14 Towards a Mechanistic Philosophy of Neuroscience

Carl F. Craver and
David M. Kaplan

1. Introducing Neuroscience

Neuroscience is an interdisciplinary research community united by the goal of understanding, predicting and controlling the functions and malfunctions of the central nervous system (CNS). The philosophy of neuroscience is the subfield of the philosophy of science concerned with the goals and standards of neuroscience, its central explanatory concepts, and its experimental and inferential methods.[1] Neuroscience is especially interesting to philosophers of science for at least three reasons. First, neuroscience is immature in comparison to physics, chemistry and much of biology and medicine. It has no unifying theoretical framework or common vocabulary for its myriad subfields. Many of its basic concepts, techniques and exemplars of success are under revision simultaneously. Neuroscience thus exemplifies a form of scientific progress in the absence of an overarching paradigm (Kuhn 1970). Second, neuroscience is a physiological science. Philosophers of biology have tended to neglect physiology (though see Schaffner 1993; Wimsatt 2007). Physiological sciences study the parts of organisms, how they are organized together into systems, how they work and how they break. Its generalities are not universal in scope. Its theories intermingle concepts from several levels. Neuroscience thus offers an opportunity to reflect on the structure of physiological science more generally. Finally, unlike other physiological sciences, neuroscientists face the challenges of relating mind to brain. The question arises whether the explanatory resources of physiological science can be extended into the domains involving consciousness, rationality and agency, or whether such phenomena call out for distinctive explanatory resources.

The philosophy of neuroscience should also interest neuroscientists. Because neuroscience is so young, and because it faces unique explanatory

demands, it is useful to reflect on what the science hopes to accomplish and how it might most efficiently achieve its objectives. Philosophers have much to contribute to that discussion.

Below, we sketch five debates in the philosophy of neuroscience. In Section 2, we discuss the debate between predictivists and mechanists about the norms of explanation. In section 3, we apply this debate to the question of whether dynamical systems explanations are legitimate, non-mechanistic types of explanation. In Section 4, we discuss the debate between mechanistic realists and mechanistic pragmatists about how to build a periodic table of the mind-brain. In Section 5, we discuss the strengths and limitations of deficit studies as an instance of experimental strategies for discovering mechanisms. In Section 6, we discuss the thesis that the brain is a computer, considering debates about how to define computation and about whether computational explanations are mechanistic explanations. In Section 7, we discuss the relative merits of reductive and integrative visions of neuroscience. A consistent theme throughout is that the philosophy of neuroscience can make considerable progress if it starts with the central idea that neuroscience is fundamentally a mechanistic science.

2. Mechanistic Explanation

Neuroscientists are united by the collective project of explaining how the nervous system and its parts work. Yet there is no consensus about what counts as a successful explanation. Consider two clearly contrasting views. Predictivists believe that any good predictive model is explanatory. Mechanists believe that, in addition, explanation requires knowledge of a mechanism.

Predictivists assert, for example, that to explain a phenomenon is to show that its occurrence follows from universal or statistical regularities, together with the relevant initial conditions. This view of explanation once dominated the philosophy of science (Hempel 1965). On a very liberal interpretation of predictivism, any mathematical or computational model that predicts all the relevant features of the phenomenon in a wide range of conditions counts as an explanation.

Mechanists, in contrast, insist that explanatory models describe the relevant causes and mechanisms in the system under study. In their view, a predictively adequate mathematical model might fail to explain how the system behaves. To explain a phenomenon, one has to know the mechanism that produces it, one has to know what its components are, what they do and how they are organized together (spatially, temporally and hierarchically) such that they give rise to the phenomenon to be explained (see Bechtel and Richardson 1993; Machamer et al. 2000; Craver 2007).

Predictivists and mechanists have argued over how to understand a central exemplar of modern neuroscience: the Hodgkin-Huxley model of the action potential. Action potentials are fleeting changes in the voltage difference across the neuronal membrane that serve as a basic unit of neural signalling. Hodgkin and Huxley hoped to discover how neurons generate action potentials, but their techniques could not discern the molecular mechanisms we now know to be involved. Neuroscientists have since discovered that action potentials are produced by voltage-sensitive ion channels that control the diffusion of sodium and potassium ions across the membrane. To understand how the debate between predictivists and mechanists plays out in this example, some historical context is needed.

Hodgkin and Huxley initially focused on a more tractable problem: to characterize changes of membrane conductance for sodium and potassium as a function of membrane voltage. They used a voltage clamp to hold the membrane voltage at particular values by balancing the resulting flow of current through the membrane. They could then calculate the membrane conductance at a particular voltage from the current required to balance the flow of current through the membrane at that voltage. Their primary modelling achievement was to generate equations that described the voltage-conductance relationships discovered with the voltage clamp. Using basic concepts from electrical theory, Hodgkin and Huxley represented the membrane as an electrical circuit, with currents carried by ions along parallel paths, each with its own battery (driving force) and variable resistor (the conductances). The electrical circuit schema could then be described mathematically, providing a formal model of the membrane that predicts much of the electrical behaviour of neuronal membranes. This mathematical model remains in wide use.

The core of the model is the total current equation (Eq. 1), expressed as a set of coupled partial differential equations describing the voltage-dependent changes in ionic conductances across the neuronal membrane:

$$I = C_M dv/dt + G_k n^4 (V - V_K) + G_{Na} m^3 h(V - V_{Na}) + G_1(V - V_1) \quad (1)$$

where I is the total current crossing the membrane, composed of a capacitive current $C_M dV/dt$, a potassium current $G_K n^4(V - V_K)$, a sodium current $G_{Na} m^3 h(V - V_{Na})$, and the leakage current $G_l(V - V_l)$, a sum of smaller currents for other ions. G_K, G_{Na} and G_l are the maximum conductances to the different ions. V is displacement of Vm from Vrest. V_K, V_{Na} and V_l are the differences between equilibrium potentials for the various ions (where diffusion and driving force balance, and no net current flows) and Vm. Crucially, the model also includes the three coefficients, n, m and h, whose values determine

how conductance varies with voltage and time (see Hodgkin and Huxley 1952).

Mechanists and predictivists debate whether aspects of the Hodgkin-Huxley model genuinely explain, or merely predict, how the membrane behaves at different voltages (Bogen 2005, 2008; Craver 2006, 2008; Machamer et al. 2000; Schaffner 2008; Weber 2005, 2008). The debate focuses on the rate coefficients n, m and h. In building the model, Hodgkin and Huxley considered a hypothetical mechanism involving the movement of activation and inactivation particles in the membrane. The coefficient n is raised to the fourth power in the rate equation for potassium conductance, one might think, because the maximum conductance depends on the probability that each of four activation particles has moved as it must. A similar interpretation is available for the term m^3h in the equation for sodium conductance. It is important to remember, however, that Hodgkin and Huxley had only vague and speculative ideas about how membranes change conductance. The idea of ion-selective channels did not appear until the 1960s, and neuroscientists debated their existence well into the 1980s (see Hille 1992). This is why Hodgkin and Huxley insist that, 'the success of the equations is no evidence in favour of the mechanism of permeability change that we tentatively had in mind when formulating them.' They explain: 'An equally satisfactory description of the voltage clamp data could no doubt have been achieved with equations of very different form, which would probably have been equally successful in predicting the electrical behaviour of the membrane.' (1952, p. 541) That is, they did not know the mechanism.

Hodgkin and Huxley were mechanists. Mechanists can agree with predictivists that explanatory models should make accurate predictions. They insist, however, on the distinction between explanatory models and phenomenal models. Phenomenal models characterize a phenomenon without explaining it, or even pretending to explain it. For example, Snell's law describes how light refracts as it passes from one medium to another, but the law does not explain why the path of light changes as it does. Theoretical neuroscientists Dayan and Abbott take this mechanistic stance when they distinguish purely descriptive models that 'summarize data compactly' from mechanistic models that 'address the question of how nervous systems operate on the basis of known anatomy, physiology, and circuitry' (2001, p. xiii).

How might predictively adequate models fail as explanations? One might build a model to predict a cause on the basis of its effects. A car's heat gauge affords accurate predictions of the engine's temperature, but it does not explain it, presumably because the heat gauge indicates, but does not cause, the engine's changes in temperature. To continue with the example, the heat gauge reading can predict radiator blowouts, but does not explain them when they

occur. In this case, the correlational relation between them is explained by a common cause (engine heat). For these reasons (and others like them) mechanists hold that explanatory models reveal causal mechanisms. Hodgkin and Huxley insist that their model merely predicts how the membrane changes conductance with voltage, because they know it does not describe the mechanism for those changes. Finally, this example illustrates the mechanist's distinction between how-possibly and how-actually explanations (Dray 1957; Hempel 1965). Hodgkin and Huxley's activation and inactivation particles are how-possibly fictions, used as heuristic tools to interpret their mathematical model. It does not count against the Hodgkin-Huxley model that the how-actually model has turned out to be quite a bit more complicated (see, for example, Doyle et al. 1998; Yu and Catterall 2003). That part of the model is intended for prediction, not explanation.

Summarizing these considerations, mechanists hold explanatory models to a strict model-to-mechanism-mapping (3M) requirement:

> (3M) A model of a target phenomenon explains that phenomenon when (a) the variables in the model correspond to identifiable components and organizational features of the target mechanism that produces, maintains, or underlies the phenomenon, and (b) the causal relations posited among these variables in the model correspond to the activities or operations among the components of the target mechanism.

Why do mechanists think these mapping relations mark a crucial difference between explanations and non-explanations? Two reasons are primary. First, mechanists think that explanation and know-how are deeply connected to one another. The mechanist reserves the term 'explanation' for descriptions that reveal how things work, precisely because knowing how things work is a reliable route to identifying ways of making systems work for us. Second, as illustrated in the above examples, predictivism violates scientific common sense about what does and does not count as an adequate explanation (see, for example, Salmon 1984). Predictions can be made on the basis of correlations, and only some correlations are explanatory. Perhaps predictivism can be restricted to block these devastating counter-examples, but the efforts to do so will only bring it closer into line with the mechanist point of view. One might require that explanations appeal to laws of nature, or that they unify diverse phenomena within a single explanatory framework. Mechanists think that such demands are ill-fitted to the explanatory domains of physiological sciences (Bechtel and Abrahamsen 2005), and they doubt that such corrections can be made to predictivism without collapsing into a version of mechanism.

3. Need for Non-mechanistic Explanation?

Some dynamicists argue that complex systems require a non-mechanistic form of explanation. Mechanistic explanations suffice for simple systems, in which parts work sequentially like a flow chart. But few target systems have such a unidirectional, sequential architecture. Neural systems often have tens to millions of parts interacting promiscuously with one another, often through reciprocal feedback connections. They exhibit non-linear changes in behaviour; minor changes in a single part or in the operating conditions have dramatic effects on the behaviour of the whole. As the number of parts and the diversity and interdependence of their interactions increases, it becomes harder to see the system as made up of discrete parts. Contrary to the demands of 3M, dynamicists argue, the phenomenon must be understood on its own level, in terms of system variables and order parameters (see Beckermann et al. 1992; Stephan 2006). For such systems, 'the primary explanatory tools' are the mathematical methods of non-linear dynamical systems theory (Chemero and Silberstein 2007).

Some mechanists respond that complex systems are, in fact, decomposable. Bechtel (1998) shows that fermentation and glycolysis, which have numerous feedback loops and non-linearities, can be broken down into systems and subsystems or parts and can be localized to specific parts of the cell. Action potential mechanisms also support decomposition. Action potentials are produced by complex feedback loops: one governing the behaviour of sodium channels and one governing the behaviour of potassium channels. They might be described as coupled oscillators. Nonetheless, it is possible to understand how those oscillators are implemented in the behaviour of decomposable subsystems in the cell (the sodium and potassium channels).

Mechanists might claim further that dynamical models fail as explanations so long as they violate the 3M requirement discussed above, for precisely the reasons mentioned in favour of that requirement.

Consider how the arguments of the previous section might be applied to a dynamical model in cognitive neuroscience, such as the Haken-Kelso-Bunz (Haken et al. 1985) model of human bimanual coordination (henceforth, HKB). Mechanists argue that such models are neither mechanistic nor explanatory of the phenomena they describe.

The HKB model (Eq. 2) describes a behavioural phenomenon arising in certain bimanual coordination tasks. Subjects are instructed to repeatedly move their index fingers up and down in time with a pacing metronome either in-phase (both fingers move up and down together) or antiphase (one finger moves up and the other moves down, and vice versa). Researchers systematically increase the tempo. Beyond a certain critical frequency, subjects can no longer maintain the antiphase movement and switch involuntarily into

in-phase movement. Subjects who begin in phase do not shift. Only in-phase movement is possible beyond the critical frequency.

Dynamicists characterize this system in terms of state-spaces and attractors representing the behaviour of two coupled oscillators. At slow tempos, the oscillators can be stably coupled to one another in both antiphase and in-phase movement modes. The state-space of the system is thus said to have two basins of attraction (or attractors) – one for antiphase and one for in-phase movement. At high tempos, only the in-phase mode is possible – the state-space has a single attractor. At the switch point, the attractor landscape of the system is said to change.

The HKB model provides an elegant mathematical description of this phenomenon, including the rate of change in the phase relationship between the left and right index fingers and the critical switch point from the antiphase to the in-phase pattern. The core of the model is a differential equation describing the coordination dynamics of these coupled components:

$$\dot{\phi} = -A \sin \phi - 2B \text{ Sin } 2\phi \qquad (2)$$

where ϕ is the phase difference (relative phase) between the two moving index fingers (when $\phi = 0$ the moving fingers are perfectly synchronous), and A and B reflect the experimentally observed oscillation frequencies of the fingers (for further discussion of the HKB model see Haken et al. 1985; Kelso 1995; Kessler and Kelso 2001).

Crucially, the modellers do not intend HKB to be a description of mechanisms or of the motor systems responsible for the experimentally observed behavioural dynamics.[2] Instead, it is a mathematically compact description of the temporal evolution of a purely behavioural dependent variable (relative phase) as a function of another purely behavioural independent variable (finger oscillation frequency). None of the variables or parameters of the HKB model map on to components and operations in the mechanism, and none of the mathematical relations or dependencies between variables map onto causal interactions between those system components (as required by 3M).[3]

Predictivists might argue that, for complex phenomena, explanation is just prediction. Mechanists will insist, however, that this view commits the predictivist either to accepting as explanations many things that are not accepted in any science as explanatory (such as mere correlations, relationships between an effect and its causes, and inferences from effects of common causes), or to proposing some means of distinguishing phenomenal redescriptions from explanations, for distinguishing how-possibly explanations from how-actually explanations, and for tracking progress in the search for deeper explanations. Dynamicists such as Kelso appear to recognize the importance of mechanistic explanation. After developing HKB, Kelso and colleagues began a research

programme to understand how this behavioural regularity results from features of the underlying organization of component neural systems and their dynamics (see, for example, Jantzen et al. 2009; Jirsa et al. 1998; Schöner 2002; Schöner and Kelso 1988). Dynamical models do not provide a separate kind of explanation subject to distinct standards. They are tools in the search for mechanistic explanations. Or so says the mechanist.

4. Functional and Structural Taxonomies

To discover a mechanism, one must first identify a phenomenon to explain. The taxonomy of cognitive capacities guides the search for their neural mechanisms. Neuroscience currently has nothing like the periodic table of chemical elements to guide its research projects: no systematic scheme of basic functional capacities of the mind or their neural and molecular mechanisms. To the extent that anything like such a taxonomy currently exists, it remains actively under revision, even in highly developed areas of neuroscience. As Bechtel and Richardson (1993) emphasize, one is often forced to revise one's taxonomy of phenomena as one learns about underlying mechanisms. Conversely, one's understanding of the mechanistic structure of the brain depends, to a large extent, on one's characterization of the phenomenon one seeks to understand. Thus, neuroscience is apparently engaged in iterative cycles of recalibration between models at different levels of organization, each of which must fit its findings to what is known about other levels. Mechanistic realists and mechanistic pragmatists differ from one another in their understanding of this process.

Mechanistic realists expect to ground the taxonomy of cognitive and neural phenomena in objective facts about the mechanistic structure of the brain. By learning which mechanisms are distinct from which others, one can carve the mind-brain at its joints. This process terminates in the periodic table of the mind-brain (or something asymptotically approaching it). This pattern of reasoning is evident among proponents of 'massive modularity', who believe the mind is subdivided into functionally distinct, domain-specific subsystems (Boyer 2002; Carruthers 2006; Fodor 1983; Pinker 1999). It is also evident in a host of investigative techniques used by neuroscience to localize functions in the brain (such as clinical dissociations, lesion studies and neuroimaging). Dissociation experiments show that the mechanisms underlying one cognitive capacity can be disrupted independently of the mechanisms underlying another. They drive taxonomic reform by splitting capacities into mechanistically distinct units. If cognitive capacities can sustain damage independently of one another, the realist claims, the putative capacity is heterogeneous, and a revised taxonomy, splitting the capacity in question, will be more useful for

prediction, explanation and control than is the taxonomy that includes the heterogeneous capacity.[4]

Pragmatists question the appropriateness of the periodic table as a metaphor for thinking about the taxonomic structure of neuroscience. They question whether there is a uniquely correct way of subdividing the functional capacities and mechanisms of the brain, and they emphasize that functional and structural decompositions depend in subtle ways on researchers' explanatory goals, pragmatic interests and theoretical orientations (see, for example, Uttal 2003). They charge that the realist faces a dilemma: either the realist view regresses, or it is viciously circular. The regress threatens because the view presupposes some method of carving the brain into distinct types of mechanisms, and that method appears no less difficult than the method of carving the mind into functionally distinct capacities. On the other horn, circularity threatens because our ability to carve the brain into distinct mechanisms requires some idea of what those mechanisms do, and this requires some commitment about which capacities require explanations. Mechanistic characterizations necessarily filter out the parts, activities and causal interactions among parts that are irrelevant to the capacity to be explained. Filtering judgements such as these are guided by a prior understanding of the capacities one hopes to understand. The pragmatist argues that there is no antecedently intelligible mechanistic structure in the brain that carves the mind into objectively correct functional units – functional and structural taxonomies are interdeterminate.

Pragmatists also charge that antecedent commitments about the correct cognitive taxonomy creep into the analysis of dissociation experiments. (Similar considerations apply, *mutatis mutandis*, to other localization techniques in cognitive neuroscience, including neuroimaging.) Neuropsychologists typically use a variety of tasks to provide a broad profile of the subject's motor, sensory and cognitive competencies and deficits. From the observed pattern of competencies and deficits, inferences can be drawn about the existence of damage to one or more neurocognitive mechanisms. Assumptions about the correct taxonomy of the mind-brain enter into task design, task analysis and diagnosis.

Furthermore, such assumptions reciprocally influence views about brain organization. Some cognitive neuropsychologists (e.g. Dunn and Kirsner 1988) argue that dissociation experiments establish that the brain has a modular architecture only on the assumption that the brain has such a modular architecture. By relaxing this assumption and allowing, for example, that cognitive dissociations might be produced through damage to single neural processing systems, one becomes more cautious about inferring distinct functions from deficit dissociations. Olton (1989) shows that apparent dissociations can also arise by varying task demands on memory tasks. Relatedly, Plaut (1995) shows that non-modular connectionist architectures can exhibit

apparent double dissociations, thus providing further reason for caution in drawing inferences about underlying mechanisms from task dissociation evidence. Lambdon Ralph and Rogers (2007), for example, show that one can produce category-specific naming deficits in connectionist networks without separate modules for the storage of particular semantic categories.

For these reasons, pragmatists argue that much of the evidence for the currently accepted taxonomy of brain functions makes nontrivial and contentious assumptions about the taxonomic structure of the mind-brain.

Regardless of how the debate between realists and pragmatists resolves, however, everyone can agree that the taxonomy of cognitive capacities is simultaneously anchored at multiple levels of organization. Churchland (1986) describes this process as the coevolution of theories at different levels. This coevolution develops as researchers with different techniques, vocabularies and even distinct standards of assessment attempt to negotiate their way to a consensus model of a mechanism. Neuroscience thus exemplifies a kind of conceptual development substantially different from that observed in more mature sciences, such as physics and chemistry.

5. Experiment and Evidence: Deficit Studies of Brain Mechanisms

Let us leave the debate between pragmatists and realists behind and assume we know which cognitive capacities require explanation and that we have tasks to distinguish them unambiguously. Can one, then, justifiably infer the organization of brain mechanisms on the basis of deficit studies (such as single and double dissociations)? Must we make additional assumptions beyond these? Are there kinds of mechanism that deficit studies alone are incapable of revealing under any reasonable set of assumptions?

Clark Glymour (1994, 2001; also see Bub 1994 for response) uses computational methods to investigate the conditions under which one can draw reliable inferences about underlying mechanisms on the basis of deficit studies. In order to apply his formal analysis, he makes a number of instructive assumptions in addition to those mentioned above (2001):

(1) That behavioural deficits are all or nothing, not graded (p. 136);
(2) That all normal individuals have the same cognitive architecture (p. 136);
(3) That the cognitive architecture in patients is a sub-graph of the normal architecture (that is, the normal architecture minus one and only one component) (p. 136);
(4) That the normal architecture has no cycles or feedback loops (p. 137);
(5) That the normal architecture is non-redundant (p. 137).

Glymour shows that if all of these are true and, in addition, we eventually see all of the patients that the cognitive architecture allows, then it is possible in many cases for cognitive scientists to converge on the correct cognitive architecture on the basis of deficit studies alone. This is a powerful result. However, some cognitive architectures are indistinguishable on the basis of deficit studies alone. For example, it is impossible for deficit studies to distinguish a cognitive architecture with one variable (V) between input and output and a cognitive architecture in which two variables (V1 and V2) lie in sequence between the input and output (Glymour 2001, pp. 144–5). Deficit studies also have difficulty distinguishing cases in which two cognitive mechanisms have been shown to be distinct from those in which a single cognitive mechanism has merely lost the resources (e.g. computational resources) required for it to perform both of its typical functions (Glymour 2001, pp. 146–8). Indeed, adding resource considerations into the models of brain functions makes even simple causal architectures exceedingly difficult to discover.

Glymour's starting assumptions, as he acknowledges, are frequently false. Patients often have multiple deficits resulting from remarkably different kinds of brain damage and disease, and the deficits lie anywhere on a spectrum from mild to global impairment (see Van Orden et al. 2001). Brains vary considerably from individual to individual (see Ojeman 1991). Brains reorganize in response to damage and disease, and new systems can sometimes take over the functions of the damaged system (see Hardcastle and Stewart 2002). Among the most dramatic examples of neural plasticity is the discovery that congenitally blind subjects use the primary visual cortex when performing demanding tactile discrimination tasks, such as braille reading, in comparison to normal subjects, who show deactivation in the same areas in equivalent tasks (Sadato et al. 1996; Cohen et al. 1997). In deficit studies, we do not study what the missing part did, but rather what the rest of the brain can do without it. Finally, brain systems are paradigmatically complex, involving multiple feedback loops at multiple levels and multiple redundant pathways that make development and recovery possible.

How does neuroscience make progress in the face of such challenges? Neuroscientists have an arsenal of techniques and strategies beyond deficit studies to study how brain mechanisms work. These techniques and strategies have non-overlapping strengths and weaknesses; they are more valuable when taken together than when taken singularly (see Bechtel 2002, 2007; Craver 2002, 2007). Deficit studies are a species of bottom-up inhibitory experiments in which one intervenes into a mechanism to inhibit one of its components, while monitoring the behaviour of the mechanism as a whole. Other examples include gene knockouts, transcranial magnetic stimulation and the use of pharmacological antagonists to inhibit systems in the brain.

Such interventions can often be delayed and reversed, allowing for subtly controlled manipulations.

Activation experiments, in contrast, involve activating a mechanism as a whole and monitoring the behaviours of its components. For example, subjects are engaged in tasks while brain activity is measured (using, for example, single-unit and multi-unit microelectrode recording and functional neuroimaging methods, including functional magnetic resonance imaging (fMRI), diffusion-tensor imaging (DTI), positron emission tomography (PET), electroencephalography (EEG), and magnetoencephalography (MEG)). Finally, stimulation experiments involve intervening to drive the components of a mechanism, while monitoring the behaviour of the mechanism as a whole. One might, for example, stimulate neurons in a brain region in order to produce specific motor patterns (Graziano et al. 2002) or to alter visual decision-making about motion direction in non-human primates (Salzman et al. 1990). Normative assessment of any single technique or strategy in isolation will yield more pessimistic results than would an analysis of the full range of mutually reinforcing techniques. Philosophers have focused on the integration of theories (Oppenheim and Putnam 1958) and the integration of fields (Darden 1991), but contemporary neuroscience affords the opportunity to consider how experimental strategies are integrated in the search for mechanisms.[5]

6. Computation and the Brain

Contemporary neuroscientists use computation to analyse, interpret and cross-validate experimental data; to generate testable predictions and refine theoretical hypotheses; and to optimize experimental designs (see, for example, Koch 1999; Koch and Segev 2000; Sejnowski 2009). Some neuroscientists and philosophers, computationalists, believe that the brain is a computer. They hold that brain systems perform computations and employ neural codes (see, for example, Rieke et al. 1999). Fictionalists think computation is in the eye of the modeller. It might be useful to describe the brain as a computer, but there is no well-defined sense in which neural systems, in fact, compute. The brain is simply a collection of causal mechanisms – full stop (Searle 1992).

To resolve the debate between computationalists and fictionalists, one must first decide what it means to compute. Grush (2001) cautions that an account of computation must navigate between the twin perils of triviality and irrelevance. On one hand, the notion cannot be left so permissive that everything computes (see Piccinini 2007a). For example, if one holds that a system computes if and only if a computer can simulate it, then brains compute. But so do thunderstorms, turbulent flows and orbiting planets. Or, if one holds that a system computes when its input-output transformations can be described

as Turing computable functions,[6] then most worldly causal processes count as computational. Churchland and Sejnowski (1992), and Sejnowski, Koch and Churchland (1988) hold something like this view: the brain computes in the sense that (a) its input-output behaviour can be interpreted in terms of some computable mathematical function, and (b) this interpretation is something we find useful or otherwise revealing. Clause (b) adds only the pragmatic thesis that, for whatever reason, computational explanations are especially revealing in a particular investigative context. This view postpones the question of why computational explanations are revealing in some cases and not others. It also looks very close to the fictionalist thesis that computing is in the eye of the beholder.

On the other hand, the notion of computation cannot be so restrictive as to make it utterly implausible from the start that brains compute. If one defines computation in terms of the behaviour of serial, von Neumann architectures or as operations on well-defined symbol strings according to a set of explicitly stored rules or lines of programme code (e.g. Cummins 1985), then the brain is not really a candidate for computing. The available models of computation from computer science (including abstract connectionist or neural network models (Rumelhart et al. 1986) are unlike brains in many respects. It would appear that computational neuroscience needs a proprietary model of neural computation.

Grush (2001), Eliasmith and Anderson (2004), and Shagrir (2006, 2010) have each attempted to develop such a proprietary notion. Grush argues that brain systems compute in the sense that they process information or transform internal representations. This view fits standard examples of computational neuroscience. Single neurons in the primary visual cortex (V1) are maximally active for stimuli presented at preferred orientations. They are commonly said to represent those orientations (Hubel and Wiesel 1968). Neurons in visual area MT fire preferentially for objects moving at specific speeds and in specific directions. Neurons in the motor cortex (M1) exhibit cosine tuning, that is, each cell is maximally responsive (fires most action potentials or spikes) for a particular direction of movement with activity declining as a function of the cosine of the angle between the preferred and actual movement direction (Georgepoulos 1982, 1986). Perhaps, as Grush claims, the proprietary notion of computation involves the manipulation of representations or the processing of information.

Why is this particular idea of computation useful to neuroscientists? Churchland (Churchland and Grush 1999) and Shagrir (2006) argue that mechanistic descriptions of neural systems in terms of firing rates of neurons across populations, for example, without any reference to what those activity patterns represent or how those representations are being transformed, would leave the functional role of those activity patterns utterly obscure. Shagrir argues that computational terms are essential when the project is to explain

how a semantic task, that is, a task defined in terms of representational content, is performed (2006, p. 393). This semantic view of computation is a received view in the philosophy of mind and in mainstream cognitive science. As Fodor (1981) demands: 'no computation without representation.'

Fictionalists hold out against these suggestions. They worry that the appeal to representation only delays the twin threats of triviality and irrelevance. To the extent that any causal system can be described as carrying information (as a river's height carries information about rainfall) or processing information (as a submerged river rock processes inputs related to current speed and outputs some transformation of it), the idea that the brain computes appears trivial. It amounts only to the claim that brain processes are causal processes, or that its various states are correlated with one another or with aspects of the external environment. Alternatively, one might attempt to apply a more restrictive notion of information that does not apply equally to weather patterns and turbulent flows.

One might require, for example, that states of the system have been selected for their being correlated with something else, or that the system has been trained up through development and learning so that certain states are so correlated (Dretske 1988, 1995). One might require that the states in question should be available for use by the organism in guiding behaviour. One might require further that information in the parochial sense refers only to states that are re-presenting (in the sense that they are duplicating some other input to the system). Critically, the fictionalist can argue that these more restrictive notions create bigger problems than they solve. If granted, such notions potentially render neuroscience obsolete for the task of identifying neural systems that implement computations. If computations are defined over representations that count as such if and only if they have the appropriate causal or evolutionary history, for example, then neuroscience will, in principle, be unable to answer questions about whether, and what, a given neuron or neuronal populations represents, or what computation a neuron performs. And this, one might reasonably think, is a far worse outcome for neuroscience than any consequences of adopting a fictionalist stance (e.g. utilizing a notion of computation with built-in subjectivity or observer-dependence). Thus, fictionalists will suspect that none of the proposed specifications will avoid both of Grush's twin perils: they will remain either overly broad, applying to systems that do not seem to be computing, or overly restrictive, ruling out systems that we think are genuinely computational.

The syntactic view of computation, in contrast, holds that computation can be defined without appealing to semantic notions (e.g. Egan 1995; Piccinnini 2007b). Piccinnini (2007b) argues that computing mechanisms are defined structurally in terms of system transformations of input strings of digits into output strings according to a rule defined over them, without reliance on

semantic interpretation of the inputs and outputs (or internal states). Anything that has these structural features is a computer. The view avoids the evil of triviality by placing strict constraints on systems that count as computers. Thunderstorms and rivers will not count as computers on this view. Concerning the claim of irrelevance, this view can respond that the account of computation comes with the credentials of the classical theory of computation (for overview, see Sipser 2005). Who better to settle the dispute about what does and does not count as a computer than the researchers responsible for developing our most rigorous, formal treatment of computation? And given that something like this abstract notion is implemented in the myriad digital computers that now surround us, there is at least a large class of things to which the account applies.

Fictionalists, however, will no doubt deny that the syntactic view satisfies Grush's challenge. For although the class of syntactic computing mechanisms is much smaller than the class of systems supporting computer simulation, there are still more computing mechanisms than a gripping computationalist hypothesis would likely countenance. DNA replication mechanisms will count as computing, as will gumball machines. Relevance is also a worry. The syntactic account is well suited to describe computation in finite state automata, Turing machines and digital computers. But there is good reason to wonder whether such a rigidly specified syntactic account can find application in the wetware of the brain. As Shagrir argues, some of the central successes of computational neuroscience, such as Robinson's (1989) network model of the oculomotor integrator, would not count as computing mechanisms, for the simple reason that the system's inputs and outputs have continuous values and so are not well-defined strings in any sense. The fictionalist will, therefore, likely object that this notion is at once too permissive (in countenancing too many systems to be interesting) and too restrictive (in limiting the application of the term 'computation' to an overly restrictive class of system).

A second debate about computation in neuroscience concerns whether computational explanations have a distinctive kind of explanatory value over and above non-computational explanations. Rather than survey the vast range of positions on this matter, we hope to motivate a view of computational explanations in neuroscience that fits neatly into the mechanistic perspective discussed in Sections 2: computational models that shoulder explanatory weight do so in virtue of being mechanistic explanations.

There is a widespread intuition that computational explanation is fundamentally distinct from mechanistic explanation. Marr (1982) expressed this view in his famous tri-level framework for cognitive science, in which information-processing and computational levels are independent of lower levels of analysis. Sejnowski, Koch, and Churchland (1988) likewise insist that 'Mechanical and causal explanations of chemical and electrical signals in the

brain are different from computational explanations' (p. 1300). They explain: 'The chief difference is that a computational explanation refers to the information content of the physical signals and how they are used to accomplish a task' (p. 1300). The mechanist will argue that one can consistently agree about the distinctiveness of computational models, and yet disagree that the form of explanation they provide is essentially distinct from other more familiar mechanistic explanations.

From a mechanistic perspective, computational explanations must satisfy the same criteria of adequacy as other mechanistic explanations (see Section 2). In particular, they should respect the realistic assumptions built into 3M: components and transformations in the computational model should map onto components and activities in the mechanism. However, mechanists need not deny that computational descriptions are distinctively valuable as interpretive tools: computational explanations of neural systems help us to keep track of the interpretive mapping between internal states of the system and mathematically representable quantities in the world (as Shagrir 2006 argues). However, mechanists insist that providing such informational mappings is insufficient for computational models to have explanatory force. After all, one might develop such mappings, not on the basis of components in the system that produce the target behaviour or cognitive function in question, but on the basis of components whose states are merely correlated with aspects of the target behaviour or cognitive function. For the mechanist, computational explanations are explanatory to the extent that they describe the causal structure of a system. Computational interpretation without causation is not explanation. Furthermore, one can explain information processing tasks mechanistically, without reference to representational states. Such explanations are, in many cases, more readily intelligible than uninterpreted mechanistic explanations would be, but hard-to-understand explanations are still explanations.

The mechanist will thus disagree with Shagrir's understanding of the role of computation in providing explanations: '[W]e adopt the computational approach because we seek to explain how a semantic task can be carried out, and computational explanations are able to do this. But what is the source of this explanatory force? Why are computing mechanisms able to explain semantic tasks? I suggest that the explanatory force of a computing mechanism derives from its correspondence to mathematical relations between the represented objects and states' (Shagrir 2006, p. 394). Mechanists grant that Shagrir identifies crucial features warranting the computational interpretation of a system (that its states map onto mathematical relations in the world). However, they argue that the explanatory force of such models lies not in the interpretation the model supports, but in the causal structure the model correctly describes.

7. Mechanism and the Unity of Neuroscience

We close on the topic with which we began. Neuroscience is a multi-field discipline with (as yet) no grand unifying theory. If the goal of neuroscience is to elucidate the workings of the mind-brain by bringing together diverse investigators studying different aspects of brain structure and function, what form should we expect this union to take? And what benchmarks measure our progress?

Some take a reductionist viewpoint, arguing that the unity of neuroscience is achieved by assimilating higher-level phenomena to lowest-level phenomena. Bickle (1998, 2003), for example, argues that the success of contemporary neuroscience consists in its ability to reveal cellular and molecular mechanisms for cognitive and mental phenomena. This position is evidenced by the fact that one can often change organism-level behaviours by intervening on genes, molecules and ions. The rest of neuroscience – what lies between behaviour and the cellular and molecular mechanisms – is of tremendous heuristic value in defining which mechanisms are explanatorily relevant, but ultimately, it is the molecular mechanisms that do all the explanatory work. The most predictively adequate and instrumentally useful neuroscience will forge mind-to-molecule linkages directly, leaving out the purely heuristic middle realm of cells and physiological systems. Bickle calls this view 'ruthless reductionism'.

Ruthless reductionism has a cousin in Weber's heteronomy thesis. Weber (2005) argues that biological generalizations have no distinctive explanatory value and that the laws of physiology are evolutionarily contingent. As such, they lack the requisite necessity to underlie explanations. This requisite necessity can be found only at the physical and chemical level.[7] Thus, biology is heteronomous, subject to the external authority of physical law. To explain phenomena in neuroscience, it is necessary to relate physiological phenomena (such as action potentials) to physico-chemical laws (such as Ohm's law and the Nernst equation). Biology merely fixes the conditions on which physical laws operate.

The integrationist alternative to these views accepts that there are adequate explanations at multiple levels of organization and insists that many adequate explanations in neuroscience must span multiple levels. Mechanists tend to be integrationists (Bechtel 2007; Craver 2005; Darden 2006; Machamer et al. 2000). They see the nervous system as composed of nested hierarchies of mechanisms within mechanisms linking behaviours in social environments to the behaviours of molecules and ions via intermediate levels of organization. For example, neuroscientists studying spatial learning in rodents have now found crucial linkages between multiple levels, from the lower-level behaviour of ion channels, through forms of synaptic plasticity among neurons, to the

behaviour of brain regions, to the coordination of social behaviour among conspecifics (Squire and Kandel 2000; see also Bickle 2008).

Integrationists deny the reductionist thesis that biological generalizations lack explanatory force. For them, any generalization that holds up under the conditions of a well-controlled experiment arguably counts as explanatory (Woodward 2003; Craver 2007), regardless of whether the regularity in question is evolutionarily contingent and regardless of whether the generalization has any direct connection with physical or chemical laws. Here, we offer three considerations favouring an integrationist perspective.

First, a reminder: no area of neuroscience is metaphysically fundamental. Neuroscience is a science of the middle range, bottoming out well before the physical bedrock that reductionists hope to someday reach. To the physicist, ion channels are complex and messy biological objects. Yet such channels are at, or near, the rock bottom of today's neuroscientific hierarchy.

Second, neuroscientists have discovered causal generalizations that satisfy the requirements of controlled experiments at multiple levels. Levels of causal regularity are to be expected in a world in which interacting parts are organized into mechanisms, within mechanisms, within mechanisms. Basic features of evolution by natural selection make it likely that nature exhibits a hierarchy of nearly decomposable levels of organization that are explanatorily salient and practically useful (Simon 1969; Steel 2008; Wimsatt 2007). Levels are not merely heuristic guides to the search of molecular mechanisms. They supply veritable handles in the causal structure of the mind-brain that can potentially be used for the purposes of curing diseases, improving function and manipulating the brain for good or for ill. Research at, and across, levels is a crucial guide to making interesting predictions (diagnoses, prognoses) and designing interventions (therapies, drug treatments). Were neuroscience to aim for an endgame of ruthless reduction or heteronomy, it would discourage the search for these higher-level forms of intervention (see, for example, Ramachandran et al. (1995, 1998) work on phantom limbs). If one thinks of intermediate levels of organization as mere shadows of lower-level activity, inefficacious abstractions about the behaviour of molecules, one is less likely to think about designing interventions that take advantage of higher levels of organization.

Finally, integrationist strategies are useful tools for concept development. By integrating descriptions at multiple levels, scientists place their concepts under constraints from other levels of organization. The action potential is taken to be a robust phenomenon, in part, because of how it figures in neurotransmitter release and in the behaviour of networks of neurons. The membrane mechanism that Hodgkin and Huxley hoped to discover is visible as a mechanism only on the assumption that the action potential is a meaningful unit to be explained in the first place. To echo an earlier theme, at the molecular

level, the world is a busy, buzzing confusion. The molecular level has to be segmented into mechanisms by subtracting out irrelevant parts and irrelevant interactions. To do that, one needs to know which higher-level phenomena to take seriously. The effort to bridge levels of organization is, in this sense, a procedure for testing whether the concept is empirically adequate. It is an epistemic benefit of integrationist thinking.

8. Conclusion

Despite exponential growth in the neurosciences over the last several decades, the goals, methods and practices of neuroscience have received scant attention from philosophers of science. This chapter surveys a number of current trends and open debates in the philosophy of neuroscience, and it shows that the mechanistic framework affords a fruitful perspective from which to develop answers to many of the core issues in philosophy of neuroscience, including explanation, methodology, computation and reduction. The 1990s, hailed as the 'Decade of the Brain', has long since come to an end. Given the exciting developments unfolding in philosophy of neuroscience in recent years, the prospects for a decade of the philosophy of the brain are now extremely bright.

Notes

1 See Bechtel et al. (2001), Machamer et al. (2001) and Bickle (2009) for recent collections dealing with these and related issues.

2 Making explicit his original aims, Kelso (1995) states: 'one of the main motivations behind these experiments was to counter the then dominant notion of motor programs, which tries to explain switching (an abrupt shift in spatiotemporal order) by a device or mechanism that contains "switches."' (1995, p. 57). Appearances notwithstanding, Kelso's motivation was not to replace one flawed mechanistic hypothesis with another. In an earlier passage, Kelso rails against *all* models of motor behaviour that invoke underlying neural mechanisms, including those that invoke motor programmes (the stored sequence of instructions to drive the muscles during a movement), asserting that '[a]ny time we posit an entity such as reference level or program and endow it with content, we mortgage scientific understanding.' (1995, pp. 33–4). For Kelso, the HKB model was a full-fledged explanation, even though it does not describe mechanisms.

3 Models such as HKB at most involve behavioural 'components', such as the phase relationship between the index fingers, but these are features of the phenomenon and should not be confused with parts of a mechanism (see Bechtel 1998). Van Gelder makes a similar point about dynamical models of cognition: 'the variables [dynamical models] posit are not low level (e.g., neural firing rates) but, rather, macroscopic quantities at roughly the level of the cognitive performance itself' (1998, p. 619). Hence, the HKB equation resembles other well-known empirical

'laws' of human motor behaviour such as Fitts' law (Fitts 1954), which quantifies the robust empirical relationship between target size and distance, movement speed and accuracy.

4 Many who embrace Boyd's (1991) Homeostatic Property Cluster account of natural kinds share this view, including Kornblith (1995), Wilson (2005) and Wilson et al. (2007). For critical discussion, see Craver (2009).

5 Sullivan (2009) claims that the unity of neuroscience is impossible, owing to the heterogeneity of experimental practices and protocols across and within research laboratories.

6 In computer science, a computable function is defined as a mathematical function in which the mapping relation between input and output can be specified by a rule or algorithm, or step-by-step procedure. See Sipser (2005), or for a seminal treatment, see Turing (1936).

7 Weber (2005) allows that there might be laws at higher levels of organization than physiology, such as at the level of population genetics.

References

Bechtel, W. (1998), 'Representations and cognitive explanations: assessing the dynamicist challenge in cognitive science', *Cognitive Science*, 22, 295–318.

—(2002), 'Aligning multiple research techniques in cognitive neuroscience: why is it important?', *Philosophy of Science*, 69, 48–58.

—(2007), *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. New Jersey: Lawrence Erlbaum.

Bechtel, W. and A. Abrahamsen (2005), 'Explanation: a mechanist alternative', *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36(2), 421–41.

Bechtel, W., P. Mandik, J. Mundale and R. S. Stufflebeam (eds) (2001), *Philosophy and the Neurosciences: A Reader*. Oxford: Blackwell.

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

Beckermann, A., H. Flohr and J. Kim (eds) (1992), *Emergence or Reduction?* Berlin: Walter de Gruyter.

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

—(2003), *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Dordrecht: Springer.

—(2008), 'The molecules of social recognition memory: implications for neuroethics and extended mind', *Consciousness and Cognition*, 17, 468–74.

—(2009), *The Oxford Handbook of Philosophy and Neuroscience*. Oxford: Oxford University Press.

Bogen, J. (2005), 'Regularities and causality; generalizations and causal explanations', *Studies in History and Philosophy of Biology and Biomedical Sciences*, 36, 397–420.

—(2008) 'The Hodgkin-Huxley equations and the concrete model: comments on Craver, Schaffner, and Weber', *Philosophy of Science*, 75, 1034–46.

Boyd, R. (1991), 'Realism, anti-foundationalism and the enthusiasm for natural kinds', *Philosophical Studies*, 61, 127–48.
Boyer, P. (2002), *Religion Explained: The Evolutionary Origins of Religious Thought*. New York: Basic Books.
Bressler, S. L., and J. A. S. Kelso (2001), 'Cortical coordination dynamics and cognition', *Trends in Cognitive Sciences*, 5, 26–36.
Bub, J. (1994), 'Testing models of cognition through the analysis of brain-damaged performance', *British Journal for the Philosophy of Science*, 45, 837–55.
Carruthers, P. (2006), *The Architecture of the Mind: Massive Modularity and the Flexibility of Thought*. Oxford: Oxford University Press.
Chemero, A., and M. Silberstein (2007), 'After the philosophy of mind: replacing scholasticism with science', *Philosophy of Science*, 75, 1–27.
Churchland, P. S. (1986), *Neurophilosophy: Toward a Unified Science of the Mind/Brain*. Cambridge, MA: MIT Press.
Churchland, P. S. and R. Grush (1999), *Computation and the Brain*, in Frank Heil and Robert Wilson (eds) *The MIT Encyclopedia of Cognitive Sciences*. Cambridge, MA: MIT Press, pp. 155–8.
Churchland, P. S., and T. J. Sejnowski (1992), *The Computational Brain*. Cambridge, MA: MIT press.
Cohen, L. G., P. Celnik, A. Pascual-Leone, B. Corwell, L. Falz, J. Dambrosia and M. Honda (1997), 'Functional relevance of cross-modal plasticity in blind humans', *Nature*, 389, 180–3.
Craver, C. F. (2002), 'Interlevel experiments and multilevel mechanisms in the neuroscience of memory', *Philosophy of Science Supplement*, 69, S83–97.
—(2005), 'Beyond reduction: mechanisms, multifield integration, and the unity of science', *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36, 373–96.
—(2006), 'When mechanistic models explain', *Synthese*, 153, 355–76.
—(2007), *Explaining the Brain*. Oxford: Oxford University Press.
—(2008), 'Physical law and mechanistic explanation in the Hodgkin and Huxley model of the action potential', *Philosophy of Science*, 75, 1022–33.
—(2009), 'Mechanisms and natural kinds', *Philosophical Psychology*, 22, 575–94.
Cummins, R. (1985), *The Nature of Psychological Explanation*. Cambridge, MA: MIT Press.
Darden, L. (1991), *Theory Change in Science: Strategies from Mendelian Genetics. Oxford: Oxford University Press.*
—(2006), *Reasoning in Biological Discoveries*. New York: Cambridge University Press.
Dayan, P. and L. F. Abbott (2001), *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Cambridge, MA: MIT Press.
Dray, W. (1957), *Law and Explanation in History*. Oxford: Oxford University Press.
Doyle, D. A., J. Cabral, R. A. Pfuetzner, A. Kuo, J. M. Gulbis, S. L. Cohen, B. T. Chait and R. MacKinnon (1998), 'The structure of the potassium channel: molecular basis of K^+ conduction and selectivity', *Science*, 280, 69–77.
Dretske, F. (1988), *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: MIT Press.
—(1995), *Naturalizing the Mind*. Cambridge, MA: MIT Press.

—(2003), 'What can we infer from double dissociations?', *Cortex*, 39, 1–7.
Egan, F. (1995), 'Computation and content', *The Philosophical Review*, 104, 181–203.
Eliasmith, C., and C. H. Anderson (2004), *Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems*. Cambridge, MA: MIT Press.
Fitts, P. M. (1954), 'The information capacity of the human motor system in controlling the amplitude of movement', *Journal of Experimental Psychology*, 47, 381–91.
Fodor, F. A. (1981), *Representations: Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, MA: MIT Press.
—(1983), *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: MIT Press.
Georgopoulos, A. P., Kalaska, J. F., Caminiti, R. and J. T. Massey (1982), 'On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex', *Journal of Neuroscience*, 2, 1527–37.
Georgopoulos, A. P., Schwartz, A.B. and R.E. Kettner (1986), 'Neuronal population coding of movement direction', *Science*, 233, 1416–19.
Glymour, C. (1994), 'On the methods of cognitive neuropsychology', *British Journal for the Philosophy of Science*, 45, 815–35.
—(2001), *The Mind's Arrows*. Cambridge, MA: MIT Press.
Graziano, M. S. A., C. S. R. Taylor and T. Moore (2002), 'Complex movements evoked by microstimulation of precentral cortex', *Neuron*, 34, 841–51.
Grush, R. (2001), 'The semantic challenge to computational neuroscience', in Machamer, P. K., R. Grush and P. McLaughlin (eds), *Theory and Method in the Neurosciences*. Pittsburgh, PA: University of Pittsburgh Press, pp. 155–72.
Grush, R., P. McLaughlin and P. K. Machamer (eds) (2001), *Theory and Method in the Neurosciences*. Pittsburgh, PA: University of Pittsburgh Press.
Haken, H., J. A. Kelso and H. Bunz (1985), 'A theoretical model of phase transitions in human hand movements', *Biological Cybernetics*, 51, 347–56.
Hardcastle, V. G. and C. M. Stewart (2002), 'What do brain data really show?', *Philosophy of Science*, 69, 72–82.
Hempel, C. G. (1965), *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.
Hille, B. (1992), *Ion Channels of Excitable Membranes* (2nd edn). Sunderland, MA: Sinauer.
Hodgkin, A. L. and A. F. Huxley (1952), 'A quantitative description of membrane current and its application to conduction and excitation in nerve', *Journal of Physiology*, 117, 500–44.
Hubel, D. H. and T. N. Wiesel (1968), 'Receptive fields and functional architecture of monkey striate cortex', *Journal of Physiology*, 195, 215.
Jantzen, K. J., F. L. Steinberg and J. A. S. Kelso (2009), 'Coordination dynamics of large-scale neural circuitry underlying rhythmic sensorimotor behavior', *Journal of Cognitive Neuroscience*, 21, 2420–33.
Jirsa, V. K., A. Fuchs and J. A. S. Kelso (1998), 'Connecting cortical and behavioral dynamics: bimanual coordination', *Neural Computation*, 10, 2019–45.
Kelso, J. A. S. (1995), *Dynamic Patterns: The Self-Organization of Brain and Behavior.* Cambridge, MA: MIT Press.
Koch, C. (1999), *Biophysics of Computation: Information Processing in Single Neurons*. Oxford: Oxford University Press.

Koch, C. and I. Segev (2000), 'The role of single neurons in information processing', *Nature Neuroscience*, 3, 1171–7.
Kornblith, H. (1995), *Inductive Inference and its Natural Ground: An Essay in Naturalistic Epistemology*. Cambridge, MA: MIT Press.
Kuhn, T. S. (1970), *The Structure of Scientific Revolutions* (2nd edn). Chicago, IL: University of Chicago Press.
Lambon Ralph M. A., C. Lowe and T. T. Rogers (2007), 'Neural basis of category-specific semantic deficits for living things: evidence from semantic dementia, HSVE and a neural network model', *Brain*, 130, 1127–37.
Machamer, P., L. Darden and C. F Craver (2000), 'Thinking about mechanisms', *Philosophy of Science*, 67, 1–25.
Machamer, P. K., R. Grush and P. McLaughlin (eds) (2001), *Theory and Method in the Neurosciences.* Pittsburgh, PA: University of Pittsburgh Press.
Marr, D. (1982), *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Henry Holt and Co.
Ojemann, G. A. (1991), 'Cortical organization of language', *Journal of Neuroscience*, 11, 2281–7.
Olton, D. S. (1989), 'Inferring psychological dissociations from experimental dissociations: the temporal context of episodic memory', in H. L. Roediger and F. I. M. Craik (eds), *Varieties of Memory and Consciousness: Essays in Honor of Endel Tulving.* Hillsdale, NJ: Lawrence Erlbaum Associates.
Oppenheim, P. and H. Putnam (1958), 'Unity of science as a working hypothesis', *Minnesota Studies in the Philosophy of Science*, 2, 3–36.
Plaut, D. C. (1995), 'Double dissociation without modularity: evidence from connectionist neuropsychology', *Journal of Clinical and Experimental Neuropsychology*, 17, 291–321.
Piccinini, G. (2007a), 'Computational modelling vs. computational explanation: is everything a Turing machine, and does it matter to the philosophy of mind?', *Australasian Journal of Philosophy*, 85, 93–115.
—(2007b), 'Computing mechanisms', *Philosophy of Science*, 74, 501–26.
Pinker, S. (1999), *How the Mind Works*. New York: W W Norton & Co.
Ramachandran, V. S. and Blakesee, S. (1998), *Phantoms in the Brain*. New York: Quill, William Morrow.
Ramachandran, V. S., D. C. Rogers-Ramachandran and S. Cobb (1995), 'Touching the phantom', *Nature*, 377, 489–90.
Rieke, F., D. Warland, R. de Ruyter van Steveninck and W. Bialek (1999), *Spikes: Exploring the Neural Code*. Cambridge, MA: MIT Press.
Robinson, D. A. (1989), 'Integrating with neurons', *Annual Review of Neuroscience*, 12, 33–45.
Rumelhart, D. E. and J. L. McClelland (1986), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (vols 1 and 2). Cambridge, MA: MIT Press.
Sadato, N., A. Pascual-Leone, J. Grafman, V. Ibañez, M. P. Deiber, G. Dold and M. Hallett (1996), 'Activation of the primary visual cortex by Braille reading in blind subjects', *Nature*, 380, 526–8.
Salzman, C. D., K. H. Britten and W. T. Newsome (1990), 'Cortical microstimulation influences perceptual judgements of motion direction', *Nature*, 346, 174–7.

Salmon, W. (1984), *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.
Schaffner, K. F. (1993), *Discovery and Explanation in Biology and Medicine*. Chicago, IL: University of Chicago Press.
—(2008), 'Theories, models, and equations in biology: the heuristic search for emergent simplifications in neurobiology', *Philosophy of Science*, 75, 1008–21.
Schöner, G. (2002), 'Dynamical systems approaches to neural systems and behavior', in N. J. Smelser and P. B. Baltes (eds), *International Encyclopedia of the Social & Behavioral Sciences*. Oxford: Pergamon, pp. 10571–5.
Schöner, G, and J. A. S. Kelso (1988), 'Dynamic pattern generation in behavioral and neural systems', *Science*, 239, 1513–20.
Searle, J. (1992), *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.
Sejnowski, T. J. (2009), 'Computational methods', in L. H. Squire (ed.), *Encyclopedia of Neuroscience*, pp. 19–22.
Sejnowski, T. J., C. Koch and P. S. Churchland (1988), 'Computational neuroscience', *Science*, 241, 1299–306.
Shagrir, O. (2006), 'Why we view the brain as a computer', *Synthese*, 153, 393–416.
Shagrir, O. (2010), 'Brains as analog-model computers', *Studies in History and Philosophy of Science A*, in press.
Simon, H. A. (1969), *The Sciences of the Artificial*. Cambridge, MA: MIT Press.
Sipser, M. (2005), *Introduction to the Theory of Computation* (2nd edn). Pacific Grove, CA: PWS Publishing Co.
Squire, L. R. and E. R. Kandel (2000), *Memory: From Mind to Molecules*. New York: Scientific American Library.
Steel, D. (2008), *Across the Boundaries: Extrapolation in Biology and Social Science*. Oxford: Oxford University Press.
Stephan, A. (2006), 'Emergentism, irreducibility, and downward causation', *Grazer Philosophische Studien*, 65, 77–93.
Sullivan, J. A. (2009), 'The multiplicity of experimental protocols: a challenge to reductionist and non-reductionist models of the unity of neuroscience', *Synthese*, 167, 511–39.
Turing, A. M. (1936), 'On computable numbers, with an application to the Entscheidungsproblem', *Proceedings of the London Mathematical Society*, 42, 230–65.
Uttal, W. R. (2003), *The New Phrenology: The Limits of Localizing Cognitive Processes in the Brain*. Cambridge, MA: MIT Press.
Van Gelder, T. (1998), 'The dynamical hypothesis in cognitive science', *Behavioral and Brain Sciences*, 21, 615–28.
Van Orden, G. C., B. F. Pennington and G. O. Stone (2001), 'What do double dissociations prove?', *Cognitive Science*, 25, 111–72.
Von Neumann, J., P. M Churchland and P. S. Churchland (2000), *The Computer and the Brain*. New Haven, CT: Yale University Press.
Weber, M. (2005), *Philosophy of Experimental Biology*. Cambridge: Cambridge University Press.
—(2008), 'Causes without mechanisms: experimental regularities, physical laws, and neuroscientific explanation', *Philosophy of Science*, 75, 995–1007.
Wilson, R. A. (2005), *Genes and the Agents of Life*. Cambridge: Cambridge University Press.

Wilson, R. A., M. Barker and I. Brigandt (2007), 'When traditional essentialism fails: biological natural kinds', *Philosophical Topics*, 35, 189–215.
Wimsatt, W. C. (2007), *Re-engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Cambridge, MA: Harvard University Press.
Woodward, J. (2003), *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.
Yu, F. H. and W. A. Catterall (2003), 'Overview of the voltage-gated sodium channel family', *Genome Biology*, 4(3), 207.