Analyse & Kritik 26/2004 (© Lucius & Lucius, Stuttgart) p. 382–397

Andrew Brook/Pete Mandik

The Philosophy and Neuroscience Movement*

Abstract: A movement dedicated to applying neuroscience to traditional philosophical problems and using philosophical methods to illuminate issues in neuroscience began about twenty-five years ago. Results in neuroscience have affected how we see traditional areas of philosophical concern such as perception, belief-formation, and consciousness. There is an interesting interaction between some of the distinctive features of neuroscience and important general issues in the philosophy of science. And recent neuroscience has thrown up a few conceptual issues that philosophers are perhaps best trained to deal with. After sketching the history of the movement, we explore the relationships between neuroscience and philosophy and introduce some of the specific issues that have arisen.

0. Introduction

The exponentially-growing body of work on the human brain of the past few decades has not only taught us a lot about how the brain does cognition, it has also had a profound influence on other disciplines that study cognition and behaviour. A notable example, interestingly enough, is philosophy. A small movement dedicated to applying neuroscience to traditional philosophical problems and using philosophical methods to illuminate issues in neuroscience began about 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

* This paper draws upon material in Andrew Brook/Pete Mandik, 'Introduction', Andrew Brook/Kathleen Akins (eds.), *Cognition and the Brain: The Philosophy and Neuroscience Movement*, pp. 1–27. © Cambridge University Press 2005, reproduced with permission.

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.

As a result of this interest, a group of philosophers and neuroscientists dedicated to informing each other's work has grown up. Many of these people now have Ph.D.-level training or the equivalent in both neuroscience and philosophy. Much of the work that has appeared has been clustered around five themes,

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

and two big issues lie in the substructure of all of them,

- the relationship of neuroscience to the philosophy of science,

and,

- whether cognitive science will be reduced to neuroscience or eliminated by it.

We will take up these themes shortly. But first, a sketch of some relevant history.

1. 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 20^{th} century, even when the idea that the mind is simply the 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. In this paper, he 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 and did 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 various visual properties with various neural processes. A decade earlier, Donald Hebb (1949) had tried to explain cognitive phenomena such as 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 more than 25,000 members.) In the mid-1970s, the term 'cognitive science' was adopted as the label for interdisciplinary studies of cognition and the idea took hold that what we mean by 'mind' is primarily a set of functions for processing information. The idea of information processing might not have been much more than a uniting metaphor without the advent of large-capacity computers. For over three decades now, real effort has been put into implementing the relevant functions in computational systems (this is one leading kind of 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 2000 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 lower-case for some reason) brain scans in particular. In the sciences, psycholo-

gist 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 and widely read). Churchland's book has three main aims:

1. to develop an account of intertheoretic reduction and specifically of the reduction of mind to brain radically different from the one in logical positivist philosophy of science;

2. to show that consciousness-based objections to psychoneural reduction do not work,
   and,

3. to show that functionalist/multiple realizability objections to psychoneural reduction do not 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. We will say something about the relevant history of each area and examine briefly what is going on currently. By and large, the topics of primary interest in the philosophy of neuroscience are topics that tie the issue of the relationship of the mind (cognition) to the brain into current philosophy of science.

Indeed, 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/Stoljar 1999, explore the relationship of neuroscience and the cognitive sciences in detail). There are a few exceptions. An important philosophical study of areas of neuroscience not directly relevant to cognition is Machamer et al. (2000), who discuss individual neurons, how neurons work, and so on. But that is the general pattern.

We now turn to the big background issues identified earlier, namely, neuroscience and the philosophy of science; and reductionism vs eliminativism in neuroscience and cognitive science.

2. Neuroscience and the Philosophy of Science

In much early philosophy of science, the notion of law is central, 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/Richardson 1993; Bechtel 2007; Machamer/Craver 2000; Craver 2007). 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/Bechtel 2002).

One main reason why neuroscience raises issues such as these in stark form is that, while there is clearly an enormous amount of structure in the brain (the human brain is made up of roughly 100,000,000,000 neurons), neuroscience has had very little success in finding general laws that all or nearly all brains instantiate. Maybe for at least the complex kinds of activity that underpin cognition, it will turn out that there are no such laws, just masses and masses of individually-distinct (though still highly structured) events.

A related challenge to logical positivist philosophy of science questions whether scientific theories are best considered to be sets of sentences at all. Paul (P. M.) Churchland (1989, 2007), 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. This challenge is directly connected to the next issue.

3. Reduction Versus Elimination

There are three general views concerning the relation between the psychological states posited by cognitive science 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, cognitive science for cognitively-delineated types of activity, neuroscience for activities described in terms of brain processes and structures (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 neurophysiological states; every cognitively-delineated type can be mapped onto some type of brain process and/or structure with nothing much left over. The history of science has been in no small part a history of such reduction, as they are (somewhat misleadingly) called (misleading because the reduced kinds still continue to exist): Chemistry has been shown to be a branch of physics, large parts of biology have been shown to be a branch of chemistry, and so on. Reductivists about cognition (or psychology generally) believe that cognition (and psychology generally) 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 that psychological states are best regarded as talking about nothing that actually exists.

To give just one example of the kind of argument mounted in support of eliminativism, phenomena identified using psychological concepts are difficult if not impossible to quantify precisely, but all successful sciences quantify the kinds of thing of interest. Eliminativist arguments are anti-reductivist in one very important way: They argue that there is no way to reduce psychological theories to neural theories and even if there were, there would be no point in doing so.

Philosophers of neuroscience generally fall into either the reductionist or the eliminativist camps. Most are reductionists of some stripe—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 most are also not very 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 accept that these concepts should be eliminated; we shouldn't even try to reduce them. 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. To give just one example, though it is a leading example, almost nobody, not even the 'high priests' of eliminativism, Paul and Patricia Churchland, have ever argued that the concept of consciousness should be eliminated. Maybe it should be shaped up a bit, trimmed back in places, but nearly everyone holds that the concept refers to something real and important.

Some philosophers of neuroscience explicitly advocate a mixture of the two. For instance, the Churchlands 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, while they have held that 'folk concepts' such as belief and desire do not name anything real, scientific psychological concepts such as representation do (so long as we keep our notion of representation neutral with respect to the various theories of what representations are). Many kinds of representation will ultimately be found to be identical to some particular kind of neural state or process (P. S. Churchland 1986).

In the space we have, 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) The first concerns 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 *our 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 were built into folk psychology, then folk psychology would probably be 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. The burgeoning results in neuroscience have thrown the issues up in high relief and sometimes have given them new content. Work on them by philosophers has helped neuroscientists develop a more precise sense of exactly how their work relates to other scientific work, cognitive science in particular. It is interesting, even a bit remarkable, that, as we noted, most people in the philosophy and neuroscience movement have arrived at roughly the same position on them. Thus, even though they form the background to most current work, we will say no more about them.

On many more specific topics, we are far from having such a settled position. We turn now to a representative sample of these topics. We identified them earlier:

- Data and theory in neuroscience.

The issue of the relationship of data to theory contains a huge group of subissues. We will restrict ourselves to two issues: Can introspection generate good data for neuroscience? And, is function in the brain localized to specific regions (often referred to as modules) or spread across wide areas of the brain.

- Neural representation and computation.

A huge topic! Here we will focus on the architecture, syntax, and semantics of neural representation

- Visuomotor transformation.

Under this heading, we will examine two issues. The first concerns the hypothesis that we have two visual systems: one for conscious perception and the other for action. The second concerns the increasingly popular hypothesis that perception and control of behaviour are interdependent.

- Colour vision

Here the big issue is over how to think about the relationship between colour experiences and the distal stimuli that elicit such experiences. One thing that is puzzling about this relationship is that the ways in which colours are *experienced* diverges quite dramatically from the ways in which their environmental triggers—the ostensible colours themselves—actually *are*.

- Consciousness

Two pressing issues among many in connection with consciousness are the issue, first, whether consciousness is just a part of cognition or something unique and, in some measure at least, beyond the reach of either cognitive science or neuroscience forever and, second, if consciousness is cognitive, what kind of cognitive process/structure is it, and if it is not, with what kind of cognitive and brain processes and structures is it associated?[1]

[1]Papers discussing all these issues in more detail can be found in Brook/Akins 2005. For other recent anthologies of articles in the growing intersection of philosophy and neuroscience see Bechtel et. al 2001, Keeley 2006, and Bickle (in press).

4. Data and Theory: Introspection, Localization, Modularity

4.1 Introspection

In a variety of ways, the advent of sophisticated imaging of brain activity has created a new reliance on introspection—it is difficult if not impossible to relate what is going on cognitively to various brain activities without self-reports from subjects about what is going on in them. Introspection has been in bad odour as a research tool for over 100 years. It has variously been claimed that

1. introspective claims are unreliable because they are not regularly replicated in others.

2. subjects confabulate (make up stories) about what is going on in themselves when they need to do so to make sense of behaviour.

3. introspection has access only to a tiny fraction of what is going on in oneself cognitively.

4. it is impossible for introspection to access brain states.

And so on. However, researchers into the brain (neuroscience) have been forced back onto introspection because often the only access that they have to the cognitive and conscious functions that they want to study in the brain is the access that the subject him or herself has. It is perhaps a bit ironic that neuroscience, the most scientific of approaches to human nature to date, has been forced to fall back onto a technique rejected as unscientific over 100 years ago!

One interesting middle position gaining some currency is that some of the limitations in introspection can be overcome by training. After training and practise introspection can come to be much more reliable than it is in its native state—perhaps even reliable enough to introspectively identify internal events in terms of the taxonomy of neuroscience, and thus introspect brain states as such (Churchland 1989; Mandik 2006).

Another way around the classical and classically unreliable appeal to introspection is to point out that not all first-person utterances are introspective reports. Perhaps when first-person utterances are expressing feelings, for example, they are or at least can be more reliable sources of data that first-person utterances that report self-observations are. Among other things, with first person utterances that express rather than report, there may no longer be a conflict between the use of subjectively-rooted utterances and the requirement that evidence be public.

4.2 Localization

A question with a long history in the study of the brain concerns how localized cognitive function is. Early localization theorists (early 1800s) included the phrenologists Gall and Spurzheim. Flourens was a severe early critic of the idea from the same period.

Localizationism re-emerged 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. However, resolution improves every year, especially in fMRI.

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 to 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/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.

However much MRI may assume and depend on the idea that cognitive function is localized in the brain, the idea faces grave difficulties. Even a system as simple and biologically basic as oculomotor control (the control system that keeps the eyes pointed in one direction as the head moves around, for example)

is the very reverse of localized. Units dispersed widely across cortex contribute to performing this function. Moreover, many of these units are also involved in many other information-processing and control activities. A further factor pointing in the same anti-localization direction is that the brain is very plastic, especially in childhood. If one area is damaged, often another area can take over the functions previously performed by the damaged area. (Since these claims arise from actually observing how the brain does things, they also undermine the old idea that we can study cognitive function without studying the brain.)

4.3 Modularity

The question of localization connects to the question of modularity, another big issue in cognitive neuroscience. Fodor (1983) advanced a strong modularity thesis concerning the cognitive architecture of peripheral systems (vision, language, touch, and the like). 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 argued 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. The latter is what the work sketched two paragraphs ago would suggest. Another body of evidence supporting the same conclusion concerns back projection. Temporal cortical areas implicated in high levels of visual processing, for example, 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.

5. Neural Representation and Computation

Neural representation and computation is a huge topic, as we said. We will start with neural representation.

The neurophilosophical questions concerning computation and representation nearly all assume a definition of computation in terms of transformation of representations. Thus, most questions concerning computation and representation are really questions about representation. Contributions to this topic can be

thought of as falling into three groups, though the boundaries between them are far from crisp. There are 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.

5.1 Architecture of Neural Representation

Here is some of the thinking afoot currently about neural architecture. Past approaches to understanding the mind, including symbolicism, connectionism, and dynamicism, rely heavily on metaphors. A much less metaphorical approach, or so it is claimed, unifies representational and dynamical descriptions of the mind. First, representation is rigorously defined by encoding and decoding relations. Then, the variables identified at higher levels are treated as state variables in control theoretical descriptions of neural dynamics. Given the generality of control theory and representation so defined, it is claimed that this approach is sufficiently powerful to unify descriptions of cognitive systems from the neural to the psychological levels. If so, contrary to dynamicist arguments (van Gelder 1998), one can have both representation and dynamics in cognitive science.

5.2 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/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 models activity in and interactions between patches of the neuron's membrane (Bower/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.

5.3 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 a century before (he called it 'the mark of the mental', only a slight exaggeration).

How do neural states come to have contents? There are three broad answers to this question that have been popular in philosophy: The isomorphism approach, the functional role approach and the informational approach. All three appear in the philosophy of neuroscience.

Proponents of the isomorphism approach construe representation as a relation of resemblance that obtains between representations and that which they represent. Such resemblances are often thought to be abstract or second-order resemblance, meaning, for instance even though a representation and what it represents might not have a first-order resemblance of being, e.g., the same colour, they may still enter into systems of relationships such that the relationships may be mapped onto one another (as in the mapping of various heights of a mercury column in a thermometer onto various temperatures). (See, for instance, Churchland 2007; Mandik et. al 2007; O'Brien/Opie 2004).

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 role 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/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.

One issue of considerable interest is how the brain does time. How the brain does objective time, actual persistence, is interesting enough but even more interesting is the subjective time of behaviour: the temporal representation that is analogous to egocentric space (in contrast to objective or allocentric space). How can times be represented in the brain so that when we recall them, we recall them as falling into a temporal order. How, for example, when we recall a series of sounds that we have heard, do we hear it as a melody rather than as a chord?

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.

5.4 Visuomotor Transformation

A specific but absolutely central topic to do with neural representation is visuomotor transformation, that is to say, how we use visual information to guide motor control.

Here the leading theory, due to Milner and Goodale (1995), is that we have two complementary visual systems, vision-for-perception and vision-for-action. They base their conclusion 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 (to the side and near the bottom of the brain) and the dorsal pathway (to the rear and near the top of the brain) in the visual system of primates. Probably no other claim in cognitive neuroscience has attracted as much attention as this one in the past ten or twelve years.

Another important body of work in visuomotor control focuses on the idea that spatial perception and motor output are interdependent. There are two broad approaches. One posits mental representations mediating between perception and action. This approach is often called representationalism. The other approach, a kind of antirepresentationalism, opposes this idea, arguing that intelligent, visually guided behaviour can be explained without positing intermediaries with representational or semantic properties between sensory input and motor output.

5.5 Colour Vision

The final two issues on which we will focus in this quick survey of issues currently at the interface between philosophy and neuroscience are colour vision and consciousness. Any complete theory of neural representation would have to contain a theory of both.

The biggest issue to do with colour vision, as we said, is the issue of how to think about the relationship of colour experience to the causal factors that produce colour experience. For example, experiences of different colours are the result of combinations of intensities of light of the three broad wavelengths of light to which the retina is sensitive (four wavelengths in some women) plus other factors. Light of three intensities at three wavelengths is nothing like redness as one experiences it. So how should we think of the relationship between the two?

Even worse, some argue that colour experience arises from processing that distorts the stimulus features that are its main causes, thereby largely constructing a world of perceived colour that has almost nothing to do with how the world actually is. For these people, perceived colour similarity is a systematic misrepresentation of the corresponding stimuli. How such systematic misrepresentation could have come to have a survival or reproductive advantage is just one of the puzzling, even baffling questions to which contemporary work in neuroscience on colour gives rise.

Most remarkably of all, we can have colour experiences that represent physically impossible colours. In a stunning example of neurophilosophy at work, Paul Churchland (2005) has shown that by exploiting shifts in experienced colour due to tiredness and habituation, experiences of colour can be brought about where the colours could not exist on standard colour wheels and other theories of the structure of colour and, moreover, would require physically-impossible states, for example, that things in one's world be emitting light and be emitting no light at the same time. Indeed, as Churchland shows, some of the colour experiences that we can have cannot even be represented by a colour sample.

6. Consciousness

Most of the philosophical interest in consciousness starts from the question of whether consciousness could possibly be a physical process of any kind, let alone a brain process. A common view in philosophy of neuroscience is that everything to do with the mind, including consciousness, will turn out to be explicable in terms of neurophysiology—not even explanatory autonomy is allowed. If consciousness is not something that neuroscience can capture, then that hallowed shibboleth of neuroscience will be false and there will be at least severe limitations on the extent to which there could ever be a science of consciousness.

In the face of claims that at least something about consciousness is not neural or even physical at all, cognitive neuroscientists and their philosopher fellow-travellers have tended to one or the other of three different kinds of reaction:

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

2. They just ignore the claim. This is the approach taken by many cognitive and neuro-scientists. Or,

3. 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; Clark 1993; Akins 1993a; 1993b; 1996; Hardcastle 1997; Baars 1988; Rosenthal 1991; Mandik, in press).

The usual way to argue the main idea in (1), that there is nothing unique or *sui generis* about consciousness, is to tackle the arguments that claim that there is and try to show that they do not work. Here is a sample of such arguments. 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 zombie thought experiments and thought experiments about colour scientists who have never experienced colour.

Zombie thought experiments are representative of the genre. Consider what philosophers call *qualia:* the introspectible aspects of conscious experiences, what it is like to be conscious of something. Those who hold that consciousness is something unique argue that there could be beings who are behaviourally, cognitively, and even physically exactly like us, yet they have no conscious experience at all. If so, conscious experience cannot be a matter of behaviour, cognition, or physical makeup.

A variant, inverted spectrum thought experiments, urge that others could have radically different conscious experience of, in this case, colour with no change in behaviour, cognition, or physical makeup. 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. If inverted spectra are possible, then the same conclusion follows as from the alleged possibility of zombies: consciousness is not safe for neuroscience.

Zombie and inverted spectra arguments strive to show that representations can have functionality as representations without consciousness. A more scientific way to argue for a similar conclusion involves 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, 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.[2]

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, it can be argued that in another sense, they are conscious of the things on which they cannot report. For example, 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 have been 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. Thus, an alternative account of the 'blindness' in these cases is that subjects are conscious of what they cannot report but are not conscious of being thus conscious (so they cannot report it). If so, far from inattentional blindness suggesting that representations can have full functionality without consciousness, the phenomenon would pull in the opposite direction. At minimum, it seems to be at least fully compatible with the idea that consciousness is a form of cognition (see Mandik 2005).

So what about philosophers and neuroscientists who ignore that challenging claim that consciousness is unique, *sui generis,* or throw science at it? Their numbers are legion and we won't attempt to examine the various alternative theories here. They range from attention theories to global work space theories to pandemonium architecture models to connectionist and dynamic systems models. Recent neuroscience has made a lot of progress in identifying the regions and systems in the brain most closely associated with consciousness of various kinds (Koch 2004 gives an excellent summary). What is important here is that nearly all this work starts from a common assumption, that consciousness is a fairly standard cognitive phenomenon that can be captured in the kind of theory that captures cognitive functioning in general, without any attempt to argue for the assumption.

Ignoring the challenging claim is risky (not to mention a bit rude). It risks leaving many—and not just dyed-in-the-wool anticognitivists—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. This is how many react to suggestions that consciousness is, for example, 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 (NCC), but it is not *what consciousness is".*

[2]For more on this fascinating group of phenomena, see Mack, `http://psyche.cs.monash.edu.au/v7/psyche-7-16-mack.html` or Mack/Rock 1998.

Throwing science at the challenge faces exactly the same risk. No matter what the scientific model, sceptics about a science of consciousness can always claim that the model is not a model of *consciousness,* that the researcher has changed the topic to something that can be understood neuroscientifically, is merely talking about correlates of consciousness (NCCs) or whatever. Moreover, it would appear that no amount of neuroscience could make this objection irrational. No matter what the scientific model of consciousness, the charge can always be levelled that the model is studying mere correlates, that it is not uncovering the nature of *consciousness.* Many now believe that the only approach with any hope of success so far as a science of consciousness is concerned is to beard the sceptics in their lair, to tackle the arguments that they advance and show that they just don't work or worse, are incoherent. To make consciousness safe for neuroscience, we would have to show one (or both) of two things. The first would be that the sceptics have given no good reason to believe that consciousness is not safe for neuroscience. The second, and perhaps stronger, would be to show that consciousness is not and could not be unique in the way required by sceptics. Both of these are pre-eminently philosophical tasks.

In general, at the interface between neuroscience and philosophy at the moment, there is a great ferment. Results in neuroscience are shedding light on, even reshaping, traditional philosophical hunches about and approaches to the mind. And neuroscience is throwing up some new issues of conceptual clarification and examination of possibilities that philosophers are better equipped to handle than anyone else. We live in interesting times!

Bibliography

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

— (1993b), A Bat Without Qualities, in: M. Davies/G. Humphreys (eds.), *Consciousness: Psychological and Philosophical Essays,* New York

— (1996), Of Sensory Systems and the 'Aboutness' of Mental States, in: *Journal of Philosophy 93(7),* 337–372

Baars, D. (1988), *A Cognitive Theory of Consciousness,* Cambridge

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

—/J. Mundale (1997), Multiple Realizability Revisited: Linking Cognitive and Neural States, in: *Philosophy of Science 66*, 175–207

—/A. Abrahamsen (2002), *Connectionism and the Mind: Parallel Processing, Dynamics, and Evolution in Networks,* Oxford

—/G. Graham (1998), The Life of Cognitive Science, in: W. Bechtel/G. Graham (eds.), *A Companion to Cognitive Science.* Oxford, 1–104

—/P. Mandik/J. Mundale/R. Stufflebeam (eds.) (2001), *Philosophy and the Neurosciences: A Reader,* Oxford

— (2007), *Mental Mechanisms,* London

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

— (in press) (ed.) *The Oxford Handbook of Philosophy and Neuroscience*, Oxford

Biro, J. (1991), Consciousness and Subjectivity in Consciousness, in: E. Villaneuva

(ed.), *Philosophical Issues,* Atascadero/CA
Block, N. (1986), Advertisement for a Semantics for Psychology, in: French, P. A. (ed.), *Midwest Studies in Philosophy*, 615—678
Bliss, T. V. P./T. Lomo (1973), Long-Lasting Potentiation of Synaptic Transmission in the Dentate Area of the Anaesthetized Rabbit Following Stimulation of The Perforant Path, in: *Journal of Physiology (London)* 232, 331–356
Bower, J./D. Beeman (1995), *The Book of GENESIS,* New York
Brentano, F. (1874), *Psychology from an Empirical Standpoint.* A. C. Pancurello/ D. B. Tyrrell/L. L. McAlister, trans., New York
Brook, A./K. Akins (eds.) (2005), *Cognition and the Brain: The Philosophy and Neuroscience Movement,* New York
— /R. Stainton] (2000), *Knowledge and Mind,* Cambridge/MA
— /P. Raymont] (forthcoming), *A Unified Theory of Consciousness,* Cambridge/MA
Caplan, D./T. Carr/J. Gould/R. Martin (1999), Language and Communication, in: Zigmond et al. 1999
Chalmers, D. (1996), *The Conscious Mind,* Oxford
Churchland, P. M. (1979), *Scientific Realism and the Plasticity of Mind,* Cambridge
— (1981), Eliminative Materialism and the Propositional Attitudes, in: *Journal of Philosophy 78*, 67–90
— (1989), *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science,* Cambridge/MA
— (1993), *Sensory Qualities,* Cambridge
— (1995), *The Engine of Reason, The Seat of the Soul ,* Cambridge/MA
— (2005), Chimerical Colours, in: Brook, A./K. Akins (eds.) (2005)
— (2007), *Neurophilosophy at Work*, Cambridge
Churchland, P. S. (1986) *Neurophilosophy,* Cambridge/MA
—/T. Sejnowski (1992), *The Computational Brain,* Cambridge/MA
Clark, A. (1993), *Associative Engines,* Cambridge/MA
Craver, C. (2007), *Explaining the Brain,* Oxford
Dennett, D. C. (1978), Why You Can't Make a Computer That Feels Pain, in: *Synthese 38*, 415–449
— (1991), *Consciousness Explained,* New York
— (1995), The Path Not Taken, in: *Behavioral and Brain Sciences 18 (2),* 252–253
Descartes R. (1649), *Les passions de l'âme,* Amsterdam, in : Adam, C./P. Tannery, *Œuvres de Descarte,* Paris, 1964–74, vol. XI..
Dretske, F. (1981), *Knowledge and the Flow of Information,* Cambridge/MA
— (1988) *Explaining Behavior,* Cambridge/MA
Feigl, H. (1958/1967), *The 'Mental' and the 'Physical'. The Essay and a Postscript,* Minneapolis
Fodor, J. A. (1974), Special Sciences (or: The Disunity of Science as a Working Hypothesis), in: *Synthese 28*, 97–115
— (1983), *The Modularity of Mind,* Cambridge/MA
Gold, I./D. Stoljar (1999), A Neuron Doctrine in the Philosophy of Neuroscience, in: *Behavioral and Brain Sciences 22 (5),* 809–830
Grush, R. (1998), Skill and Spatial Content, in: *Electronic Journal of Analytic Philosophy 6(6)*, `http://ejap.louisiana.edu/EJAP/1998/grusharticle98.html`
— (2001), The Semantic Challenge to Computational Neuroscience, in: P. Machamer/R. Grush/P. McLaughlin (eds.) (2001), *Theory and Method in the Neurosciences,* Pittsburgh/PA
— (2002), Cognitive Science, in: P. Machamer/M. Silberstein (eds.). *Guide to Phi-*

losophy of Science, Oxford
Hardcastle, V. G. (1997), When a Pain Is Not, in: *Journal of Philosophy 94(8),* 381–406
Hardin, C. L. (1988), *Colour for Philosophers,* Indianapolis/ID
Hebb, D. (1949), *The Organization of Behavior,* New York
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, in: *Dialogue 20*, 38–59
Hubel, D./T. Wiesel (1962), Receptive Fields, Binocular Interaction and Functional Architecture In the Cat's Visual Cortex, in: *Journal of Physiology (London) 195*, 215–243
Jackson, F. (1986), What Mary didn't Know, in: *Journal of Philosophy 83(5),* 291–295
Kandel, E. (1976), *Cellular Basis of Behavior,* San Francisco
Keeley, Brian (ed.) (2006), *Paul M. Churchland (Contemporary Philosophy in Focus),* Cambridge
Koch, C. (2004), *Quest for Consciousness,* Englewood/CO
Lehky, S. R./T. Sejnowski (1988), Network Model of Shape-from-Shading: Neural Function Arises from Both Receptive and Projective Fields, in: *Nature 333*, 452-454
Lettvin, J. Y./H. R. Maturana/W. S. MCulloch/W. H. Pitts (1959), What the Frog's Eye Tells the Frog's Brain, in: *Proceedings of the IRF 47 (11),* 1940–1951
Machamer, P./L. Darden/C. Craver (2000), Thinking about Mechanisms, in: *Philosophy of Science 67*, 1–25
Mack, A. (?????), Inattentional Blindness, `http://psyche.cs.monash.edu.au/v7/psyche-7-16-mack.htm`
—/I. Rock (1998), *Inattentional Blindness,* Cambridge/MA
Mandik, P. (2001), Mental Representation and the Subjectivity of Consciousness, in: *Philosophical Psychology 14 (2),* 179–202
— (2005), Phenomenal Consciousness and the Allocentric-Egocentric Interface, in: R. Buccheri et al. (eds.), Endophysics, Time, Quantum and the Subjective World, ???????????????
— (2006), The Instrospectibility of Brain States as Such, in: Keeley 2006
— (in press), The Neurophilosophy of Subjectivity, in: J. Bickle (ed.), *The Oxford Handbook of Philosophy and Neuroscience,* Oxford
—/W. Bechtel (2002), Philosophy of Science, in: L. Nadel (ed.), *The Encyclopaedia of Cognitive Science,* London
—/M. Collins/A. Vereschagin (2007), Evolving Artificial Minds and Brains, in: A. Schalley/D. Khlentzos (eds.), *Mental States, Vol. 1: Nature, Function, Evolution,* Amsterdam
McCauley, R. (2001), Explanatory Pluralism and the Co-evolution of Theories of Science, in: W. Bechtel/P. Mandik/J. Mundale/R. S. Stufflebeam (eds.), *Philosophy and the Neurosciences: A Reade,* Oxford
McGinn, C. (1991), *The Problem of Consciousness: Essays Towards a Resolution,* Oxford
Milner, A. D./M. A. Goodale (1995), *The Visual Brain in Action,* Oxford
Mundale, J./W. Bechtel| (1996), Integrating Neuroscience, Psychology, and Evolutionary Biology through a Teleological Conception of Function, in: *Minds and Machines 6*, 481–505
Nagel, T. (1971), Brain Bisection and the Unity of Consciousness, in: *Synthese 22*, 396-413

Newell, A. (1980), Physical Symbol Systems, in: *Cognitive Science 4*, 135–183

O'Brien, G./J. Opie (2004), Notes Toward a Structuralist Theory of Mental Representation, in: H.Clapin/P.Staines/P.Slezak (eds.), Representation in Mind: New Approaches to Mental Representation, Amsterdam

Place, U. T. (1956), Is Consciousness a Brain Process? In: *The British Journal of Psychology 47(1),* 44–50

Putnam, H. (1967), Psychological Predicates, in: W. H. Capitan/D. D. Merrill (eds.), *Art, Mind and Religion,* Pittsburgh

Rosenthal, D. (1991), *The Nature of Mind,* Oxford

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

Stich, S. (1983), *From Folk Psychology to Cognitive Scienc,* Cambridge/MA

Stufflebeam, R./W. Bechtel (1997), PET: Exploring the Myth and the Method, in: *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

van Gelder, T. (1998), Mind as Motion: Explorations in the Dynamics of Cognition, in: *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, in: *Review of Metaphysics 29*, 3–18

— (1978), Inferring Functional Localization from Neurological Evidence, in: E. Walker (ed.), *Explorations in the Biology of Language,* Cambridge/MA

Weiskrantz, L. (1986), *Blindsight: A Case Study and Implication,* Oxford

Zigmond, M./F. Bloom/S. Landis/J. Roberts/L. Squire (eds.) (1999), *Fundamental Neuroscience,* San Diego