The Philosophy of Neuroscience

First published Mon Jun 7, 1999; substantive revision Tue May 25, 2010

Over the past three decades, philosophy of science has grown increasingly "local." Concerns have switched from general features of scientific practice to concepts, issues, and puzzles specific to particular disciplines. Philosophy of neuroscience is a natural result. This emerging area was also spurred by remarkable recent growth in the neurosciences. Cognitive and computational neuroscience continues to encroach upon issues traditionally addressed within the humanities, including the nature of consciousness, action, knowledge, and normativity. Empirical discoveries about brain structure and function suggest ways that "naturalistic" programs might develop in detail, beyond the abstract philosophical considerations in their favor.

The literature distinguishes "philosophy of neuroscience" and "neurophilosophy." The former concerns foundational issues within the neurosciences. The latter concerns application of neuroscientific concepts to traditional philosophical questions. Exploring various concepts of representation employed in neuroscientific theories is an example of the former. Examining implications of neurological syndromes for the concept of a unified self is an example of the latter. In this entry, we will assume this distinction and discuss examples of both.

1. Before and After *Neurophilosophy*

Contrary to some opinion, actual neuroscientific discoveries have exerted little influence on the details of materialist philosophies of mind. The "neuroscientific milieu" of the past four decades has made it harder for philosophers to adopt dualism. But even the "type-type" or "central state" identity theories that rose to brief prominence in the late 1950s (Place, 1956; Smart, 1959) drew upon few actual details of the emerging neurosciences. Recall the favorite early example of a psychoneural identity claim: pain is identical to C-fiber firing. The "C fibers" turned out to be related to only a single aspect of pain transmission (Hardcastle, 1997). Early identity theorists did not emphasize psychoneural identity hypotheses, admitting that their "neuro" terms were placeholders for concepts from future neuroscience. Their arguments and motivations were philosophical, even if the ultimate justification of the program was held to be empirical.

The apology for this lacuna by early identity theorists was that neuroscience at that time was too nascent to provide any plausible identities. But potential identities were afoot. David Hubel and Torsten Wiesel's (1962) electrophysiological demonstrations of the receptive field properties of visual neurons had been reported with great fanfare. Using their techniques, neurophysiologists began discovering neurons throughout visual cortex responsive to increasingly abstract features of visual stimuli: from edges to motion direction to colors to properties of faces and hands. More notably, Donald Hebb had published *The Organization of Behavior* (1949) a decade earlier. Therein he offered detailed explanations of psychological phenomena in terms of known neural mechanisms and anatomical circuits. His psychological explananda included features of perception, learning, memory, and even emotional disorders. He offered these explanations as potential identities. (See the Introduction to his 1949). One philosopher who did take note of some available neuroscientific detail was Barbara Von Eckardt-Klein (1975). She discussed the identity theory with respect to sensations of touch and pressure, and incorporated then-current hypotheses about neural coding of sensation modality, intensity, duration, and location as theorized by Mountcastle, Libet, and Jasper. Yet she was a glaring exception. By and large, available neuroscience at the time was ignored by both philosophical friends and foes of early identity theories.

Philosophical indifference to neuroscientific detail became "principled" with the rise and prominence of functionalism in the 1970s. The functionalists' favorite argument was based on multiple realizability: a given mental state or event can be realized in a wide variety of physical types (Putnam, 1967; Fodor, 1974). So a detailed understanding of one type of realizing physical system (e.g., brains) will not shed light on the fundamental nature of mind. A psychological state-type is autonomous from any single type of its possible realizing physical mechanisms. (See the entry on "Multiple Realizability" in this Encyclopedia, linked below.) Instead of neuroscience, scientifically-minded philosophers influenced by functionalism sought evidence and inspiration from cognitive psychology and "program-writing" artificial intelligence. These disciplines abstract away from

underlying physical mechanisms and emphasize the "information-bearing" properties and capacities of representations (Haugeland, 1985). At this same time neuroscience was delving directly into cognition, especially learning and memory. For example, Eric Kandel (1976) proposed presynaptic mechanisms governing transmitter release rate as a cell-biological explanation of simple forms of associative learning. With Robert Hawkins (1984) he demonstrated how cognitivist aspects of associative learning (e.g., blocking, second-order conditioning, overshadowing) could be explained cell-biologically by sequences and combinations of these basic forms implemented in higher neural anatomies. Working on the post-synaptic side, neuroscientists began unraveling the cellular mechanisms of long term potentiation (LTP) (Bliss and Lomo, 1973). Physiological psychologists quickly noted its explanatory potential for various forms of learning and memory.[1]Yet few "materialist" philosophers paid any attention. Why should they? Most were convinced functionalists. They believed that the "engineering level" details might be important to the clinician, but were irrelevant to the theorist of mind.

A major turning point in philosophers' interest in neuroscience came with the publication of Patricia Churchland's *Neurophilosophy* (1986). The Churchlands (Patricia and Paul) were already notorious for advocating eliminative materialism (see the next section). In her (1986) book, Churchland distilled eliminativist arguments of the past decade, unified the pieces of the philosophy of science underlying them, and sandwiched the philosophy between a five-chapter introduction to neuroscience and a 70-page chapter on three then-current theories of brain function. She was unapologetic about her intent. She was introducing philosophy of science to neuroscientists and neuroscience to philosophers. Nothing could be more obvious, she insisted, than the relevance of empirical facts about how the brain works to concerns in the philosophy of mind. Her term for this interdisciplinary method was "co-evolution" (borrowed from biology). This method seeks resources and ideas from anywhere on the theory hierarchy above or below the question at issue. Standing on the shoulders of philosophers like Quine and Sellars, Churchland insisted that specifying some point where neuroscience ends and philosophy of science begins is hopeless because the boundaries are poorly defined. Neurophilosophers would pick and choose resources from both disciplines as they saw fit.

Three themes predominate Churchland's philosophical discussion: developing an alternative to the logical empiricist theory of intertheoretic reduction; responding to property-dualistic arguments based on subjectivity and sensory qualia; and responding to anti-reductionist multiple realizability arguments. These projects have remained central to neurophilosophy over the past decade. John Bickle (1998) extends the principal insight of Clifford Hooker's (1981) post-empiricist theory of intertheoretic reduction. He quantifies key notions using a model-theoretic account of theory structure adapted from the structuralist program in philosophy of science (Balzer, Moulines, and Sneed, 1987). He also makes explicit the form of argument scientists employ to draw ontological conclusions (cross-theoretic identities, revisions, or eliminations) based on the nature of the intertheoretic reduction relations obtaining in specific cases. For example, physicists concluded that visible light, a theoretical posit of optics, is electromagnetic radiation within specified wavelengths, a theoretical posit of electromagnetism: a cross-theoretic

ontological identity. In another case, however, chemists concluded that phlogiston did not exist: an elimination of a kind from our scientific ontology. Bickle explicates the nature of the reduction relation in a specific case using a semi-formal account of ‘intertheoretic approximation’ inspired by structuralist results. Paul Churchland (1996) has carried on the attack on property-dualistic arguments for the irreducibility of conscious experience and sensory qualia. He argues that acquiring some knowledge of existing sensory neuroscience increases one's ability to ‘imagine’ or ‘conceive of’ a comprehensive neurobiological explanation of consciousness. He defends this conclusion using a thought-experiment based on the history of optics and electromagnetism. Finally, the literature critical of the multiple realizability argument has begun to flourish. Although the multiple realizability argument remains influential among nonreductive physicalists, it no longer commands the universal acceptance it once did. Replies to the multiple realizability argument based on neuroscientific details have appeared. For example, William Bechtel and Jennifer Mundale (1999) argue that neuroscientists use psychological criteria in brain mapping studies. This fact undercuts the likelihood that psychological kinds are multiply realized. (For a review of recent developments see the final sections of the entry on ‘Multiple Realizability’ in this Encyclopedia, linked below.)

2. Eliminative Materialism and Philosophy Neuralized

Eliminative materialism (EM) is the conjunction of two claims. First, our common sense ‘belief-desire’ conception of mental events and processes, our ‘folk psychology,’ is a false and misleading account of the causes of human behavior. Second, like other false conceptual frameworks from both folk theory and the history of science, it will be replaced by, rather than smoothly reduced or incorporated into, a future neuroscience. According to Churchland, folk psychology is the collection of common homilies about the causes of human behavior. You ask me why Marica is not accompanying me this evening. I reply that her grant deadline is looming. You nod sympathetically. You understand my explanation because you share with me a generalization that relates beliefs about looming deadlines, desires about meeting professionally and financially significant ones, and ensuing free-time behavior. It is the collection of these kinds of homilies that EM claims to be flawed beyond significant revision. Although this example involves only beliefs and desires, folk psychology contains an extensive repertoire of propositional attitudes in its explanatory nexus: hopes, intentions, fears, imaginings, and more. To the extent that scientific psychology (and neuroscience!) retains folk concepts, EM applies to it as well.

EM is physicalist in the classical sense, postulating some future brain science as the ultimately correct account of (human) behavior. It is eliminative in predicting the future removal of folk psychological kinds from our post-neuroscientific ontology. EM proponents often employ scientific analogies (Feyerabend 1963; Churchland, 1981). Oxidative reactions as characterized within elemental chemistry bear no resemblance to phlogiston release. Even the “direction” of the two processes differ. Oxygen is gained when

an object burns (or rusts), phlogiston was said to be lost. The result of this theoretical change was the elimination of phlogiston from our scientific ontology. There is no such thing. For the same reasons, according to EM, continuing development in neuroscience will reveal that there are no such things as beliefs and desires as characterized by common sense.

Here we focus only on the way that neuroscientific results have shaped the arguments for EM. Surprisingly, only one argument has been strongly influenced. (Most arguments for EM stress the failures of folk psychology as an explanatory theory of behavior.) This argument is based on a development in cognitive and computational neuroscience that might provide a genuine alternative to the representations and computations implicit in folk psychological generalizations. Many eliminative materialists assume that folk psychology is committed to propositional representations and computations over their contents that mimic logical inferences (Paul Churchland, 1981; Stich, 1983; Patricia Churchland, 1986).[2] Even though discovering such an alternative has been an eliminativist goal for some time, neuroscience only began delivering on this goal over the past fifteen years. Points in and trajectories through vector spaces, as an interpretation of synaptic events and neural activity patterns in biological neural networks are key features of this new development. This argument for EM hinges on the differences between these notions of cognitive representation and the propositional attitudes of folk psychology (Churchland, 1987). However, this argument will be opaque to those with no background in contemporary cognitive and computational neuroscience, so we need to present a few scientific details. With these details in place, we will return to this argument for EM (five paragraphs below).

At one level of analysis the basic computational element of a neural network (biological or artificial) is the neuron. This analysis treats neurons as simple computational devices, transforming inputs into output. Both neuronal inputs and outputs reflect biological variables. For the remainder of this discussion, we will assume that neuronal inputs are frequencies of action potentials (neuronal “spikes”) in the axons whose terminal branches synapse onto the neuron in question. Neuronal output is the frequency of action potentials in the axon of the neuron in question. A neuron computes its total input (usually treated mathematically as the sum of the products of the signal strength along each input line times the synaptic weight on that line). It then computes a new activation state based on its total input and current activation state, and a new output state based on its new activation value. The neuron's output state is transmitted as a signal strength to whatever neurons its axon synapses on. The output state reflects systematically the neuron's new activation state.[3]

Analyzed at this level, both biological and artificial neural networks are interpreted naturally as*vector-to-vector transformers*. The input vector consists of values reflecting activity patterns in axons synapsing on the network's neurons from outside (e.g., from sensory transducers or other neural networks). The output vector consists of values reflecting the activity patterns generated in the network's neurons that project beyond the net (e.g., to motor effectors or other neural networks). Given that neurons' activity depends partly upon their total input, and total input depends partly on synaptic weights (e.g., presynaptic neurotransmitter release rate, number and efficacy of postsynaptic receptors,

availability of enzymes in synaptic cleft), the capacity of biological networks to change their synaptic weights make them *plastic* vector-to-vector transformers. In principle, a biological network with plastic synapses can come to implement any vector-to-vector transformation that its composition permits (number of input units, output units, processing layers, recurrency, cross-connections, etc.) (Churchland, 1987).

The anatomical organization of the cerebellum provides a clear example of a network amendable to this computational interpretation. Consider Figure 1. The cerebellum is the bulbous convoluted structure dorsal to the brainstem. A variety of studies (behavioral, neuropsychological, single-cell electrophysiological) implicate this structure in motor integration and fine motor coordination. Mossy fibers (axons) from neurons outside the cerebellum synapse on cerebellar granule cells, which in turn project to parallel fibers. Activity patterns across the collection of mossy fibers (frequency of action potentials per time unit in each fiber projecting into the cerebellum) provide values for the input vector. Parallel fibers make multiple synapses on the dendritic trees and cell bodies of cerebellular Purkinje neurons. Each Purkinje neuron "sums" its post-synaptic potentials (PSPs) and emits a train of action potentials down its axon based (partly) on its total input and previous activation state. Purkinje axons project outside the cerebellum. The network's output vector is thus the ordered values representing the pattern of activity generated in each Purkinje axon. Changes to the efficacy of individual synapses on the parallel fibers and the Purkinje neurons alter the resulting PSPs in Purkinje axons, generating different axonal spiking frequencies. Computationally, this amounts to a different output vector to the same input activity pattern (plasticity).[4]

This interpretation puts the useful mathematical resources of *dynamical systems* into the hands of computational neuroscientists. *Vector spaces* are an example. For example, learning can be characterized fruitfully in terms of changes in synaptic weights in the network and subsequent reduction of error in network output. (This approach goes back to Hebb, 1949, although within the vector-space interpretation that follows.) A useful representation of this account is on a *synaptic weight-error space*, where one dimension represents the global error in the network's output to a given task, and all other dimensions represent the weight values of individual synapses in the network. Consider Figure 2. Points in this multi-dimensional state space represent the global performance error correlated with each possible collection of synaptic weights in the network. As the weights change with each performance (in accordance with a biologically-implemented learning algorithm), the global error of network performance continually decreases. Learning is represented as synaptic weight changes correlated with a descent along the error dimension in the space (Churchland and Sejnowski, 1992). Representations (concepts) can be portrayed as *partitions* in multi-dimensional vector spaces. An example is a *neuron activation* vector space. See Figure 3. A graph of such a space contains one dimension for the activation value of each neuron in the network (or some subset). A point in this space represents one possible pattern of activity in all neurons in the network. Activity patterns generated by input vectors that the network has learned to group together will cluster around a (hyper-) point or subvolume in the activity vector space. Any input pattern sufficiently similar to this group will produce an activity pattern lying in geometrical proximity to this point or

subvolume. Paul Churchland (1989) has argued that this interpretation of network activity provides a quantitative, neurally-inspired basis for prototype theories of concepts developed recently in cognitive psychology.

Using this theoretical development, Paul Churchland (1987, 1989) has offered a novel argument for EM. According to this approach, activity vectors are the central kind of representation and vector-to-vector transformations are the central kind of computation in the brain. This contrasts sharply with the *propositional* representations and *logical/semantic* computations postulated by folk psychology. Vectorial content is unfamiliar and alien to common sense. This cross-theoretic difference is at least as great as that between oxidative and phlogiston concepts, or kinetic-corpuscular and caloric fluid heat concepts. Phlogiston and caloric fluid are two "parade" examples of kinds eliminated from our scientific ontology due to the nature of the intertheoretic relation obtaining between the theories with which they are affiliated and the theories that replaced these. The structural and dynamic differences between the folk psychological and emerging cognitive neuroscientific kinds suggest that the theories affiliated with the latter will also correct significantly the theory affiliated with the former. This is the key premise of an eliminativist argument based on predicted intertheoretic relations. And these intertheoretic contrasts are no longer just an eliminativist's goal. Computational and cognitive neuroscience has begun to deliver an alternative kinematics for cognition, one that provides no structural analogue for the propositional attitudes.

Certainly the replacement of propositional contents by vectorial alternatives implies significant correction to folk psychology. But does it justify EM? Even though this central feature of folk-psychological posits finds no analogues in one hot theoretical development in recent cognitive and computational neuroscience, there might be other aspects of cognition that folk psychology gets right. Within neurophilosophy, concluding that a cross-theoretic identity claim is true (e.g., folk psychological state F is identical to neural state N) or that an eliminativist claim is true (there is no such thing as folk psychological state F) depends on the nature of the intertheoretic reduction obtaining between the theories affiliated with the posits in question (Hooker, 1981; Churchland, 1986; Bickle, 1998). But the underlying account of intertheoretic reduction recognizes a spectrum of possible reductions, ranging from relatively "smooth" through "significantly revisionary" to "extremely bumpy".[5] Might the reduction of folk psychology and a "vectorial" neurobiology occupy the middle ground between "smooth" and "bumpy" intertheoretic reductions, and hence suggest a "revisionary" conclusion? The reduction of classical equilibrium thermodynamics to statistical mechanics to microphysics provides a potential analogy. John Bickle (1992, 1998, chapter 6) argues on empirical grounds that such a outcome is likely. He specifies conditions on "revisionary" reductions from historical examples and suggests that these conditions are obtaining between folk psychology and cognitive neuroscience as the latter develops. In particular, folk psychology appears to have gotten right the grossly-specified functional profile of many cognitive states, especially those closely related to sensory input and behavioral output. It also appears to get right the "intentionality" of many cognitive states—the object that the state is of or about—even though cognitive neuroscience eschews its implicit linguistic explanation of this feature.

Revisionary physicalism predicts significant *conceptual change* to folk psychological concepts, but denies total elimination of the caloric fluid-phlogiston variety.

The philosophy of science is another area where vector space interpretations of neural network activity patterns has impacted philosophy. In the Introduction to his (1989) book, *A Neurocomputational Perspective*, Paul Churchland asserts that it will soon be impossible to do serious work in the philosophy of science without drawing on empirical work in the brain and behavioral sciences. To justify this claim, in Part II of the book he suggests neurocomputational reformulations of key concepts from this area. At the heart is a neurocomputational account of the structure of scientific theories (1989, chapter 9). Problems with the orthodox "sets-of-sentences" view have been well-known for over three decades. Churchland advocates replacing the orthodox view with one inspired by the "vectorial" interpretation of neural network activity. Representations implemented in neural networks (as discussed above) compose a system that corresponds to important distinctions in the external environment, are not explicitly represented as such within the input corpus, and allow the trained network to respond to inputs in a fashion that continually reduces error. These are exactly the functions of theories. Churchland is bold in his assertion: an individual's theory-of-the-world is a specific point in that individual's error-synaptic weight vector space. It is a configuration of synaptic weights that partitions the individual's activation vector space into subdivisions that reduce future error messages to both familiar and novel inputs. (Consider again Figure 2 and Figure 3.) This reformulation invites an objection, however. Churchland boasts that his theory of theories is preferable to existing alternatives to the orthodox "sets-of-sentences" account—for example, the *semantic* view (Suppe, 1974; van Fraassen, 1980)—because his is closer to the "buzzing brains" that use theories. But as Bickle (1993) notes, neurocomputational models based on the mathematical resources described above are a long way into the realm of abstractia. Even now, they remain little more than novel (and suggestive) applications of the mathematics of quasi-linear dynamical systems to simplified schemata of brain circuitries. Neurophilosophers owe some account of identifications across ontological categories before the philosophy of science community will accept the claim that theories are points in high-dimensional state spaces implemented in biological neural networks. (There is an important methodological assumption lurking in this objection, however, which we will discuss toward the end of the next paragraph.)

Churchland's neurocomputational reformulations of scientific and epistemological concepts build on this account of theories. He sketches "neuralized" accounts of the theory-ladenness of perception, the nature of concept unification, the virtues of theoretical simplicity, the nature of Kuhnian paradigms, the kinematics of conceptual change, the character of abduction, the nature of explanation, and even moral knowledge and epistemological normativity. Conceptual redeployment, for example, is the activation of an already-existing prototype representation—the centerpoint or region of a partition of a high-dimensional vector space in a trained neural network—to a novel type of input pattern. Obviously, we can't here do justice to Churchland's many and varied attempts at reformulation. We urge the intrigued reader to examine his suggestions in their original form. But a word about philosophical methodology is in order. Churchland is *not*attempting

“conceptual analysis” in anything resembling its traditional philosophical sense and neither, typically, are neurophilosophers.(This is why a discussion of neurophilosophical reformulations fits with a discussion of EM.) There are philosophers who take the discipline's ideal to be a relatively simple set of necessary and sufficient conditions, expressed in non-technical natural language, governing the application of important concepts (like justice, knowledge, theory, or explanation). These analyses should square, to the extent possible, with pretheoretical usage. Ideally, they should preserve synonymy. Other philosophers view this ideal as sterile, misguided, and perhaps deeply mistaken about the underlying structure of human knowledge (Ramsey, 1992). Neurophilosophers tend to reside in the latter camp. Those who dislike philosophical speculation about the promise and potential of nascent science in an effort to reformulate (“*reform*-ulate”) traditional philosophical concepts have probably already discovered that neurophilosophy is not for them. But the charge that neurocomputational reformulations of the sort Churchland attempts are “philosophically uninteresting” or “irrelevant” because they fail to provide “adequate analyses” of theory, explanation, and the like will fall on deaf ears among many contemporary philosophers, as well as their cognitive-scientific and neuroscientific friends.

Before we leave the neurophilosophical applications of this theoretical development from recent cognitive/computational neuroscience, one more point of scientific detail is in order. Many *neural*modelers no longer treat the neuron as the basic computational unit in the brain. *Compartmental modeling* enables computational neuroscientists to mimic activity in and interactions between patches of neuronal membrane (Bower and Beeman, 1995). This permits modelers to control and manipulate a variety of subcellular factors that determine action potentials per time unit (including the topology of membrane structure in individual neurons, variations in ion channels across membrane patches, field properties of post-synaptic potentials depending on the location of the synapse on the dendrite or soma). Modelers can “custom build” the neurons in their target circuitry without sacrificing the ability to study circuit properties of networks. For these reasons, many serious computational *neuro*scientists now work at a level that treats neurons as structured computational devices. With compartmental modeling, not only are simulated neural networks interpretable as vector-to-vector transformers. The neurons composing them are, too.

Philosophy of science and scientific epistemology are not the only areas where philosophers have lately urged the relevance of neuroscientific discoveries. Kathleen Akins (1996) argues that a “traditional” view of the senses underlies the variety of sophisticated “naturalistic” programs about intentionality. (She cites the Churchlands, Daniel Dennett, Fred Dretske, Jerry Fodor, David Papineau, Dennis Stampe, and Kim Sterelny as examples, with extensive references.) Current neuroscientific understanding of the mechanisms and coding strategies implemented by sensory receptors shows that this traditional view is mistaken. The traditional view holds that sensory systems are “veridical” in at least three ways. (1) Each signal in the system correlates with a small range of properties in the external (to the body) environment. (2) The structure in the relevant relations between the external properties the receptors are sensitive to is preserved in the structure of the relations between the resulting sensory states. And (3) the sensory system reconstructs faithfully,

without fictive additions or embellishments, the external events. Using recent neurobiological discoveries about response properties of thermal receptors in the skin as an illustration, Akins shows that sensory systems are "narcissistic" rather than "veridical." All three traditional assumptions are violated. These neurobiological details and their philosophical implications open novel questions for the philosophy of perception and for the appropriate foundations for naturalistic projects about intentionality. Armed with the known neurophysiology of sensory receptors, for example, our "philosophy of perception" or of "perceptual intentionality" will no longer focus on the search for correlations between states of sensory systems and "veridically detected" external properties. This traditional philosophical (and scientific!) project rests upon a mistaken "veridical" view of the senses. Neuroscientific knowledge of sensory receptor activity also shows that sensory experience does not serve the naturalist well as a "simple paradigm case" of an intentional relation between representation and world. Once again, available scientific detail shows the naivity of some traditional philosophical projects.

Focusing on the anatomy and physiology of the pain transmission system, Valerie Hardcastle (1997) urges a similar negative implication for a popular methodological assumption. Pain experiences have long been philosophers' favorite cases for analysis and theorizing about conscious experience generally. Nevertheless, every position about pain experiences has been defended recently: eliminativism, a variety of objectivist views, relational views, and subjectivist views. Why so little agreement, despite agreement that pain experiences are the place to start an analysis or theory of consciousness? Hardcastle urges two answers. First, philosophers tend to be uninformed about the neuronal complexity of our pain transmission systems, and build their analyses or theories on the outcome of a single component of a multi-component system. Second, even those who understand some of the underlying neurobiology of pain tend to advocate gate-control theories.[6]But the best existing gate-control theories are vague about the neural mechanisms of the gates. Hardcastle instead proposes a dissociable dual system of pain transmission, consisting of a pain sensory system closely analogous in its neurobiological implementation to other sensory systems, and a descending pain inhibitory system. She argues that this dual system is consistent with recent neuroscientific discoveries and accounts for all the pain phenomena that have tempted philosophers toward particular (but limited) theories of pain experience. The neurobiological uniqueness of the pain inhibitory system, contrasted with the mechanisms of other sensory modalities, renders pain processing atypical. In particular, the pain inhibitory system dissociates pain sensation from stimulation of nociceptors (pain receptors). Hardcastle concludes from the neurobiological uniqueness of pain transmission that pain experiences are atypical conscious events, and hence not a good place to start theorizing about or analyzing the general type.

3. Neuroscience and Psychosemantics

Developing and defending theories of content is a central topic in current philosophy of mind. A common desideratum in this debate is a theory of cognitive representation

consistent with a physical or naturalistic ontology. We'll here describe a few contributions neurophilosophers have made to this literature.

When one perceives or remembers that he is out of coffee, his brain state possesses intentionality or "aboutness." The percept or memory is about one's being out of coffee; it represents one as being out of coffee. The representational state has content. A psychosemantics seeks to explain what it is for a representational state to be about something: to provide an account of how states and events can have specific representational content. A physicalist psychosemantics seeks to do this using resources of the physical sciences exclusively. Neurophilosophers have contributed to two types of physicalist psychosemantics: the Functional Role approach and the Informational approach. For a description of these and other theories of mental content, see the entries on causal theories of mental content, mental representation, and teleological theories of mental content.

The core claim of a functional role semantics holds that a representation has its content in virtue of relations it bears to other representations. Its paradigm application is to concepts of truth-functional logic, like the conjunctive 'and' or disjunctive 'or.' A physical event instantiates the 'and' function just in case it maps two true inputs onto a single true output. Thus it is the relations an expression bears to others that give it the semantic content of 'and.' Proponents of functional role semantics propose similar analyses for the content of all representations (Block 1986). A physical event represents birds, for example, if it bears the right relations to events representing feathers and others representing beaks. By contrast, informational semantics ascribe content to a state depending upon the causal relations obtaining between the state and the object it represents. A physical state represents birds, for example, just in case an appropriate causal relation obtains between it and birds. At the heart of informational semantics is a causal account of information (Dretske, 1981, 1988). Red spots on a face carry the information that one has measles because the red spots are caused by the measles virus. A common criticism of informational semantics holds that mere causal covariation is insufficient for representation, since information (in the causal sense) is by definition always veridical while representations can misrepresent. A popular solution to this challenge invokes a teleological analysis of 'function.' A brain state represents *X* by virtue of having the function of carrying information about being caused by *X* (Dretske 1988). These two approaches do not exhaust the popular options for a psychosemantics, but are the ones to which neurophilosophers have contributed.

Paul Churchland's allegiance to functional role semantics goes back to his earliest views about the semantics of terms in a language. In his (1979) book, he insists that the semantic identity (content) of a term derives from its place in the network of sentences of the entire language. The functional economies envisioned by early functional role semanticists were networks with nodes corresponding to the objects and properties denoted by expressions in a language. Thus one node, appropriately connected, might represent birds, another feathers, and another beaks. Activation of one of these would tend to spread to the others. As 'connectionist' network modeling developed, alternatives arose to this one-representation-per-node 'localist' approach. By the time Churchland (1989) provided a

neuroscientific elaboration of functional role semantics for cognitive representations generally, he too had abandoned the 'localist' interpretation. Instead, he offered a 'state-space semantics'.

We saw in the section just above how (vector) state spaces provide a natural interpretation for activity patterns in neural networks (biological and artificial). A state-space semantics for cognitive representations is a species of a functional role semantics because the individuation of a particular state depends upon the relations obtaining between it and other states. A representation is a point in an appropriate state space, and points (or subvolumes) in a space are individuated by their relations to other points (locations, geometrical proximity). Churchland (1989, 1995) illustrates a state-space semantics for neural states by appealing to sensory systems. One popular theory in sensory neuroscience of how the brain codes for sensory qualities (like color) is the *opponent process account* (Hardin 1988). Churchland (1995) describes a three-dimensional activation vector state-space in which every color perceivable by humans is represented as a point (or subvolume). Each dimension corresponds to activity rates in one of three classes of photoreceptors present in the human retina and their efferent paths: the red-green opponent pathway, yellow-blue opponent pathway, and black-white (contrast) opponent pathway. Photons striking the retina are transduced by the receptors, producing an activity rate in each of the segregated pathways. A represented color is hence a triplet of activation frequency rates. As an illustration, consider again Figure 3. Each dimension in that three-dimensional space will represent average frequency of action potentials in the axons of one class of ganglion cells projecting out of the retina. Each color perceivable by humans will be a region of that space. For example, an orange stimulus produces a relatively low level of activity in both the red-green and yellow-blue opponent pathways (*x*-axis and *y*-axis, respectively), and middle-range activity in the black-white (contrast) opponent pathway (*z*-axis). Pink stimuli, on the other hand, produce low activity in the red-green opponent pathway, middle-range activity in the yellow-blue opponent pathway, and high activity in the black-white (contrast) opponent pathway.[7] The location of each color in the space generates a 'color solid.' Location on the solid and geometrical proximity between regions reflect structural similarities between the perceived colors. Human gustatory representations are points in a four-dimensional state space, with each dimension coding for activity rates generated by gustatory stimuli in each type of taste receptor (sweet, salty, sour, bitter) and their segregated efferent pathways. When implemented in a neural network with structural and hence computational resources as vast as the human brain, the state space approach to psychosemantics generates a theory of content for a huge number of cognitive states.[8]

Jerry Fodor and Ernest LePore (1992) raise an important challenge to Churchland's psychosemantics. Location in a state space alone seems insufficient to fix a state's representational content. Churchland never explains why a point in a three-dimensional state space represents a *color*, as opposed to any other quality, object, or event that varies along three dimensions.[9]. Churchland's account achieves its explanatory power by the interpretation imposed on the dimensions. Fodor and LePore allege that Churchland never specifies how a dimension comes to represent, e.g., degree of saltiness, as opposed to yellow-blue wavelength opposition. One obvious answer appeals to the stimuli that form

the ‘external’ inputs to the neural network in question. Then, for example, the individuating conditions on neural representations of colors are that opponent processing neurons receive input from a specific class of photoreceptors. The latter in turn have electromagnetic radiation (of a specific portion of the visible spectrum) as their activating stimuli. However, this appeal to ‘external’ stimuli as the ultimate individuating conditions for representational content makes the resulting approach a version of informational semantics. Is this approach consonant with other neurobiological details?

The neurobiological paradigm for informational semantics is the *feature detector*: one or more neurons that are (i) maximally responsive to a particular type of stimulus, and (ii) have the function of indicating the presence of that stimulus type. Examples of such stimulus-types for visual feature detectors include high-contrast edges, motion direction, and colors. A favorite feature detector among philosophers is the alleged fly detector in the frog. Lettvin *et al*. (1959) identified cells in the frog retina that responded maximally to small shapes moving across the visual field. The idea that these cells' activity functioned to detect flies rested upon knowledge of the frogs' diet. (Bechtel 1998 provides a useful discussion.) Using experimental techniques ranging from single-cell recording to sophisticated functional imaging, neuroscientists have recently discovered a host of neurons that are maximally responsive to a variety of stimuli. However, establishing condition (ii) on a feature detector is much more difficult. Even some paradigm examples have been called into question. David Hubel and Torsten Wiesel's (1962) Nobel Prize-winning work establishing the receptive fields of neurons in striate cortex is often interpreted as revealing cells whose function is edge detection. However, Lehky and Sejnowski (1988) have challenged this interpretation. They trained an artificial neural network to distinguish the three-dimensional shape and orientation of an object from its two-dimensional shading pattern. Their network incorporates many features of visual neurophysiology. Nodes in the trained network turned out to be maximally responsive to edge contrasts, but did not appear to have the function of edge detection. (See Churchland and Sejnowski 1992 for a review.)

Kathleen Akins (1996) offers a different neurophilosophical challenge to informational semantics and its affiliated feature-detection view of sensory representation. We saw in the previous section how Akins argues that the physiology of thermoreception violates three necessary conditions on ‘veridical’ representation. From this fact she draws doubts about looking for feature detecting neurons to ground a psychosemantics generally, including thought contents. Human thoughts about flies, for example, are sensitive to numerical distinctions between particular flies and the particular locations they can occupy. But the ends of frog nutrition are well served without a representational system sensitive to such ontological refinements. Whether a fly seen now is numerically identical to one seen a moment ago need not, and perhaps cannot, figure into the frog's feature detection repertoire. Akins' critique casts doubt on whether details of sensory transduction will scale up to provide an adequate unified psychosemantics. It also raises new questions for human intentionality. How do we get from activity patterns in “narcissistic” sensory receptors, keyed not to “objective” environmental features but rather only to effects of the stimuli on the patch of tissue innervated, to the human ontology replete with enduring objects with

stable configurations of properties and relations, types and their tokens (as the "fly-thought" example presented above reveals), and the rest? And how did the development of a stable, rich ontology confer survival advantages to human ancestors?

4. Consciousness Explained?

Consciousness has re-emerged as a topic in philosophy of mind and the cognitive and brain sciences over the past three decades. Instead of ignoring it, many physicalists now seek to explain it (Dennett, 1991). Here we focus exclusively on ways that neuroscientific discoveries have impacted philosophical debates about the nature of consciousness and its relation to physical mechanisms. (See links to other entries in this encyclopedia below for broader discussions about consciousness and physicalism.)

Thomas Nagel (1974) argues that conscious experience is subjective, and thus permanently recalcitrant to objective scientific understanding. He invites us to ponder 'what it is like to be a bat' and urges the intuition that no amount of physical-scientific knowledge (including neuroscientific) supplies a complete answer. Nagel's intuition pump has generated extensive philosophical discussion. At least two well-known replies make direct appeal to neurophysiology. John Biro (1991) suggests that part of the intuition pumped by Nagel, that bat experience is substantially different from human experience, presupposes systematic relations between physiology and phenomenology. Kathleen Akins (1993a) delves deeper into existing knowledge of bat physiology and reports much that is pertinent to Nagel's question. She argues that many of the questions about bat subjectivity that we still consider open hinge on questions that remain unanswered about neuroscientific details. One example of the latter is the function of various cortical activity profiles in the active bat.

More recently philosopher David Chalmers (1996) has argued that any possible brain-process account of consciousness will leave open an 'explanatory gap' between the brain process and properties of the conscious experience.[10] This is because no brain-process theory can answer the "hard" question: Why should that particular brain process give rise to conscious experience? We can always imagine ("conceive of") a universe populated by creatures having those brain processes but completely lacking conscious experience. A theory of consciousness requires an explanation of how and why some brain process causes consciousness replete with all the features we commonly experience. The fact that the hard question remains unanswered shows that we will probably never get a complete explanation of consciousness at the level of neural mechanism. Paul and Patricia Churchland (1997) have recently offered the following diagnosis and reply. Chalmers offers a*conceptual argument*, based on our ability to imagine creatures possessing brains like ours but wholly lacking in conscious experience. But the more one learns about how the brain produces conscious experience—and a literature is beginning to emerge (e.g., Gazzaniga, 1995)—the harder it becomes to imagine a universe consisting of creatures with brain processes like ours but lacking consciousness. This is not just bare assertion. The Churchlands appeal to some neurobiological detail. For example, Paul Churchland (1995) develops a

neuroscientific account of consciousness based on recurrent connections between thalamic nuclei (particularly “diffusely projecting” nuclei like the intralaminar nuclei) and cortex.[11] Churchland argues that the thalamocortical recurrency accounts for the selective features of consciousness, for the effects of short-term memory on conscious experience, for vivid dreaming during REM (rapid-eye movement) sleep, and other “core” features of conscious experience. In other words, the Churchlands are claiming that when one learns about activity patterns in these recurrent circuits, one can't “imagine” or “conceive of” this activity occurring without these core features of conscious experience. (Other than just mouthing the words, “I am now imagining activity in these circuits without selective attention/the effects of short-term memory/vivid dreaming/...”).

A second focus of skeptical arguments about a complete neuroscientific explanation of consciousness is sensory *qualia*: the introspectable qualitative aspects of sensory experience, the features by which subjects discern similarities and differences among their experiences. The colors of visual sensations are a philosopher's favorite example. One famous puzzle about color qualia is the alleged conceivability of spectral inversions. Many philosophers claim that it is conceptually possible (if perhaps physically impossible) for two humans not to differ neurophysiologically, while the color that fire engines and tomatoes appear to have to one subject is the color that grass and frogs appear to have to the other (and vice versa). A large amount of neuroscientifically-informed philosophy has addressed this question. (C.L. Hardin 1988 and Austen Clark 1993 are noteworthy examples.) A related area where neurophilosophical considerations have emerged concerns the metaphysics of colors themselves (rather than color experiences). A longstanding philosophical dispute is whether colors are objective properties existing external to perceivers or rather identifiable as or dependent upon minds or nervous systems. Some recent work on this problem begins with characteristics of color experiences: for example, that color similarity judgments produce color orderings that align on a circle (Clark 1993). With this resource, one can seek mappings of phenomenology onto environmental or physiological regularities. Identifying colors with particular frequencies of electromagnetic radiation does not preserve the structure of the hue circle, whereas identifying colors with activity in opponent processing neurons does. Such a tidbit is not decisive for the color objectivist-subjectivist debate, but it does convey the type of neurophilosophical work being done on traditional metaphysical issues beyond the philosophy of mind. (For more details on these issues, see the entry on Color in this Encyclopedia, linked below.)

We saw in the discussion of Hardcastle (1997) two sections above that neurophilosophers have entered disputes about the nature and methodological import of pain experiences. Two decades earlier, Dan Dennett (1978) took up the question of whether it is possible to build a computer that feels pain. He compares and notes tension between neurophysiological discoveries and common sense intuitions about pain experience. He suspects that the incommensurability between scientific and common sense views is due to incoherence in the latter. His attitude is wait-and-see. But foreshadowing Churchland's reply to Chalmers, Dennett favors scientific investigations over conceivability-based philosophical arguments.

Neurological deficits have attracted philosophical interest. For thirty years philosophers have found implications for the unity of the self in experiments with commissurotomy patients (Nagel 1971).[12]In carefully controlled experiments, commissurotomy patients display two dissociable seats of consciousness. In chapter 5 of her (1986) book, Patricia Churchland scouts philosophical implications of a variety of neurological deficits. One deficit is blindsight. Some patients with lesions to primary visual cortex report being unable to see items in regions of their visual fields, yet perform far better than chance in forced guess trials about stimuli in those regions. A variety of scientific and philosophical interpretations have been offered. Ned Block (1988) worries that many of these conflate distinct notions of consciousness. He labels these notions 'phenomenal consciousness' ('P-consciousness') and 'access consciousness' ('A-consciousness'). The former is the 'what it is like'-ness of experience. The latter is the availability of representational content to self-initiated action and speech. Block argues that P-consciousness is not always representational whereas A-consciousness is. Dennett (1991, 1995) and Michael Tye (1993) are skeptical of non-representational analyses of consciousness in general. They provide accounts of blindsight that do not depend on Block's distinction.

We break off our brief overview of neurophilosophical work on consciousness here. Many other topics are worth neurophilosophical pursuit. We mentioned commissurotomy and the unity of consciousness and the self, which continues to generate discussion. Qualia beyond those of color and pain have begun to attract neurophilosophical attention (Akins 1993a, 1993b, 1996; Clark 1993), as has self-consciousness (Bermudez 1998).

5. Location of Cognitive Function: From Lesion Studies to Recent Neuroimaging

One of the first issues to arise in the 'philosophy of neuroscience' (before there was a recognized area) was the localization of cognitive functions to specific neural regions. Although the 'localization' approach had dubious origins in the phrenology of Gall and Spurzheim, and was challenged severely by Flourens throughout the early nineteenth century, it re-emerged in the study of aphasia by Bouillaud, Auburtin, Broca, and Wernicke. These neurologists made careful studies (where possible) of linguistic deficits in their aphasic patients followed by brain autopsies post mortem.[13] Broca's initial study of twenty-two patients in the mid-nineteenth century confirmed that damage to the *left cortical hemisphere* was predominant, and that damage to the second and third frontal convolutions was necessary to produce speech production deficits. Although the anatomical coordinates Broca postulated for the 'speech production center' do not correlate exactly with damage producing production deficits, both this area of frontal cortex and speech production deficits still bear his name ('Broca's area' and 'Broca's aphasia'). Less than two decades later Carl Wernicke published evidence for a second language center. This area is anatomically distinct from Broca's area, and damage to it produced a very different set of aphasic symptoms. The cortical area that still bears his name ('Wernicke's area') is located around the first and second convolutions in temporal cortex, and the aphasia that bears his

name (‘Wernicke's aphasia’) involves deficits in language comprehension. Wernicke's method, like Broca's, was based on lesion studies: a careful evaluation of the behavioral deficits followed by post mortem examination to find the sites of tissue damage and atrophy. Lesion studies suggesting more precise localization of specific linguistic functions remain a cornerstone to this day in aphasic research.

Lesion studies have also produced evidence for the localization of other cognitive functions: for example, sensory processing and certain types of learning and memory. However, localization arguments for these other functions invariably include studies using animal models. With an animal model, one can perform careful behavioral measures in highly controlled settings, then ablate specific areas of neural tissue (or use a variety of other techniques to block or enhance activity in these areas) and remeasure performance on the same behavioral tests. But since we lack an animal model for (human) language production and comprehension, this additional evidence isn't available to the neurologist or neurolinguist. This fact makes the study of language a paradigm case for evaluating the logic of the lesion/deficit method of inferring functional localization. Philosopher Barbara Von Eckardt (1978) attempts to make explicit the steps of reasoning involved in this common and historically important method. Her analysis begins with Robert Cummins' early analysis of functional explanation, but she extends it into a notion of *structurally adequate*functional analysis. These analyses break down a complex capacity C into its constituent capacitiesc_1, c_2,…, c_n, where the constituent capacities are consistent with the underlying structural details of the system. For example, human speech production (complex capacity C) results from formulating a speech intention, then selecting appropriate linguistic representations to capture the content of the speech intention, then formulating the motor commands to produce the appropriate sounds, then communicating these motor commands to the appropriate motor pathways (constituent capacitiesc_1, c_2,…, c_n). A functional-localization hypothesis has the form: brain structure S in organism (type) O has constituent capacity c_i, where c_i is a function of some part of O. An example might be: Broca's area (S) in humans (O) formulates motor commands to produce the appropriate sounds (one of the constituent capacities c_i). Such hypotheses specify aspects of the structural realization of a functional-component model. They are part of the theory of the neural realization of the functional model.

Armed with these characterizations, Von Eckardt argues that inference to a functional-localization hypothesis proceeds in two steps. First, a functional deficit in a patient is hypothesized based on the abnormal behavior the patient exhibits. Second, localization of function in normal brains is inferred on the basis of the functional deficit hypothesis plus the evidence about the site of brain damage. The structurally-adequate functional analysis of the capacity connects the pathological behavior to the hypothesized functional deficit. This connection suggests four adequacy conditions on a functional deficit hypothesis. First, the pathological behavior P (e.g., the speech deficits characteristic of Broca's aphasia) must result from failing to exercise some complex capacity C (human speech production). Second, there must be a structurally-adequate functional analysis of how people exercise capacity C that involves some constituent capacity c_i (formulating motor commands to produce the appropriate sounds). Third, the operation of the steps described by the

structurally-adequate functional analysis minus the operation of the component performing c_i(Broca's area) must result in pathological behavior P. Fourth, there must not be a better available explanation for why the patient does P. Arguments to a functional deficit hypothesis on the basis of pathological behavior is thus an instance of argument to the best available explanation. When postulating a deficit in a normal functional component provides the best available explanation of the pathological data, we are justified in drawing the inference.

Von Eckardt applies this analysis to a neurological case study involving a controversial reinterpretation of agnosia.[14] Her philosophical explication of this important neurological method reveals that most challenges to localization arguments either argue only against the localization of a particular type of functional capacity or against generalizing from localization of function in one individual to all normal individuals. (She presents examples of each from the neurological literature.) Such challenges do not impugn the validity of standard arguments for functional localization from deficits. It does not follow that such arguments are unproblematic. But they face difficult factual and methodological problems, not logical ones. Furthermore, the analysis of these arguments as involving a type of functional analysis and inference to the best available explanation carries an important implication for the biological study of cognitive function. Functional analyses require functional theories, and structurally adequate functional analyses require checks imposed by the lower level sciences investigating the underlying physical mechanisms. Arguments to best available explanation are often hampered by a lack of theoretical imagination: the available explanations are often severely limited. We must seek theoretical inspiration from any level of theory and explanation. Hence making explicit the ‘logic’ of this common and historically important form of neurological explanation reveals the necessity of joint participation from all scientific levels, from cognitive psychology down to molecular neuroscience. Von Eckardt (1978) anticipated what came to be heralded as the ‘co-evolutionary research methodology,’ which remains a centerpiece of neurophilosophy to the present day.

Over the last two decades, evidence for localization of cognitive function has come increasingly from a new source: the development and refinement of neuroimaging techniques. The form of localization-of-function argument appears not to have changed from that employing lesion studies (as analyzed by Von Eckardt). Instead, these imaging technologies resolve some of the methodological problems that plague lesion studies. For example, researchers do not need to wait until the patient dies, and in the meantime probably acquires additional brain damage, to find the lesion sites. Two functional imaging techniques are prominent: positron emission tomography, or PET, and functional magnetic resonance imaging, or fMRI. Although these measure different biological markers of functional activity, both now have a resolution down to around 1mm.[15] As these techniques increase spatial and temporal resolution of functional markers and continue to be used with sophisticated behavioral methodologies, the possibility of localizing specific psychological functions to increasingly specific neural regions continues to grow.[16]

6. A Result of the Co-evolutionary Research Ideology: Cognitive and Computational Neuroscience

What we now know about the cellular and molecular mechanisms of neural conductance and transmission is spectacular. (For those in doubt, simply peruse for five minutes a recent volume of*Society for Neuroscience Abstracts*.) The same evaluation holds for all levels of explanation and theory about the mind/brain: maps, networks, systems, and behavior. This is a natural outcome of increasing scientific specialization. We develop the technology, the experimental techniques, and the theoretical frameworks within specific disciplines to push forward our understanding. Still, a crucial aspect of the total picture gets neglected: the relationship between the levels, the 'glue' that binds knowledge of neuron activity to subcellular and molecular mechanisms, network activity patterns to the activity of and connectivity between single neurons, and behavior to network activity. This problem is especially glaring when we focus on the relationship between 'cognitivist' psychological theories, postulating information-bearing representations and processes operating over their contents, and the activity patterns in networks of neurons. Co-evolution between explanatory levels still seems more like a distant dream rather than an operative methodology.

It is here that some neuroscientists appeal to 'computational' methods (Churchland and Sejnowski 1992). If we examine the way that computational models function in more developed sciences (like physics), we find the resources of *dynamical systems* constantly employed. Global effects (such as large-scale meteorological patterns) are explained in terms of the interaction of 'local' lower-level physical phenomena, but only by dynamical, nonlinear, and often chaotic sequences and combinations. Addressing the interlocking levels of theory and explanation in the mind/brain using computational resources that have worked to bridge levels in more mature sciences might yield comparable results. This methodology is necessarily interdisciplinary, drawing on resources and researchers from a variety of levels, including higher levels like experimental psychology, 'program-writing' and 'connectionist' artificial intelligence, and philosophy of science.

However, the use of computational methods in neuroscience is not new. Hodgkin, Huxley, and Katz (1952) incorporated values of voltage-dependent potassium conductance they had measured experimentally in the squid giant axon into an equation from physics describing the time evolution of a first-order kinetic process. This equation enabled them to calculate best-fit curves for modeled conductance versus time data that reproduced the S-shaped (sigmoidal) function suggested by their experimental data. Using equations borrowed from physics, Rall (1959) developed the cable model of dendrites. This theory provided an account of how the various inputs from across the dendritic tree interact temporally and spatially to determine the input-output properties of single neurons. It remains influential today, and has been incorporated into the GENESIS software for programming neurally realistic networks (Bower and Beeman 1995). More recently, David Sparks and his colleagues have shown that a vector-averaging model of activity in neurons of superior

colliculi correctly predicts experimental results about the amplitude and direction of saccadic eye movements (Lee, Rohrer, and Sparks 1988). Working with a more sophisticated mathematical model, Apostolos Georgopoulos and his colleagues have predicted direction and amplitude of hand and arm movements based on averaged activity of 224 cells in motor cortex. Their predictions have borne out under a variety of experimental tests (Georgopoulos *et al*. 1986). We mention these particular studies only because we are familiar with them. We could multiply examples of the fruitful interaction of computational and experimental methods in neuroscience easily by one-hundred-fold. Many of these extend back before 'computational neuroscience' was a recognized research endeavor.

We've already seen one example, the vector transformation account, of neural representation and computation, under active development in cognitive neuroscience. Other approaches using 'cognitivist' resources are also being pursued.[17] Many of these projects draw upon 'cognitivist' characterizations of the phenomena to be explained. Many exploit 'cognitivist' experimental techniques and methodologies. Some even attempt to derive 'cognitivist' explanations from cell-biological processes (e.g., Hawkins and Kandel 1984). As Stephen Kosslyn (1997) puts it, cognitive neuroscientists employ the 'information processing' view of the mind characteristic of cognitivism without trying to separate it from theories of brain mechanisms. Such an endeavor calls for an interdisciplinary community willing to communicate the relevant portions of the mountain of detail gathered in individual disciplines with interested nonspecialists: not just people willing to confer with those working at related levels, but researchers trained in the methods and factual details of a variety of levels. This is a daunting requirement, but it does offer some hope for philosophers wishing to contribute to future neuroscience. Thinkers trained in both the 'synoptic vision' afforded by philosophy and the factual and experimental basis of genuine graduate-level science would be ideally equipped for this task. Recognition of this potential niche has been slow among graduate programs in philosophy, but there is some hope that a few programs are taking steps to fill it. (See, e.g., "Other Internet Resources," linked below.)

7. Recent Developments in the Philosophy of Neuroscience

The distinction between "philosophy of neuroscience" and "neurophilosophy" has become clearer, due primarily to more questions now being pursued in both areas. Philosophy of neuroscience still tends to pose traditional questions from philosophy of science specifically about neuroscience. Such questions include: What is the nature of neuroscientific explanation? And, what is the nature of discovery in neuroscience? Answers to these questions can be pursued either descriptively (how does neuroscience proceed?) or normatively (how should neuroscience proceed)? Normative projects in philosophy of neuroscience can be deconstructive, by criticizing claims made by neuroscientists. For example, philosophers of neuroscience might criticize the conception of personhood

assumed by researchers in cognitive neuroscience (cf. Roskies 2009). Normative projects can also be constructive, by proposing theories of neuronal phenomena or methods for interpreting neuroscientific data. These latter projects are often integrated with theoretical neuroscience. For example, Chris Eliasmith and Charles Anderson developed an approach to constructing neurocomputational models in their book *Neural Engineering* (2003). In separate publications, Eliasmith has argued that the framework introduced in *Neural Engineering* provides both a normative account of neural representation and a framework for unifying explanation in neuroscience (cf. Eliasmith 2009; Eliasmith 2009).

Neurophilosophy still applies findings from the neurosciences to traditional, mainstream philosophical questions. Examples now include: What is an emotion? (Prinz 2004) What is the nature of desire? (Schroeder 2004) How is social cognition made possible? (Goldman 2006)What is the neural basis of moral cognition? (Prinz 2007) What is the neural basis of happiness? (Flanagan 2009) Neurophilosophical answers to these questions are constrained by what neuroscience reveals about nervous systems. For example, in his book *Three Faces of Desire*, Timothy Schroeder argues that our commonsense conception of desire attributes to it three capacities: (1) the capacity to reinforce behavior when satisfied, (2) the capacity to motivate behavior, and (3) the capacity to determine sources of pleasure. Based on evidence from the literature on dopamine function and reinforcement learning theory, Schroeder argues that reward processing is the basis for all three capacities. Thus, reward is the essence of desire.

At present, there is a trend in neurophilosophy to look toward neuroscience for guidance in moral philosophy. That should be evident from the themes we've just mentioned. Simultaneously, there has arisen interest in moralizing about neuroscience and neurological treatment (see Levy 2007; Roskies 2009). The new field of neuroethics combines both interest in the relevance of neuroscience data for understanding moral cognition and the relevance of moral philosophy for regulating the application of knowledge from neuroscience. The regulatory branch of neuroethics focuses on the ethics of treatment for people who suffer from neurological impairments, the ethics of attempts to enhance human cognitive performance (Schneider 2009), the ethics of applying "mind reading" technology to problems in forensic science (Farah and Wolpe 2004), and the ethics of animal experimentation in neuroscience (Farah 2008).

Other recent trends, now in philosophy of neuroscience, include renewed interest in the nature of mechanistic explanations, given the widespread use of the term among neuroscientists. In his book,*Explaining the Brain* (2005), Carl Craver contends that mechanistic explanations in neuroscience are causal and typically multi-level. For example, the explanation of the neuronal action potential involves the action potential itself, the cell in which it occurs, electro-chemical gradients, and the proteins through which ions flow. Here we have a composite entity (a cell) causally interacting with neurotransmitters at its receptors. Parts of the cell engage in various activities (the opening and closing of ligand-gated and voltage-gated ion channels) to produce a pattern of change (the depolarizing current constituting the action potential). The mechanistic explanation of the action potential countenances entities at the cellular, molecular, and atomic levels, each of which

are causally relevant to producing the action potential. This causal relevance can be confirmed by altering any one of these variables (e.g. the density of ion channels in the cell membrane) to generate alterations in the action potential, and by verifying the consistency of the purported invariance between the variables. (For challenges to Craver's account of mechanistic explanation in neuroscience, specifically concerning the action potential, see Weber 2008 and Bogen 2005.)

According to epistemic norms shared by neuroscientists, good explanations in neuroscience are good mechanistic explanations, and good mechanistic explanations are those that pick out invariant relationships between mechanisms and the phenomena they control. (For fuller treatment of invariance in causal explanations throughout science, see James Woodward 2003.) Craver's account raises questions about the place of reduction in neuroscience. John Bickle (2003) suggests that the working concept of reduction in the neurosciences consists of the discovery of systematic relationships between interventions at lower levels of organization (as they are recognized in cellular and molecular neuroscience) and higher level behavioral effects (as they are described in psychology). Bickle calls this perspective "reductionism-in-practice" to contrast it with the concepts of intertheoretic or metaphysical reduction that have been the focus of many debates in the philosophy of science and philosophy of mind. Despite Bickle's reformulation of reduction, mechanists generally resist the "reductionist" label. Is mechanism a kind of reductionism-in-practice? Or does mechanism, as a position on neuroscientific explanation, assume some type of autonomy for psychology? If it does, reductionists can challenge mechanists on this assumption. On the other hand, Bickle's reductionism-in-practice clearly departs from inter-theoretic reduction, as the latter is understood in philosophy of science. As Bickle himself acknowledges, his latest reductionism was inspired heavily by mechanists' criticisms of his earlier "new wave" account. Mechanists can challenge Bickle that his departure from the traditional accounts has also led to a departure from the interests that motivated those accounts. (See Polger 2004 for a related challenge.)

The role of temporal representation in conscious experience and the kinds of neural architectures sufficient to represent objects in time has generated recent interest. In the tradition of Husserl's phenomenology, Dan Lloyd (2002, 2003) and Rick Grush (2001, 2009) have separately drawn attention to the tripartite temporal structure of phenomenal consciousness as an explanandum for neuroscience. This structure consists of a subjective present, an immediate past, and an expectation of the immediate future. For example, one's conscious awareness of a tune is not just of a time-slice of tune-impression, but of a note that a moment ago was present, another that is now present, and an expectation of subsequent notes in the immediate future. As this experience continues, what was a moment ago temporally immediate is now retained as a moment in the immediate past, what was expected either occurred or didn't in what has now become the experienced present, and a new expectation has formed of what will come. One's experience is not static, even though the experience is of a single object (the tune).

According to Lloyd, the tripartite structure of consciousness raises a unique problem for analyzing fMRI data and designing experiments. The problem stems from the tension

between the sameness in the object of experience (e.g. the same tune through its progression) and the temporal fluidity of experience itself (e.g. the transitions between heard notes). A standard means of analyzing fMRI data consists in averaging several data sets and subtracting an estimate of baseline activation from the composites (discussed in an earlier section of this entry). This is done to filter noise from the task-related hemodynamic response. But this practice ignores much of the data necessary for studying the neural correlates of consciousness. It produces static images that neglect the relationships between data points in the time course. Lloyd instead applies a multivariate approach to studying fMRI data, under the assumption that a recurrent network architecture underlies the temporal processing that gives rise to experienced time. A simple recurrent network has an input layer, an output layer, a hidden layer, and an additional layer that copies the prior activation state of either the hidden layer or the output layer. Allowing the output layer to represent a predicted outcome, the input layer can then represent a current state and the additional layer a prior state. This assignment mimics the tripartite temporal structure of experience in a network architecture. If the neuronal mechanisms underlying conscious experience are approximated by recurrent network architecture, one prediction is that current neuronal states carry information about immediate future and prior states. Applied to fMRI, the model predicts that time points in an image series will carry information about prior and subsequent time points. The results of Lloyd's (2002) analysis of 21 subjects' data sets, sampled from the publicly accessible National fMRI Data Center, support the prediction.

Grush's (2001, 2004) interest in temporal representation is part of his broader systematic project addressing a semantic problem for computational neuroscience, namely: how do we demarcate study of the brain as an information processor from the study of any other complex causal process? This question leads back into the familiar territory of psychosemantics, but now the starting point is internal to the practices of computational neuroscience. The semantic problem is thereby rendered an issue in philosophy of neuroscience, insofar as it asks: what does (or should) ‘computation’ mean in computational neuroscience?

Grush's solution draws on concepts from modern control theory. In addition to a controller, a sensor, and a goal state, certain kinds of control systems employ a *process model* of the actual process being controlled. A process model can facilitate a variety of engineering functions, including overcoming delays in feedback and filtering noise. The accuracy of a process model can be assessed relative to its “plug-compatibility” with the actual process. Plug-compatibility is a measure of the degree to which a controller can causally couple to a process model to produce the same results it would produce by coupling with the actual process. Note that plug-compatibility is not an information relation.

To illustrate a potential neuroscientific implementation, Grush considers a controller as some portion of the brain's motor systems (e.g., premotor cortex). The sensors are the sense organs (e.g., stretch receptors on the muscles). A process model of the musculoskeletal system might exist in the cerebellum (see Kawato 1999). If the controller portion of the motor system sends spike trains to the cerebellum in the same way that it sends spikes to

the musculoskeletal system, and if in return the cerebellum receives spike trains similar to real peripheral feedback, then the cerebellum emulates the musculoskeletal system (to the degree that the mock feedback resembles real peripheral feedback). The proposed unit over which computational operations range is the neuronal realization of a process model and its components, or in Grush's terms an “emulator” and its “articulants.”

The details of Grush's framework are too sophisticated to present in short compass. (For example, he introduces a host of conceptual devices to discuss the representation of external objects.) But in a nutshell, he contends that understanding temporal representation begins with understanding the emulation of the timing of sensorimotor contingencies. Successful sequential behavior (e.g., spearing a fish) depends not just on keeping track of where one is in space, but where one is in a temporal order of movements and the temporal distance between the current, prior, and subsequent movements. Executing a subsequent movement can depend on keeping track of whether a prior movement was successful and whether the current movement is matching previous expectations. Grush posits emulators—process models in the central nervous system—that anticipate, retain, and update mock sensorimotor feedback by timing their output proportionally to feedback from an actual process (see Grush forthcoming).

Lloyd's and Grush's approaches to studying temporal representation are varied in their emphases. But they are unified in their implicit commitment to localizing cognitive functions and decomposing them into subfunctions using both top-down and bottom-up constraints. (See Bechtel and Richardson 1993 for more details on this general explanatory strategy.) Both develop mechanistic explanations that pay little regard to disciplinary boundaries. One of the principal lessons of Bickle's and Craver's work is that neuroscientific practice in general is structured in this fashion. The ontological consequences of adopting this approach are now being actively debated.

Given that philosophy of neuroscience, as other branches of philosophy of science, has both descriptive and normative aims, it is critical to develop methods for accurate estimation of current norms and practices in neuroscience. Appeals to intuition will not suffice, nor will single paradigm case studies do the job because those case studies may fail to be representative. For example, an attempt to reconstruct the conditions under which the mechanism of the action potential was discovered may tell us little about the nature of discovery for other neural mechanisms. But, large case samples may be difficult to log and analyze. Furthermore, without protocols to guide such reconstructions, the conclusions are susceptible to hidden biases, i.e. cherry picking of data to support one's conclusions. Recent work by Alcino Silva, Anthony Landreth, and John Bickle makes concrete proposals for undertaking large-scale studies of the explanatory norms and the growth of causal knowledge in neuroscience (see Silva, Landreth, and Bickle forthcoming). They outline a framework for classifying, documenting and analyzing experiments in neuroscience with practical applications for planning relevant future experiments.

Bibliography

- Akins, K., 1993a, “What Is It Like to be Boring and Myopic?” in *Dennett and His Critics*, B. Dahlboom (ed.), New York: Basil Blackwell, pp. 124–160.
- —— 1993b, “A Bat Without Qualities,” in *Consciousness: Psychological and Philosophical Essays*, M. Davies and G. Humphreys (eds.), New York: Basil Blackwell, pp. 258–273.
- —— 1996, “Of Sensory Systems and the ‘Aboutness’ of Mental States,” *Journal of Philosophy*, 93: 337–372.
- Aston-Jones, G., Desimone, R., Driver, J., Luck, S., and Posner, M., 1999, “Attention,” in Zigmond*et al*., 1385–1410.
- Balzer, W., Moulines, C. U., and Sneed, J., 1987, *An Architectonic for Science*, Dordrecht: Reidel.
- Bechtel, W., 1998, “Representations and Cognitive Explanations: Assessing the Dynamicist Challenge in Cognitive Science,” *Cognitive Science*, 22: 295–318.
- Bechtel, W., and J. Mundale, 1999, “Multiple Realizability Revisited: Linking Cognitive and Neural States,” *Philosophy of Science*, 66: 175–207.
- Bechtel, W., and R. Richardson, 1993, *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*, Princeton, NJ: Princeton University Press.
- Bechtel, W., P. Mandik, J. Mundale, and R.S. Stufflebeam, 2001, *Philosophy and the Neurosciences: A Reader*, Oxford: Blackwell.
- Bermudez, J.L., 1998, *The Paradox of Self-Consciousness*, Cambridge, MA: MIT Press.
- Bickle, J., 1992, “Revisionary Physicalism,” *Biology and Philosophy*, 7: 411–430.
- —— 1995, “Psychoneural Reduction of the Genuinely Cognitive: Some Accomplished Facts,”*Philosophical Psychology*, 8: 265–285.
- —— 1998, *Psychoneural Reduction: The New Wave*, Cambridge, MA: MIT Press.
- ——, 2003, *Philosophy and Neuroscience: A Ruthlessly Reductive Account*, Norwell, MA: Kluwer Academic Press.
- —— (ed.), 2009, *The Oxford Handbook of Philosophy and Neuroscience*, New York: Oxford University Press.
- Biro, J., 1991, “Consciousness and Subjectivity,” in *Philosophical Issues*, E. Villaneuva (ed.), Atascadero, CA: Ridgeview, pp. 113–133.
- Bliss,T.V.P. and T. Lomo, 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: 331–356.
- Block, N, 1986, “Advertisement for a Semantics for Psychology,” in *Midwest Studies in Philosophy*, P. French, *et al*. (eds.), 10: 617–678.

- ——— ,1988, “On a Confusion About a Function of Consciousness,” *Behavioral and Brain Sciences*, 18: 227–247.
- Bogen, J., 2005, “Regularities and Causality: Generalizations and Causal Explanations,” *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36: 397–420.
- Bower, J. and D. Beeman, 1995, *The Book of GENESIS*, New York: Springer-Verlag.
- Caplan, D., T. Carr, J. Gould, and R. Martin, 1999, “Language and Communication,” in Zigmond *et al*. 1999, pp. 1329–1352.
- Chalmers, D., 1996, *The Conscious Mind*, Oxford: Oxford University Press.
- Churchland, P., 1986, *Neurophilosophy*, Cambridge, MA: MIT Press.
- Churchland, P.S. and T. Sejnowski, 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.
- ———, 1987, *Matter and Consciousness*, revised edition. Cambridge, MA: MIT Press.
- ———, 1989, *A Neurocomputational Perspective*, Cambridge, MA: MIT Press.
- ———, 1995, *The Engine of Reason, The Seat of the Soul*, Cambridge, MA: MIT Press.
- ———, 1996, “The Rediscovery of Light,” *Journal of Philosophy*, 93: 211–228.
- Churchland, P.M., and P.S. Churchland, 1997, “Recent Work on Consciousness: Philosophical, Empirical and Theoretical,” *Seminars in Neurology*, 17: 101–108.
- Clark, A., 1993, *Sensory Qualities*, Cambridge: Cambridge University Press.
- Craver, C., 2007, *Explaining the Brain: What the Science of the Mind-Brain Could Be*, Oxford University Press.
- Dennett, D.C., 1978, “Why You Can't Make a Computer That Feels Pain,” *Synthese*, 38: 415–456.
- ———, 1991, *Consciousness Explained*, New York: Little Brown.
- ———, 1995, “The Path Not Taken,” *Behavioral and Brain Sciences*, 18: 252–253.
- Dretske, F. ,1981, *Knowledge and the Flow of Information*, Cambridge, MA: MIT Press.
- ———, 1988, *Explaining Behavior*, Cambridge, MA: MIT Press.
- Eliasmith, C, and C.H. Anderson, 2003, *Neural Engineering*, Cambridge, MA: MIT Press.
- ———, 2009, “Neurocomputational Models: Theory, Application, Philosophical Consequences,” in Bickle (ed.) 2009, pp. 346–369.

- Farah, M.J., 2008, “Neuroethics and the problem of other minds: Implications of neuroscience evidence for the moral status of brain-damaged patients and nonhuman animals,” *Neuroethics*, 1: 9–18.
- Farah, M.J. and P.R Wolpe, 2004, “Monitoring and manipulating the human brain: New neuroscience technologies and their ethical implications,” *Hastings Center Report*, 34: 35–45.
- Feyerabend, P., 1963, “Mental Events and the Brain,” *Journal of Philosophy*, 60: 295–296.
- Flanagan, O., 2009, “Neuro-eudaimonics, or Buddhists lead neuroscientists to the seat of happiness,” in Bickle (ed.) 2009, pp. 582–600.
- Fodor, J., 1974, “Special Sciences,” *Synthese*, 28: 77–115.
- ——, 1981, *RePresentations*, Cambridge, MA: MIT Press.
- ——, 1987, *Psychosemantics*, Cambridge, MA: MIT Press.
- Fodor, J. and E. LePore, 1992, *Holism: A Shopper's Guide*, Cambridge, MA: MIT Press.
- Gazzaniga, M. (ed.), 1995, *The Cognitive Neurosciences*, Cambridge, MA: MIT Press.
- Georgopoulos, A., A. Schwartz, and R. Kettner, 1986, “Neuronal Population Coding of Movement Direction,” *Science*, 233: 1416–1419.
- Grush, R., 2001, “The Semantic Challenge to Computational Neuroscience,” in *Theory and Method in the Neurosciences*, Peter Machamer, Rick Grush and Peter McLaughlin (eds.), Pittsburgh, PA: University of Pittsburgh Press, pp. 155–172.
- Grush, R., 2004, “The Emulation Theory of Representation: Motor Control, Imagery, and Perception,” *Behavioral and Brain Sciences*, 27: 377–442.
- Grush, R., forthcoming, “Brain Time and Phenomenological Time,” in *Cognition and the Brain: The Philosophy and Neuroscience Movement*, Kathleen Akins, Andrew Brook and Steven Davis (eds.), Cambridge: Cambridge University Press, pp. 160–207.
- Hardcastle, V.G., 1997, “When a Pain Is Not,” *Journal of Philosophy*, 94: 381–409.
- Hardin, C.L., 1988, *Color for Philosophers*, Indianapolis:Hackett.
- Haugeland, J., 1985, *Artificial Intelligence: The Very Idea*, Cambridge, MA: MIT Press.
- Hawkins, R. and E. Kandel, 1984, “Is There a Cell-Biological Alphabet for Learning?”*Psychological Review*, 91: 375–391.
- Hebb, D.O., 1949, *The Organization of Behavior*, New York: Wiley.
- Hirstein, W., 2003, *Brain Fiction*, Cambridge, MA: MIT Press.
- Hodgkin, A.L., Huxley, A.F., and Katz, B., 1952, “Measurement of current-voltage relations in the membrane of the giant axon of *Loligo*”, *Journal of Physiology* 116(4): 442–448.
- 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, 201–236, 496–529.

- Horgan, T. and G. Graham, 1991, “In Defense of Southern Fundamentalism,” *Philosophical Studies*, 62: 107–134.
- Hubel, D. and T. Wiesel, 1962, “Receptive Fields, Binocular Interaction and Functional Architecture In the Cat's Visual Cortex,” *Journal of Physiology* (London), 160: 106–154.
- Jackson, F. and P. Pettit, 1990, “In Defense of Folk Psychology,” *Philosophical Studies*, 59: 31–54.
- Kandel, E., 1976, *Cellular Basis of Behavior*, San Francisco: W.H. Freeman.
- Kawato, M., 1999, “Internal Models for Motor Control and Trajectory Planning,” *Current Opinion in Neurobiology*, 9: 18–27.
- Kolb, B. and I. Whishaw, 1996, *Fundamentals of Human Neuropsychology*, 4th edition, New York:W.H. Freeman.
- Kosslyn, S., 1997, “Mental Imagery,” in S. Gazzaniga (ed.), *Conversations in the Cognitive Neurosciences*, Cambridge, MA: MIT Press, pp. 37–52.
- Lee, C.W., R. Rohrer, D. and Sparks, 1988, “Population Coding of Saccadic Eye Movements by Neurons in the Superior Colliculus,” *Nature*, 332: 357–360.
- Lehky, S.R. and T. Sejnowski, 1988, “Network Model of Shape-from-Shading: Neural Function Arises from Both Receptive and Projective Fields,” *Nature*, 333: 452–454.
- Lettvin, J.Y., H.R. Maturana, W.S. McCulloch, W.H. and Pitts, 1959, “What the Frog's Eye Tells the Frog's Brain,” *Proceedings of the IRF*, 47: 1940–1951.
- Levine, J., 1983, “Materialism and Qualia: The Explanatory Gap,” *Pacific Philosophical Quarterly*, 64: 354–361.
- Levy, N., 2007, *Neuroethics: Challenges for the 21st Century*, Cambridge: Cambridge University Press.
- Llinás, R., 1975, “The Cortex of the Cerebellum,” *Scientific American*, 232: 56–71
- Llinás, R., and P.S. Churchland, (eds.), 1996, *The Mind-Brain Continuum*, Cambridge, MA: MIT Press.
- Lloyd, D., 2002, “Functional MRI and the Study of Human Consciousness,” *Journal of Cognitive Neuroscience*, 14: 818–831.
- Lloyd, D., 2003, *Radiant Cool: A Novel Theory of Consciousness*, Cambridge, MA: MIT Press.
- Magistretti, P., 1999, “Brain Energy Metabolism,” in Zigmond *et al*. (eds.) 1999, pp. 389–413.
- Nagel, T., 1971, “Brain Bisection and the Unity of Consciousness,” *Synthese*, 25: 396–413.
- —— (1974) “What Is It Like to Be A Bat?” *Philosophical Review*, 83: 435–450.
- Place, U.T., 1956, “Is Consciousness a Brain Process?” *The British Journal of Psychology*, 47: 44–50.
- Polger, T., 2004, *Natural Minds*, Cambridge, MA: MIT Press.

- Prinz, J., 2007, *The Emotional Construction of Morals*, Oxford: Oxford University Press.
- Putnam, H., 1967, "Psychological Predicates," in *Art, Mind, and Religion*, Capitan and Merrill (eds.), Pittsburgh: University of Pittsburgh Press, pp. 49–54.
- Rall, W., 1959, "Branching Dendritic Trees and Motoneuron Membrane Resistivity," *Experimental Neurology*, 1: 491–527.
- Ramsey, W., 1992, "Prototypes and Conceptual Analysis," *Topoi*, 11: 59–70
- Roskies, A., 2009, "What's 'Neu' in Neuroethics?" in Bickle (ed.) 2009, pp. 454–472.
- Rumelhart, D., G. Hinton, and J. McClelland, 1986, "A Framework for Parallel Distributed Processing," in Rumelhart and McClelland (eds.), *Parallel Distributed Processing*, Cambridge, MA: MIT Press, pp. 45–76.
- Sacks, O., 1985, *The Man Who Mistook his Wife for a Hat*, New York: Summit Books
- Schaffner, K., 1992, "Philosophy of Medicine," in *Introduction to the Philosophy of Science*, M. Salmon, J. Earman, C. Glymour, J. Lennox, P. Machamer, J. McGuire. J. Norton, W. Salmon, and K. Schaffner (eds.), Englewood Cliffs, NJ: Prentice-Hall, pp. 310-345.
- Schneider, S., 2009, "Future Minds: Transhumanism, Cognitive Enhancement and the Nature of Persons," in *University of Pennsylvania Bioethics Reader*, A. Caplan and V. Radvisky, (eds.), pp. 95–110. New York: Springer.
- Schroeder, T., 2004, *Three Faces of Desire*, Oxford: Oxford University Press.
- Silva, A.J., Landreth, A.W. and Bickle, J., (forthcoming), *Engineering the Next Revolution in Neuroscience*, New York: Oxford University 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 W. Bechtel, 1997, "PET: Exploring the Myth and the Method," *Philosophy of Science* (Supplement), 64 (4): S95-S106.
- Suppe, F., 1974, *The Structure of Scientific Theories*, Urbana: University of Illinois Press.
- Tye, M., 1993, "Blindsight, the Absent Qualia Hypothesis, and the Mystery of Consciousness," in*Philosophy and the Cognitive Sciences*, C. Hookway (ed.), Cambridge: Cambridge University Press, pp. 19–40.
- Van Fraassen, B.C., 1980, *The Scientific Image* Oxford University Press.
- 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*Explorations in the Biology of Language*, E. Walker (ed.), Cambridge, MA: MIT Press, pp. 27–66.
- Weber, M., 2008, "Causes without Mechanisms: Experimental Regularities, Physical Laws, and Neuroscientific Explanation," *Philosophy of Science*, 75: 995–1007.

- Woodward, J., 2003, *Making Things Happen: A Theory of Causal Explanation*, Oxford: Oxford University Press.
- Zigmond, M., F. Bloom, S. Landis, J. Roberts, L. and Squire, L. (eds.), 1999, *Fundamental Neuroscience*, San Diego: Academic Press.