, Wm. (2002) Philosophy of Science. In: Lynn Nadel, ed. *The Encyclopaedia of Cognitive Science.* London: Macmillan.

McCauley, R. (2001) Explanatory Pluralism and the Co-evolution of Theories of Science. In: Wm. Bechtel, P. Mandik, J. Mundale, and R. S. Stufflebeam, eds. *Philosophy and the Neurosciences: A Reader.* Oxford: Basil Blackwell.

McGinn, C. (1991) *The problem of consciousness: essays towards a resolution* . Oxford: Basil Blackwell.

Milner, A.D. & Goodale, M.A. (1995) *The Visual Brain in Action,* Oxford: Oxford University Press.

Mundale, J., and Bechtel, W. (1996) Integrating Neuroscience, Psychology, and Evolutionary Biology through a teleological conception of function. *Minds and Machines,* 6, 481–505.

Nagel, T. (1971) Brain Bisection and the Unity of Consciousness. *Synthese* 22, 396-413.

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

Place, U.T. (1956) Is Consciousness a Brain Process? *The British Journal of Psychology* XLVII(1), 44-50.

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

Smart, J.J.C. (1959) Sensations and Brain Processes. *Philosophical Review* 68, 141-156.

Stich, S. (1983) *From Folk Psychology to Cognitive Science.* Cambridge, MA: MIT Press.

Stufflebeam, R. and Bechtel, Wm. (1997) PET: Exploring the Myth and the Method. *Philosophy of Science (Supplement)*, 64(4): S95-S106.

Tye, M. (1993) Blindsight, the Absent Qualia Hypothesis, and the Mystery of Consciousness. In C. Hookway, ed., *Philosophy and the Cognitive Sciences.* Cambridge: Cambridge University Press.

van Gelder, T. (1998) Mind as Motion: Explorations in the Dynamics of Cognition. *Journal of Consciousness Studies* 5(3), 381-383.

von Eckardt Klein, B. (1975). Some Consequences of Knowing Everything (Essential) There is to Know About one's Mental States. *Review of Metaphysics* 29, 3-18.

von Eckardt Klein, B. (1978) Inferring Functional Localization from Neurological Evidence. In E. Walker, ed. *Explorations in the Biology of Language.* Cambridge, MA: MIT Press.

Weiskrantz, L. (1986). *Blindsight: A Case Study and Implications.* Oxford: Clarendon Press .

Zigmond, M., Bloom, F., Landis, S., Roberts, J., and Squire, L., eds. 1999. *Fundamental Neuroscience.* San Diego: Academic Press.