5 Churchland on Connectionism
AARRE LAAKSO AND GARRISON W. COTTRELL

INTRODUCTION

Paul Churchland cemented his appointment as Ambassador of Connectionism to Philosophy with the 1986 publication of his paper "Some reductive strategies in cognitive neurobiology." However, as Churchland tells the story in the preface to his collection of papers, *A Neurocomputational Perspective*, his relationship with connectionism began three years earlier, when he became acquainted with the model of the cerebellum put forward by Andras Pellionisz and Rodolfo Llinas (1979). The work of Pellionisz and Llinas foreshadows many of the arguments that Churchland makes. They argue that functions of the brain are represented in multidimensional spaces, that neural networks should therefore be treated as "geometrical objects" (323), and that "the internal language of the brain is vectorial" (330). The Pellionisz and Llinas paper also includes an argument for the superiority of neural network organization over von Neumann computer organization on the grounds that the network is more reliable and resistant to damage, a theme to which Churchland often returns.

Over the years, Churchland has applied connectionism to several areas of philosophy, notably: philosophy of mind, epistemology, philosophy of science, and ethics. Churchland's arguments in these areas have a common structure. First, he shows that the predominant positions in the field are (a) based on an assumption that the fundamental objects of study are propositions and logical inferences, and (b) have significant internal difficulties largely attributable to that assumption. Second, he presents a re-construal of the field based on connectionism, giving a "neurocomputational perspective" on the fundamental issues in the field. Finally, he argues that his connectionist alternative fares better than the predominant position on a number of criteria, and explores its further consequences. This is certainly not a formula, since Churchland always considers the particulars of each field, but it is a pattern.

In this paper, we explicate these arguments in a little more detail, and try to give some indication of what we think Churchland has right and what he has wrong. The explication is brief because we are attempting to cover, in just a few pages, topics about which Churchland has written hundreds of pages. The idea is simply to give enough context that the reader may understand the gist of Churchland's arguments and the role that certain problematic claims play in his larger program.

In evaluating Churchland's relation to connectionism, we attempt to skirt the more obviously and centrally *philosophical* issues and concentrate more on the *empirical* issues. Hence, we focus more on what Churchland claims about connectionism (and, in part, about cognitive neuroscience and cognitive psychology) than on what he claims about philosophy of mind, epistemology, philosophy of science, or ethics. This is partly an attempt to balance the perspectives taken on Churchland's work in this volume, and partly a broader attempt to balance the perspectives taken on Churchland's work in the literature as a whole. The distinction between philosophical and empirical issues is indeterminate, however. Churchland is a naturalist – he believes that philosophy and science are continuous – and this is evident in his writing. A critical part of his thesis is that certain claims about connectionism (about its properties, its uses and applications, and its consequences) are in themselves claims about philosophical issues. Moreover, we agree with him – and with Quine (1951) – that there is no fundamental distinction between philosophy and science, that rather there is a continuum of issues from those most peripheral and amenable to change to those most central and resistant to change. So, while we attempt to remain near the edges of the web of belief, we sometimes inevitably slip closer to the center.

We do not cover the basic principles of connectionist networks, such as what a hidden unit is, or what backpropagation means. These topics are dealt with in depth in many other works – see, for example, Ballard (1999), Bishop (1995), Hertz, Krogh, and Palmer (1991) – and Churchland himself often gives explanations that are engaging and clear enough for the layperson. Instead, we save our comments for those cases where we feel that Churchland has the facts wrong about connectionism, or has treated it incompletely.

THE PHILOSOPHICAL CONTEXT

In this section, we discuss Churchland's use of connectionism to ground novel approaches to the philosophy of mind, epistemology, philosophy of science, and ethics. In each case, we first present the "classical" or

"sentential" theory, then Churchland's connectionist alternative. We make an effort to highlight those features of Churchland's argument that we will return to later for a more critical view. At the same time, we try not to be too judgmental at this stage. The goal is to get Churchland's ideas on the table. We then "dig in" to a few choice bits in later sections.

Philosophy of Mind

Churchland contrasts his novel, connectionist philosophy of mind with views that he variously calls "orthodox," "classic," "propositional," and "sentential." Churchland covers a variety of different views with this umbrella, but the most central of them is Fodor's (1975) "language of thought" hypothesis, viz., that thoughts and beliefs are sentences in an internal "language of thought"; thoughts are occurrent sentences (those being actively considered), while beliefs are stored sentences. A second, related, view is that thinking (which Churchland sometimes refers to more generally as "cognitive activity") is performed by updating the occurrent belief sentences by a process of logical inference. A third is that learning is the rule-governed revision of a set of stored (belief) sentences. We will call this view "Propositionalism."

Churchland achieves a certain generality in his arguments by limiting his discussion of the opposing views to these general claims, which (at least directly) imply nothing about the computational architecture of a thinking machine. However, one of Propositionalism's strengths is that it is well-suited for a computational implementation. According to the view that we will call "Computationalism," the sentences that express particular thoughts are actually sequences of tokens in the registers of a von Neumann computer. Likewise, the sentences that express beliefs are sequences of tokens in the memory of such a computer. Thinking and learning are computational processes (programs) that operate on thought and belief tokens in virtue of their form and syntax. The programs implement algorithms that transform thoughts and beliefs according to their form and syntax, while preserving their semantics (meaning and truth values).

For Propositionalists, the affinity of Computationalism and Propositionalism is one of the principle virtues of Propositionalism. It explains how it is possible for machines (including, potentially, biological machines such as human beings) to think. To account for thinking, we need not suppose that there is some sort of nonmaterial substance, or that thought is identical to behavior, or to neural activity, or any of a thousand other problematic things. Rather, thinking is running a program, and we all understand more or less what that is.

For Churchland, on the other hand, the affinity of Computationalism and Propositionalism is one of the principle vices of Propositionalism. The brain is organized very differently from a von Neumann machine. Computationalism is the best implementation account that Propositionalism has to offer, and Propositionalism is therefore completely disconnected from any detailed, neuroscientific account of how the brain actually functions. While the mind might be organized like a computer, the brain is not. There have been attempts to show how the compositional structures essential to symbolic computation might be implemented in a more biologically plausible architecture. Some notable examples are: Cottrell's (1985) implementation of Reiter's default logic for inheritance hierarchies with exceptions in a spreading activation network; Touretzky's (1990) proposal for a method to implement the fundamental Lisp data structure in Boltzmann machines; Smolensky's (1990) proposal for implementing compositional representations as tensor products; Pollack's (1990) recursive auto-associative memory architecture; and Elman's (1991) demonstration that recurrent networks can process complex embeddings in grammatical structure. These and other examples of "implementational connectionism" (e.g., Plate 1995, Derthick 1987, Touretzky & Hinton 1985, Ballard 1986) can be viewed as attempts to demonstrate that an essentially Computationalist model of the mind can be implemented in a connectionist network.

Churchland offers a more radical view: connectionism as an "alternative cognitive paradigm" (Churchland 1989e: xiv), not merely a biologically plausible implementation mechanism for a Computationalist model of the mind but a truly novel model of the mind itself. Where Computationalism takes the computational architecture of cognition to be the von Neumann computer, Churchland takes it to be a connectionist network. The claims of his view, which we will call "Connectionism,"[1] fall out of this fundamental change. Where the Computationalist takes thoughts to be instantiated as sequences of tokens in the central processor of a computer, the Connectionist takes thoughts to be instantiated as patterns of activation in the units of a neural network. The Computationalist takes thinking to be instantiated as the transformation of sets of thought tokens according to a program that is sensitive to their structure, whereas the Connectionist takes thinking to be instantiated as the transformation of patterns of activation in the units of a neural network according to the weighted connectivity between them.

The analogy between Computationalism and Connectionism is somewhat more complicated for belief and learning. We have seen that the Computationalist takes beliefs to be instantiated as sequences of tokens

in the memory of a computer. Some Connectionists take beliefs to be instantiated in the weighted patterns of connectivity between the units in a neural network. Churchland, for example, embraces this view of Connectionism in "On the nature of theories" (1990). Other Connectionists take beliefs to correspond to the partitions of activation patterns that the connection weights determine, or, in recurrent networks, the stable patterns of activation – attractors – that are determined by the weights. Churchland reconsiders the question and adopts this latter view in "Learning and conceptual change" (1989a: 234–234). We have also defended the "partitioning of activation space" interpretation of belief in previous work (Laakso & Cottrell 2000). The interpretation of learning in Connectionism also depends on the position one takes with respect to belief. As we have seen, the Computationalist takes learning to be instantiated as the transformation of sets of belief tokens according to a program that is sensitive to their structure. The Connectionist sympathetic to the beliefs-are-weights view takes learning to be instantiated as the updating of the weighted patterns of connectivity between the units in a neural network according to an algorithm that is sensitive to their values. The Connectionist sympathetic to the beliefs-are-partitions view takes learning to be instantiated as the transformation of the partitions (or attractors) in activation space according to an algorithm that is sensitive to the weights that determine those partitions.

There is also an analogy between Propositionalism and what we will call "Vectorialism." Vectorialism is to Connectionism what Propositionalism is to Computationalism. Propositionalism asserts that thoughts are occurrent sentences in an internal "language of thought," whereas Vectorialism asserts that thoughts are vectors in an internal neural activation coding. Propositionalism asserts that beliefs are stored sentences in the "language of thought," whereas Vectorialism asserts that beliefs are matrices of connection weights – or equivalence classes of activation vectors, depending on one's view of what constitutes belief in a connectionist network – in an internal neural connectivity coding. Propositionalism asserts that thinking is logical inference, whereas Vectorialism asserts that thinking is the changing of activation vectors by matrix multiplication and nonlinear transformations. Propositionalism asserts that learning is the rule-governed revision of a set of beliefs, whereas Vectorialism asserts that learning is the mathematically governed revision of a matrix of connection weights – or of a set of equivalence classes of activation vectors, again depending on one's view of what constitutes belief in a connectionist network. Vectorialism can also be stated in an alternative geometric and kinematic language, one that Churchland sometimes uses. That is, thoughts (activation vectors) may also

be conceptualized as points in activation space; beliefs (weight matrices) may also be conceptualized as points in weight space; thinking (updating activation vectors) may also be conceptualized as motion in activation space; and learning (updating weight matrices) may also be conceptualized as motion in weight space.

We have coined the term Vectorialism because there is no widely used term for the view we have just described. It is possible to be a Connectionist without being a Vectorialist, as the examples of "implementational connectionism" that we mentioned above demonstrate. Churchland sometimes uses the term "state-space semantics" to encompass this and other parts of his view, and we have followed him in previous work (Laakso & Cottrell 2000). However, the term "semantics" arguably does not apply at this level – Vectorialism is a claim about the form of mental representations, not about their contents. Prinz (this volume) is one of many who have pointed out that "state-space semantics" is a misnomer in the absence of an adequate theory of how vectors in state space get their contents. Hence the need to use another term. Whether Churchland offers an adequate theory of content for state-space semantics is a distinctly philosophical question that we do not address here. For the same reason, we do not consider here the broader question of whether it is possible to offer an adequate theory of content for state space semantics independent of Churchland (but see Cottrell, Bartell, & Haupt 1990, for one example).

These features of the different accounts are presented in Table 5.1, which thus provides a brief summary of Churchland's position and distinguishes it from the Computationalist orthodoxy.

Table 5.1 *Comparison of Propositionalism, Computationalism, Vectorialism, and Connectionism as Approaches to the Philosophy of Mind*

| | Orthodox View | | Churchland's View | |
	Propositionalism	*Computationalism*	*Vectorialism*	*Connectionism*
Thoughts	sentences	symbolic tokens	numeric vectors	activations
Beliefs	sentences	symbolic tokens	numeric matrices/ classes of vectors	connectivity weights/partitions
Thinking	logical inference	algorithmic updating	vector transformations	changing activations
Learning	rule-governed revision	algorithmic updating	matrix transformations/ class changes	weight changes/ partition changes

One of the principal virtues that Churchland sees in Connectionism is its biological plausibility. It seems natural to think of units in a connectionist network as simplistic models of neurons, and connections as simplistic models of synapses. As Churchland writes, Connectionism "puts cognitive theory firmly in contact with neurobiology, which adds a very strong set of constraints on the former, to its substantial long-term advantage" (1990: 98). In a future section, we consider just how biologically plausible connectionism really is, but for now it is safe to say that intuitively it seems quite plausible, certainly much more plausible than the declarative and procedural mechanisms characteristic of "good old fashioned" artificial intelligence (GOFAI).

Churchland sees many virtues in Connectionism besides biological plausibility. One is its natural account of categorization. He notes that connectionist networks such as the rocks-from-mines detector (Gorman & Sejnowski 1988) and NETtalk (Sejnowski & Rosenberg 1987) develop rich, structured internal representations that both enable them to exhibit impressive behavior and correspond to real structure in their input. Churchland often explains these categorization feats in terms of "prototype representations" in the hidden unit activation space.

Another virtue that Churchland sees in Connectionism is its natural account of similarity as corresponding to proximity in state space. He writes, with evident gusto, "a state-space representation embodies the *metrical* relations between distinct possible positions within it, and thus embodies the representation of *similarity* relations between distinct items thus represented" (1986: 299, emphasis in original). He uses this feature of Connectionism to give an account of qualia, considering the example of color in depth, but also with reference to taste, olfaction and audition (1989d: 221).

Speed is another virtue that Churchland sees in Connectionism. Connectionist networks can operate very quickly, because of their massive parallelism. The declarative and procedural programs of GOFAI, on the other hand, can be very slow. Moreover, they are usually not amenable to parallelization. (This is of course part and parcel of their biological implausibility.) When such programs do achieve real-time speeds, it is generally in virtue of exploiting the remarkable speed of modern computing hardware. Neurons do not operate nearly as quickly as transistors, so the rapidity of cognition is achieved by parallelism. Connectionism models this computational strategy more closely than GOFAI.

Another virtue of Connectionism is "functional persistence" (as Churchland usually calls it) or "graceful degradation" (a more common term that Churchland also sometimes uses). The brain is remarkably resilient

to trauma, including injury and disease. There are limits, of course, but the anecdotes of famous clinical cases, like Phineas Gage for example, are remarkable not only because of the highly specific and unusual deficits that they document but also because of the remarkable amount of function that is preserved despite large-scale trauma. Connectionist networks can exhibit a similar resilience: their function is often largely preserved despite a simulated "loss" of some of their units or connections.

Churchland also praises Connectionism for being applicable to non-human animals. This is a consequence of its biological plausibility; since connectionist networks are plausible models of the operation of biological neural networks, and since the fundamental computational principles in biological neural networks are the same across species, Connectionism not only explains human cognition, but also explains cognition in other species. This is a claim that Propositionalism cannot make. As implausible as it is that human beings think by manipulating sentential representations in an internal language of thought, it is even more implausible that nonhumans do so. For some Propositionalists, this is a virtue of their account, because it provides a theory of cognition on which thought is uniquely human (see, for example, Bickerton 1995). For most cognitive scientists, the notion is ludicrous.

Epistemology

Although Churchland mentions epistemology frequently, it is almost always in the context of a broader discussion of either philosophy of science or philosophy of mind. For Churchland, epistemology is essentially a bridge between philosophy of mind and philosophy of science. That is, Churchland's attack on traditional epistemological theories follows immediately from his views on the philosophy of mind; and his views on the philosophy of science are, in turn, grounded in his epistemology. Hence, it is possible to summarize Churchland's Connectionist epistemology rather quickly.

Recall that, on Churchland's Connectionist philosophy of mind, beliefs are not sentences in an internal language of thought but vectors in a high-dimensional connectionist weight space. It follows immediately that *knowledge* is not a set of stored sentences (that happen to be true and justified, or something to that effect), but rather a set of stored connection weights. Similarly, on Churchland's Connectionist philosophy of mind, learning is not a process of rule-governed updating of stored belief-sentences, but a process of mathematically governed updating of stored belief-weights.

Table 5.2 *Comparison of Deductivism and Connectionism as Approaches to the Philosophy of Science*

	Deductivism (Orthodox View)	*Connectionism (Churchland's View)*
Knowledge	sets of sentences	prototypes
Learning	logical inference	changing weighted connectivity
Theories	sets of sentences	prototypes
Explanatory Understanding	logical inference	categorization

Again, it follows immediately that knowledge is not acquired by the rule-governed updating of stored belief-sentences, but by the mathematically-governed updating of stored belief-weights.

Philosophy of Science

As he did in the domain of philosophy of mind, Churchland also contrasts his novel, Connectionist epistemology and philosophy of science with views that he variously calls "orthodox," "classic," "propositional," and "sentential." This is another big umbrella, but the most important of the theories covered by it is the deductive-nomological or hypothetico-deductive view, according to which (a) a theory is a set of propositions, and (b) scientific inference and understanding proceed by logical inference. For brevity, we will refer to this view as Deductivism (see Table 5.2).

Deductivism has a number of well-known logical weaknesses. Among them are the paradoxes of confirmation, the problem of explanatory asymmetry, the problem of irrelevant explanation, and the problem of accidental universals. Specific versions of Deductivism also have certain logical problems, such as the indeterminacy of falsification on Popperian theories and the fact that laws were assigned negligible credibility on Carnapian accounts. Churchland discusses these and other problems with Deductivism in detail (1989d, 1990), so we will not dwell on them here.

Churchland's criticism of Deductivism focuses on its empirical implausibility, above and beyond its logical problems. One of the most important empirical issues with Deductivism is timing. People often come to understand an explanation in a very rapid flash of insight. The nearly instantaneous speed of learning and explanatory understanding seems inconsistent with the hypothesis that explanatory understanding is the outcome

of a lengthy process of logical inference. One part of the inconsistency stems from the fact that, on a Deductivist account, grasping a new concept requires first looking up the *relevant* laws or facts. The relevant laws or facts are presumably discovered by some sort of a search, and searches are notoriously slow. The second part of the inconsistency stems from the fact that, even once the relevant basic laws have been retrieved, the cognizer must then deduce the appropriate conclusion. Logical inference is also a computationally intensive operation, one that frequently requires a great deal of backtracking. In fact, logical inference can itself be viewed as a kind of search. So, on the Deductivist account, understanding and learning require *two* searches: one to locate the relevant premises in the space of all known facts, and another to locate the relevant deduction in the space of all possible inferences from those facts. This makes a mystery of our frequent experience of learning and explanatory understanding as rapid and effortless.

Another empirical issue with Deductivism is that the laws and inferences so painstakingly (and yet so rapidly) found are, often, completely inaccessible to the cognizer. People are generally unable to articulate the laws that underlie explanations of phenomena that they appear to understand. They also are generally unable to perform or recite logical inference to anywhere near the degree of rigor and completeness that Deductivism requires. Nonhuman animals also appear to be capable of some forms of causal understanding, but are presumably incapable of storing propositional representations of laws and performing logical inference on them, let alone articulating the laws and the inferences.

The same is true of young infants, and this gives Deductivism a kind of bootstrapping problem. If learning and understanding are characterized by applying the rules of logical inference to propositional premises, and if young infants can neither store propositional premises nor use the rules of logical inference, then how do they *learn* to do so? Evidently, there must be some other account of learning or development that explains our coming to have the abilities that Deductivism requires. Deductivism, however, gives no clues as to what the other account might be. Even if it did, the idea that there should be two different kinds of learning and understanding (a Deductivist account for adults and a – so to speak – *pre*-Deductivist account for infants and perhaps nonhuman animals) seems inelegant at best.

A final empirical issue with Deductivism is that it provides no account of learning or understanding skills, as opposed to facts. However, knowing-*how* is as much a part of our cognitive armory as knowing-*that*. They are, in fact, interrelated in complex ways, as shown by many studies of context-dependent learning. An explanation of skill learning is particularly

important for the philosophy of science in light of Kuhnian observations that implicit knowledge of appropriate practice is an important part of science (Kuhn 1962). While Kuhn may have overstated the importance of skills, it is now widely acknowledged that some part of scientific understanding consists of acquiring appropriate skills.

Some of the most significant problems with Deductivism are neither logical nor psychological per se, but normative. One of the goals of an account of explanation is to determine when changes to a theory are justified; similarly, one of the goals of an account of knowledge is to determine when learning produces *justified* beliefs. However, Deductivism does not meet these criteria.

For one thing, Deductivism cannot justify massive conceptual change. According to Deductivism, all explanation occurs within a framework limited by basic laws and the rules of inference. The laws of inference justify drawing novel conclusions from the basic laws, but they do not warrant changes to the laws themselves. However, fundamental shifts in the basic explanatory axioms often accompany major advances in explanatory understanding (in science) and learning (in individuals).

Nearly everyone who cares about science agrees that scientific theories should be "simple" and "elegant," but almost no one agrees about what those terms really mean or *why* they are important. While Deductivism can perhaps give an account of what "simplicity" means, it does not explain why it is important in a scientific theory. A simple definition of simplicity in Deductivist terms would be the total number of propositions that are required to state the laws governing some field; a slightly more sophisticated view might consider the total number of terms and logical operators in the laws. Regardless, Deductivism does not provide a means for justifying claims that one theory is superior to another on the grounds of simplicity; it does not explain why simplicity is important.

A corollary of the problem about justifying massive conceptual change is that Deductivism cannot give a realist account of scientific progress. Formally speaking, false premises can form just as good a basis for inference as true ones – no amount of inference alone can distinguish false premises from true. (The same is not true for inconsistent premises, but consistency is a very weak normative standard.) However, Deductivism offers no grounds for justifying one set of laws (premises) over another above and beyond their capacity to generate (by logical inference) statements that are true by observation. It was just this property that led us to say that Deductivism provides no means for justifying massive conceptual change, that is, no means for justifying revision of the premises that serve as laws. In much the same way, Deductivism provides in itself no grounds for preferring one

set of fundamental laws (premises) over another. As far as Deductivism is concerned, a false set of premises that is consistent with observation is just as good as a true set of premises. Deductivism alone provides no grounds for preferring true theories over false ones.

Much as Churchland offered Connectionism as an "alternative cognitive paradigm" in the philosophy of mind, so he offers Connectionism as (what we might call) an alternative explanatory paradigm in the philosophy of science. The idea is that explanatory understanding should be thought of not as a product of arriving at a new logical inference but as a product of learning a new category – that a person's grasping a scientific explanation can be modeled by a connectionist network categorizing its input or, equivalently (as Churchland sees it) activating a prototype vector. Of course, coming to understand a scientific theory is *more* than just making a category judgment. It is, among other things, learning to understand a wide variety of things in a certain way and coming to see commonalities among those things, including ones you have never seen before, that you would not otherwise have grasped. In learning a new category, however, a connectionist network does more than simply label the things that it has already been exposed to; it develops an internal representation that can be used to classify new things it has never seen before, and that can potentially be used in other ways in other sorts of computations. Connectionist representations "amplify" knowledge by supporting generalization and complex, context-sensitive processing (1989d: 212).

Churchland argues that Connectionism in the philosophy of science overcomes many of the problems with Deductivism. We have seen that Deductivism offers no explanation of why simplicity is an explanatory virtue. Connectionism, by contrast, has a natural account of why explanatory simplicity is a virtue: an overly complex connectionist network (one with many more hidden units than are required to categorize inputs effectively) will "memorize" the mapping between its inputs and outputs, and fail to generalize to novel inputs. A sufficiently simple connectionist network (one with just enough hidden units to categorize inputs effectively) will achieve both acceptable performance on known inputs *and* effective generalization to novel inputs. An overly simple connectionist network (one with too few hidden units) will be unable to learn to categorize its inputs effectively. Hence, Connectionism can explain why simplicity is a virtue in a scientific explanation: it allows for better generalization to future observations. Connectionism can also explain why too much simplicity is undesirable: there is a natural tradeoff between accurately describing known observations and accurately predicting new observations.

Connectionism also applies to many more types of explanation than Deductivism. We have seen that the Deductivist account of explanatory understanding does not fit scientific (causal) explanations particularly well, but there are many other types of explanations that it does not fit at all. Deductivism offers no account of inductive explanation, for example, or of moral, ethical, or legal explanation. Connectionism, on the other hand, provides a very general account of explanation as a process of concept formation, and therefore applies just as well to these other sorts of explanations as it does to scientific explanation.

A Connectionist account of explanation has other virtues as well. It accounts for our nearly instantaneous grasp of new explanations by the rapidity of parallel processing. It explains our inability to articulate laws or appreciate extended deductive arguments (because we are not using them). It also avoids many technical difficulties with the Deductivist account of scientific explanation – such as the problems of explanatory asymmetry, irrelevant explanations, and accidental universals – which have puzzled philosophers of science for decades.

Churchland admits that his Connectionist account of explanatory understanding does not provide a full account of what explanation itself means. That, however, is not his goal. For Churchland, the challenging question is how cognitive beings come to understand scientific explanations, not what explanations "really are" in some metaphysical sense.

Churchland also draws some broader morals from his Connectionist account of explanatory understanding. He claims that viewing explanatory understanding as vector processing rules out the possibility of finding unassailable "observational" foundations on which to ground the rest of science: all observation, indeed, all perception, is conceptual in the sense that it involves the same sort of vector processing operations. There is no "raw input" to the nervous system that has not been transduced by some part of the sensory system, which is a neural network. This also explains the remarkable plasticity of human beings and cognizers in general – because they are neural networks, they can adapt and change the very means by which they conceive things.

Ethics

Churchland also endeavors to draw *moral* conclusions from Connectionism, specifically to use Connectionism to ground an ethics that neither dismisses moral "knowledge" as bias nor grounds it in abstract rationality (1989b, 1995). Conceptually, social and moral knowledge consists in knowing how

to behave in social situations and how to react to moral dilemmas. It develops by learning to categorize social situations and moral questions appropriately using the pattern classification abilities of a connectionist network. Training consists in coming to react appropriately to social situations (to exhibit socially acceptable, if not advantageous, behaviors), according to the lights of the society in which one grows up. This is not merely becoming socialized to the currently prevailing moral platitudes, because there is room for disagreement (activation of different moral categories in different individuals) and for improvement over time (not only on an individual basis but also on a societal basis, as laws are codified and so on).

As we might expect, Churchland contrasts his Connectionist position with an orthodoxy that explains moral knowledge in terms of a set of sentential rules. The major traditions in ethics may be distinguished by the nature of the rules that they posit and the sources of moral authority that they acknowledge. All such traditions both prescribe and proscribe behavior according to some set of laws. On Churchland's view, by contrast, moral behavior is not prescribed by a set of laws but is is caused by a set of prototypes of "good" and "bad" behavior. Of course, to represent it as a single binary opposition is to oversimplify it drastically. Still, the point remains: for Churchland, ethical and social guidelines are prototypes, not rules. For Churchland, moral disagreements are typically not disagreements over what set of moral rules to follow, but rather, over which moral prototype most closely matches the present situation.

SOME EMPIRICAL ISSUES

Having laid out the basic philosophical context in which Churchland positions his work and explained the main uses to which he puts Connectionism, we now turn to some empirical issues raised by Churchland's claims. We begin with his claim that Connectionism has the virtue of providing a natural account of semantic similarity, in terms of proximity in activation space. In the following section, we consider Churchland's identification of volumes in the hidden-unit activation space of connectionist networks with "prototypes." Finally, we consider whether connectionism really is as biologically plausible as Churchland claims it is.

Similarity

As we noted in the previous section, one of the principle virtues that Churchland sees in Connectionism is a natural account of similarity. On

Churchland's account, perceptual and conceptual similarity is distance between activation vectors in a connectionist network. To determine how similar A is to B, we measure the hidden unit activations used to represent A and the hidden unit activations used to represent B, and then we calculate the distance between them.

A natural first question about this approach is: distance according to what metric? There are many ways of measuring distance, and even more ways of measuring dissimilarity of vectors. The "standard" Euclidean distance between two n-dimensional vectors $\vec{x}, \vec{y} \in \mathbb{R}^n$:

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{1}$$

is only one of a family of norms known as the Minkowski metrics:

$$d = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} \tag{2}$$

for $p = 1, 2, \ldots$, each one of which defines a different possible measure of dissimilarity. Note furthermore that for the special case where $p = \infty$:

$$\lim_{p \to \infty} \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p} = \max |x_i - y_i| \tag{3}$$

which is (qualitatively) yet another measure of dissimilarity.

Each one of these satisfies the mathematical definition of a *metric*, which is a general formalization of what we intuitively call "distance." Specifically, a metric must satisfy the triangle inequality (the distance from X to Y to Z is never shorter than from X directly to Z), be symmetric (the distance between X and Y must be the same as the distance between Y and X), be nonnegative, and satisfy the identity of indiscernibles (if the distance between X and Y is 0, then X and Y must be the same) and its converse (the distance between X and itself must be 0). The latter three conditions (nonnegativity, the identity of indiscernibles, and its converse) are sometimes together called *minimality*.

Besides the common examples we have given in Equations (1)–(3), there are many other less common metrics that could be defined as well. Consider, for example, the trivial example where we define distance as 0 for identical points and 1 otherwise. We know of no *a priori* reason to prefer any one of these metrics over any other for the purpose of measuring representational

similarity. To be empirically adequate, the choice of metric would need to be based on psychological considerations, that is, the results of experiments probing the properties of the cognitive similarity metric.

Those experiments have been done, and it turns out that human semantic similarity judgments do not satisfy the conditions on *any* metric meeting the definition above (Tversky 1977, Tversky & Gati 1978). For one thing, semantic similarity is not symmetric: human subjects reliably judge less prominent things to be more similar to more prominent ones than the reverse (e.g., North Korea is more similar to China than China is to North Korea). People judge their friends to be more similar to themselves than they are to their friends (Holyoak & Gordon 1983). It is also possible to find exceptions to minimality. For example, subjects find that the letter *S* is more similar to itself than the letter *W* is to itself, judging by reaction time in a same-different task (Podgorny & Garner 1979). Subjects also find the letter *M* to be more similar to the letter *H* than it is to itself, judging by inter-letter confusions in a recognition task (Gilmore, Hersh, Caramazza, & Griffin 1979).

Semantic similarity also violates the triangle inequality. An apple is similar to a banana (their "similarity distance" is short, because they are both edible fruits), and a banana is similar to a boomerang (their "similarity distance" is also short, this time because they have similar shapes). Hence, by the triangle inequality, the "similarity distance" between an apple and a boomerang (the "direct route" in this case) should also be short – less than the sum of the distances between apple and banana, on the one hand, and banana and boomerang, on the other. However, the "similarity distance" between an apple and a boomerang is quite large, because they have very little in common. In human subjects, similarity is always judged "with respect to" something – apples are similar to bananas with respect to edibility but not shape, and bananas are similar to boomerangs with respect to shape but not edibility. Humans are able to adjust the features on which they base their similarity judgments depending on the context. Equating similarity with simple distance between activation vectors in a connectionist network affords no analogous ability for adjusting the relative salience of features depending on context: the distance just is what it is.

There is also a question as to which activation vectors should be included in assessing similarity. There are often three levels of activations in a connectionist network (inputs, hidden units, and outputs), and similarity may be assessed at any of these levels, or any combination of them, including all of them simultaneously. The question is even more acute for biological neural networks, which have many layers of processing. Churchland

Table 5.3 *Activations of All Units in a Hypothetical XOR Network*

Input 1	Input 2	Hidden 1 (OR)	Hidden 2 (AND)	Output (XOR)
0	0	0	0	0
0	1	1	0	1
1	0	1	0	1
1	1	1	1	0

often writes as though the relevant activations are *all* of the activations in the network. This can lead to some counterintuitive results. Consider, for example, a feedforward network that computes the Boolean *XOR* function by combining one hidden unit that computes *OR* and one hidden unit that computes *AND*. The activations of the units in this hypothetical example are shown in Table 5.3. The Hamming distances (the sums of the number of different bits in each vector – a metric based on the Minkowski 1-norm described by Equation (2) above, for the case where $p = 1$) between the activations of *all* of the units for each pair of inputs is shown in Table 5.4. Note in Table 5.4 that the network activations for input pattern (0, 1) and the network activations for input pattern (1, 0) are Hamming-distance 2 apart, whereas the network activations for input pattern (0, 0) and the network activations for input pattern (1, 1) are Hamming-distance 4 apart. However, the input patterns (0, 1) and (1, 0) are the same Hamming-distance from each other, as are the input patterns (0, 0) and (1, 1) (i.e., 2 in every case), and the output patterns for (0, 1) and (1, 0) are identical (both 1) as are the output patterns for (0, 0) and (1, 1) (both 0). So, in this case, two pairs of patterns that are equally dissimilar at the inputs and equally similar at the outputs have different overall similarities. This suggests that it is important to consider the layer at which the patterns are compared. If we want to use distance between activation patterns as a similarity metric, then we need to

Table 5.4 *Hamming Distances Between Activations of All Units for All Possible Pairs of Input Patterns in a Hypothetical XOR Network*

Input Pattern A	Input Pattern B (0, 0)	(0, 1)	(1, 0)	(1, 1)
(0, 0)	0	3	3	4
(0, 1)	3	0	2	3
(1, 0)	3	2	0	3
(1, 1)	4	3	3	0

specify which patterns are to be compared; comparing all of them is likely to lead to uninformative results.

There are also differences that do not strictly violate the metric axioms but nevertheless conflict with the properties of common metrics like Euclidean distance. For example, most metrics strictly limit the number of points that can have a common nearest neighbor, whereas human similarity judgments often rate many items as most similar to a single item (Tversky & Hutchinson 1986). In Euclidean space, the maximum number of points that can have the same nearest neighbor[2] i in 1D is 2 (a third point will either have one of the other two as its nearest neighbor, if it falls outside them on the line, or be the nearest neighbor of one of the other two, if it falls inside them on the line). If we disallow ties, then the maximum number of points that can have the same nearest neighbor i in 2D is 5. (The vertices of a regular pentagon with i at the center will all be closer to i than to each other, whereas some of the vertices of a hexagon with i at the center will be at least as close to each other as to i.) In human similarity judgments, by contrast, many items often have the same nearest neighbor (most similar item); in particular, people often associate all (or nearly all) exemplars of a basic-level category most closely with the category itself (Tversky & Hutchinson 1986). For example, in data reported by Rosch and Mervis (1975), subjects rated the category name "fruit" as most related to all but 2 of 20 common instances of fruit (the exceptions being "lemon," which was more related to "orange," and "date," which was more related to "olive"). The fact that human similarity judgments exhibit this sort of dense nearest-neighbor structure, which metric models of similarity cannot capture, suggests that the metric models are incorrect or, at the least, incomplete.

There are nonmetric theories of semantic similarity, as well as more sophisticated metric theories. The non-metric theories include models based on matching features, such as the "contrast model" proposed by Tversky (1977), models based on aligning representations, such as Goldstone's SIAM (1994), and models based on transforming representations, such as that recently advocated by Hahn, Chater, and Richardson (2002). This is not to say that we should give up entirely on accounting for semantic similarity in terms of distance. There are several proposals on offer for basing an account of semantic similarity on distance *with some additional apparatus*, such as spatial density (Krumhansl 1978) or attentional bias (Nosofsky 1991). Hence, it may be possible to defend the claim that semantic similarity corresponds to proximity in activation space in some sense. However, doing so requires some account of how proximity is augmented
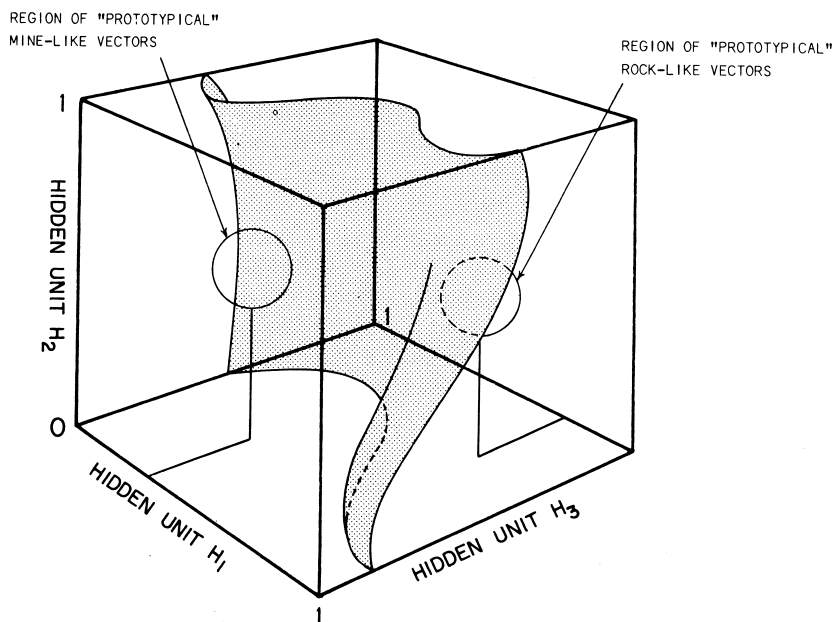
Figure 5.1: Churchland's impression of prototypes in the activation state space of a trained neural network (from "The nature of theories").

in order to adequately model the empirical data. For an excellent review of historical developments and outstanding issues in the study of conceptual similarity, see Goldstone and Son (2005).

Prototypes

We noted several times in the previous section that Churchland often writes about regions in activation state space and *prototypes* as if they are identical. Perhaps the earliest example of this is the following: "under the steady pressure of a learning algorithm . . . the network slowly but spontaneously generates a set of internal representations [that] take the form of a set or system of similarity spaces, and the central point or volume of such a space constitutes the network's representation of a *prototypical* [category member]" (1988, p. 123). The description is consistent with a diagram that Churchland often uses, beginning in "On the nature of theories" (1990), and shown here as Figure 5.1.

It seems clear from Figure 5.1 that Churchland intends us to take his description of a prototype as a "central point or volume" in activation space
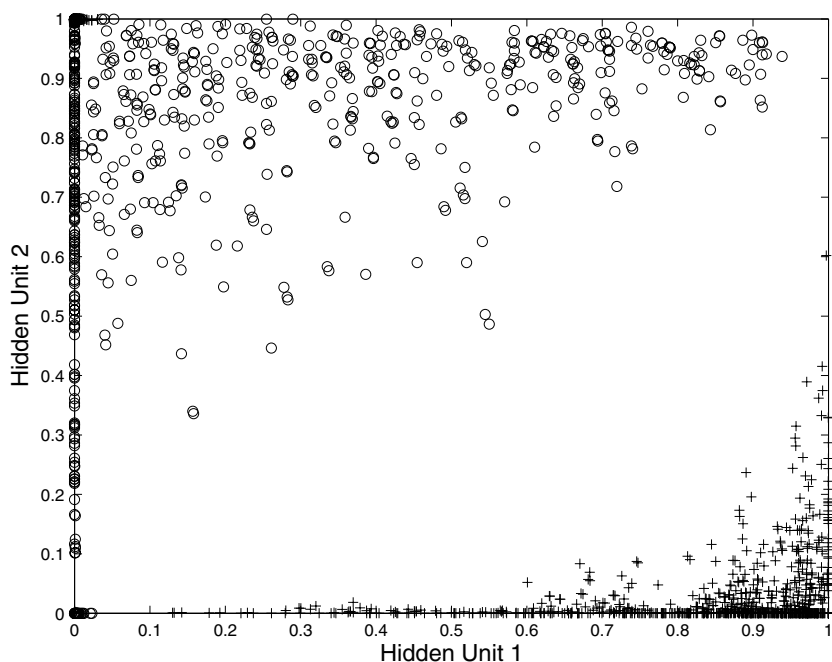
Figure 5.2:   Actual distribution of hidden unit activations in a trained connectionist network.

as literally true. The volumes that he labels as prototypes are indeed at or close to the centers of the regions separated by the hypersurface depicted in the figure. Churchland is not alone – it has become part of philosophical lore that connectionist networks naturally learn and use prototype representations of this kind. Prinz (this volume), for example, asserts that connectionist networks spontaneously form prototypes in activation space of just the sort that Churchland depicts in Figure 5.1.

However, this is not really how connectionist networks work, at least not feedforward networks trained by backpropagation. To demonstrate this, we trained a feedforward network by backpropagation on a classification problem and plotted the actual locations of points in its activation state space, shown in Figure 5.2. The problem was to discriminate poisonous from non-poisonous mushrooms in a hypothetical sample corresponding to 23 species in the Agaricus and Lepiota family, based on 22 nominally valued physical attributes such as the shapes of their caps and the colors of their stalks (Schlimmer 1987a, 1987b). The 22 nominally valued attributes were represented locally in the input by converting them to real values uniformly

distributed in the interval [0, 1]. For example, the "cap shape" attribute, which could have values of "bell," "conical," "convex," "flat," "knobbed," or "sunken" – six possible values – was represented in our inputs by values from the set {0, 0.2, 0.4, 0.6, 0.8, 1.0}. The targets were represented by 0 (edible) or 1 (poisonous). The network had two layers, with two units at the hidden layer and one unit at the output layer, with logistic activation functions at both layers. The training data consisted of 1624 randomly selected patterns, the validation data consisted of a different 1624 randomly selected patterns, and the test data consisted of a third non-overlapping set of 1624 randomly selected patterns. The network was trained by backpropagation on the training data until the mean squared error on the validation data fell below 0.01. In the example shown, this happened after about 80 epochs, and the mean squared error on the test set was 0.0162 after training.

Figure 5.2 shows the activations of the hidden units on the test patterns after training, with the activation vectors representing edible mushrooms marked with circles (o) and those representing poisonous mushrooms marked with plusses (+).

The actual hidden unit activations in Figure 5.2 don't look very much like the hypothesized prototypes in Figure 5.1. Clearly the network depicted in Figure 5.2 has learned to distinguish edible from poisonous mushrooms by distributing their respective hidden-unit activations in such a way that it is easy to separate them. It has, so to speak, "pushed" the edible mushrooms into the upper-left corner of activation space and the poisonous mushrooms into the lower-right corner of activation space, enabling it to "draw a line" between the hidden unit activations, separating the two categories. If there are prototypes in this space, they are in the *corners* of the activation state space, not in the centers of the spaces separated by the discriminating line that the output units draw between, roughly (0.1, 0) and (1.0, 0.6). This is not an accidental artifact of a single renegade run starting with unfortunate random connection weights and winding up in a local minimum. It happens every time the network is trained on this problem, even when the network is afforded even more "extra" room in activation space by giving it three or more hidden units.

In any case, even though the corners of the hidden unit activation space are where backprop "tries to" represent data, it stretches the imagination to construe the corners of such activation space as prototypes. In its ordinary usage in psychology, a prototype is a template for a concept, such that putative exemplars of the concept can be judged according to their similarity to the prototype. It is commonly assumed in the psychological literature that

prototypes are the central tendencies (in the statistical sense, e.g., averages) of their category instances, not only physically but also psychologically, much as Churchland depicts them in Figure 5.1. The prototype represents the best (i.e., most central) instance of the category, and other instances of the category are nearer to or farther from the prototype in psychological space as a function of how similar they are to the prototype.

However, there is nothing to indicate that the network depicted in Figure 5.2 either (a) represents the "best" edible mushroom – the central or average edible mushroom – at or near (0, 1), or (b) interprets distance from (0, 1) as indicating the "degree" of edibility. In general, backpropagation adjusts the weights to the output layer to discriminate the inputs by means of a linear transformation of the hidden unit activations, and adjusts the weights to the hidden layer to maximize the (output layer) discriminant by non-linearly transforming the input patterns into hidden-unit patterns (Bishop 1995: 228). To the extent that activations in the corners of hidden unit activation space are semantically interpretable, then, we could consider them to be the most discriminable exemplars – the ones that are easiest to tell apart. They are psychological *extremes* rather than psychological prototypes.

There is a kind of network, called a *radial basis function network* (RBFN), that more closely resembles Churchland's idea of prototypes in hidden-unit activation space (Bishop 1995). In the standard connectionist models that Churchland usually uses for his examples, units compute a non-linear function (normally a threshold or sigmoid) of the scalar product of the sum of their inputs with a weight. The computation performed by each unit in a such a typical feedforward connectionist network is very straightforward. Each unit j computes a function from $\mathbb{R}^n$ (a vector of the activations of the n units $i_1, \ldots, i_n$ feeding into j) into $\mathbb{R}$ (the activation of unit j), of the form:

$$z_j = g\left(\sum_{i=1}^{n} w_{ij} x_i + w_0\right) \tag{4}$$

where $x_1, \ldots, x_n$ are the activations of the input units $i_1, \ldots, i_n$; $w_{1j}, \ldots, w_{nj} \in \mathbb{R}$ are the weights on the connections from the input units to j; w_0 is a "bias"; and $g : \mathbb{R} \to \mathbb{R}$ is the "activation function," usually the logistic function $g(x) = 1/(1 + e^{-x})$ or the hyperbolic tangent function $g(x) = \tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$. Still simpler models take the activation function to be a threshold function and output a single bit. Most modern models use sigmoid activation functions, however, both because sigmoidal gates have real-valued outputs, giving multilayer networks of sigmoidal gates more computational power than those with threshold gates,

and because the backpropagation learning algorithm requires a differentiable function. Normally, all of the units at a given level of the network compute this function simultaneously. In practice, the activations are calculated serially because of the limitations of simulators running on ordinary hardware. This is irrelevant, however, because the *model* is that they are calculated in parallel, and this is how they are actually calculated when parallel hardware is available.

In a radial basis function network, by contrast, the activation of a hidden unit is calculated by comparing the vector of input activations to a prototype vector. Each hidden unit j in an RBFN computes a function from $\vec{x} \in \mathbb{R}^n$ (an input vector) into $\mathbb{R}$ (the activation of unit j), of the form:

$$z_j(\vec{x}) = \phi(d(\vec{x}, \vec{\mu}_j)) \tag{5}$$

where $\vec{\mu}_j$ is a vector determining the center of the basis function for hidden-layer unit j, and the function $d(\cdot) : \mathbb{R}^n, \ \mathbb{R}^n \to \mathbb{R}$ is a distance function, usually the Euclidean distance (1), between the input vector $\vec{x}$ and the center of the basis function for hidden-layer unit j at $\vec{\mu}_j$. The basis function $\phi(\cdot) : \ \mathbb{R} \to \mathbb{R}$ is usually a spherical Gaussian:

$$\phi(d_j) = \exp\left(-\frac{d_j^2}{2\sigma_j^2}\right) \tag{6}$$

where σ_j is a "width" parameter determining the smoothness of the basis function for hidden-layer unit j.

The activation y_k of an output unit k in an RBFN is a simple linear combination of the basis functions:

$$y_k(\vec{z}) = \sum_{j=1}^{n} w_{jk} z_j \tag{7}$$

In the first stage of training an RBFN, the parameters of the basis function (6) – the centers $\vec{\mu}$ and the widths $\vec{\sigma}$ – are set by an unsupervised learning technique, usually by considering the basis functions to be components of a Gaussian mixture model and optimizing them using the expectation maximization (EM) algorithm. Once the basis function parameters have been fixed, the weights to the output units can be quickly determined using singular value decomposition.

There is a lot to like about RBFNs. They are fast and easy to train. They have nice mathematical properties, such as the fact that their hidden unit activations can be interpreted as the posterior probabilities of the presence of features in the input space, and their output layer weights can
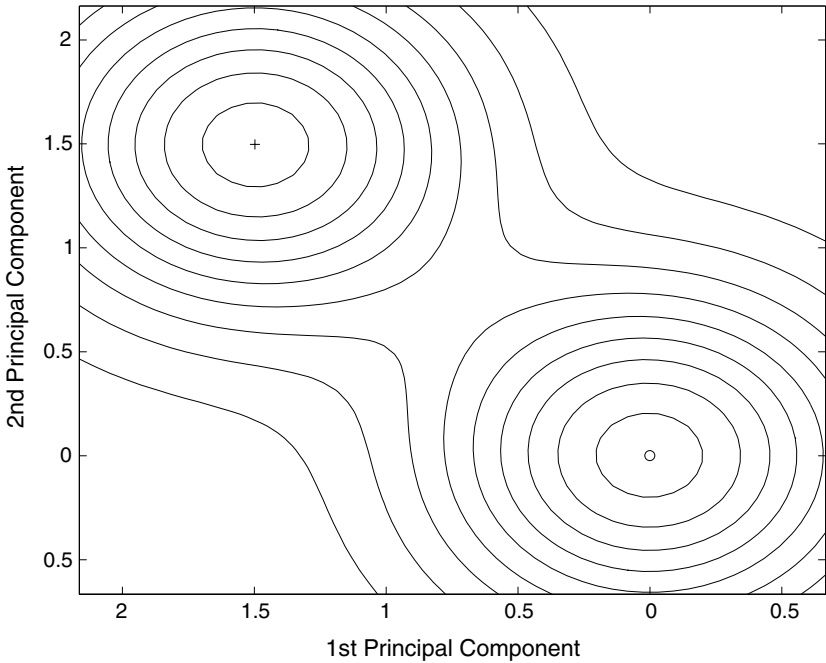
Figure 5.3:   Isoprobability contours for the basis functions in a radial basis function network.

be interpreted as posterior probabilities of membership in the output class given the presence of the features represented at the hidden layer (Bishop 1995).

RBFNs also have nice "psychological" properties. Although it is difficult to find a way to interpret the hidden units of a backprop network as representing prototypes, it is natural to interpret the hidden units of an RBFN as representing prototypes. To demonstrate this, we trained an RBFN with two hidden units and one output unit on the same data used to train the backprop network whose hidden unit activations are shown in Figure 5.2. Because the basis functions are fit to the input data, they have 22 dimensions. To visualize them, we found the first two principal components of their centers and plotted two-dimensional isoprobability curves around them. The resulting graph is shown in Figure 5.3.

Intuitively, the basis functions shown in Figure 5.3 are much more like the "prototypes" that Churchland depicts in Figure 5.1 than the hidden unit activation distributions in Figure 5.2. Their centers represent the "best" examples of edible and poisonous mushrooms, respectively. Furthermore,

distance from the centers represents systematic differences in the system's certainty: the closer an exemplar is to the "edible" center, the more probable that it is edible, and the closer an exemplar is to the "poisonous" center, the more probable that it is inedible.

However, the performance of RBFNs using basis units corresponding to intuitive prototypes may not match what one would like. For example, the RBFN depicted in Figure 5.3 achieved mean squared error of only about 0.4, even after 1000 iterations of EM during the training stage. (Recall that the backpropagation network depicted in Figure 5.2 achieved mean squared error of less than 0.01 after fewer than 100 epochs!) Adding more basis units would certainly improve the performance of the RBFN. However, doing so would entail giving up our interpretation of the basis units as prototypes for edible and poisonous mushrooms. Since each basis unit corresponds to a prototype, adding more basis units means adding more prototypes. It is reasonable to suppose that a human mycologist might have multiple prototypes for edible and poisonous mushrooms (an Edible-A prototype, an Edible-B prototype, a Poisonous-A prototype, and so on). At the extreme, an Agaricus-Lepiota expert probably would have a prototype for each species (a prototype for *Lepiota clypeolaria*, a prototype for *Lepiota procera*, and so on). An RBFN with many basis units might be suitable not only for classifying mushrooms on these sorts of fine-grained distinctions but also for modeling the subordinate (species-level) prototype structure of the human expert. However, there would be no natural way to aggregate the many basis units in such an RBFN into just two prototypes corresponding to the superordinate categories, edible and poisonous. It seems that neither a standard feedforward network trained by backpropagation nor an RBFN is a particularly good model of a prototype theory of human categorization.

On the other hand, prototype theory may not be a good explanation of human categorization. Reaction times in human categorization experiments decrease with the distance in psychological space from the stimulus representation to the decision bound that separates exemplars of contrasting categories; subjects are quicker to categorize stimuli the further they are from the category boundary (Ashby, Boynton, & Lee 1994). The natural prediction of prototype theory, by contrast, is that reaction time should increase with the distance between the stimulus and the nearest prototype. Barsalou (1985, 1991) has pointed out that, at least in the case of some goal-directed categories, the "ideal points" in psychological space may be at the extremes, rather than at the prototypes. That is, people sometimes categorize things based on extremes rather than on

central tendencies. For example, the best exemplar of "drinking water" is one that has zero contaminants, rather than one that has the average amount of contaminants found in a wide sample of potable water. Furthermore, Palmeri and Nosofsky (2001) have shown that prototypes in physical feature space (i.e., the central tendencies of the features of stimuli presented as category instances) may sometimes behave as if they were at extreme points in psychological space relative to other category instances. Although Churchland may have been wrong to assume that the hidden-unit activations of connectionist networks trained on categorization tasks are like prototypes, this evidence suggests at least the possibility that he may have been right to argue that they are good models of human categories.

Biological Plausibility

As discussed in the previous section, there are certain kinds of "connectionist neural networks" that are completely implausible biologically, even though they have excellent mathematical and computational properties, not to mention nice conceptual glosses. However, many philosophers, following Churchland, believe that ordinary feedforward connectionist networks consisting of sigmoidal gating neurons are at least somewhat biologically plausible, if highly abstract. Doubts about the biological plausibility of *back-propagation* are legion in the philosophical literature about Connectionism. So are doubts about the *psychological* plausibility of connectionism. However, just about everyone seems to accept the claim that the networks themselves (if not the supervised training algorithms) are biologically plausible, at least "broadly speaking" (see, for example, Prinz, this volume). In this section, we take a close look at just how broadly we must speak in order to justify the claim that connectionism is biologically plausible.

Churchland himself discusses many biologically implausible aspects of connectionism in some detail (1990: 86–92). The differences that Churchland notes include the fact that biological neural networks are not fully interconnected (as many connectionist models are) and the fact that individual biological neurons generally do not have both inhibitory and excitatory postsynaptic effects (while individual units in connectionist networks may have inhibitory connections to some units and excitatory connections to others). While none of the biologically implausible factors that Churchland recognizes subvert his claim that activation vectors are the fundamental units of representation, there are reasons to believe that this is not true in biological neural networks.

A biological neuron emits short ($\approx 1 - 2$ milliseconds) voltage pulses (with amplitudes of about 100 millivolts) called "action potentials" or "spikes." A spike in a neuron j is triggered by a complex process that starts when spikes in other neurons reach synapses that connect to j. A synapse transforms the spike in the presynaptic neuron i into a postsynaptic potential of about 1 millivolt in j, lasting from 1 millisecond to several seconds or longer. The postsynaptic potential may be excitatory (tend to increase the probability that the postsynaptic cell j fires) or inhibitory (tend to decrease the probability that the postsynaptic cell j fires), depending on whether it increases or decreases the membrane potential of j. The neuron j emits an action potential when some threshold, typically a time-dependent function of membrane potential, is crossed.

A very detailed model of a single neuron – the most famous is the Hodgkin-Huxley model of the giant axon of the squid – attempts to capture this process with as much detail and accuracy as possible. Such models often take into account the equilibrium potential, the specific properties of numerous ion channels (channels with sodium, potassium, calcium, and other ions, most with multiple sub-types operating at different time scales, and some with multiple sub-types sensitive to different voltage thresholds), the different types of synapses (with different neurotransmitters, different receptors, different time scales, and so on), the spatial layout of the dendritic tree (which results in a non-uniform distribution of membrane potential, inducing additional current along the membrane as well as across it), and the specific shape, amplitude, and duration of both postsynaptic potentials and action potentials. (See Gerstner & Kistler 2002 for an excellent review.)

There are simpler, so-called "phenomenological" models of the neuron. Philosophers are likely to find themselves either amused or shocked to find the term "phenomenological" applied to a mathematical model of a single neuron, but they can rest assured that the use of the term in this context is intended neither to beg any important questions (about how consciousness might arise from neural activity) nor to stipulate any sort of philosophical methodology (be it Husserlian or otherwise) toward answering that question. In this usage, which is common in the sciences, "phenomenological" means merely "relating to a phenomenon." A phenomenological model simulates some phenomenon without attempting to capture its underlying causes. Phenomenological models of the neuron attempt to reproduce some aspect of a neuron's behavior (e.g., the timing of spikes) while abstracting away from the biophysical and biochemical details. The sacrifices in accuracy and detail are balanced by gains in simplicity and comprehensibility.

The connectionist model of the neuron as a sigmoid gate described by Equation (4) is a very simple phenomenological model. On the standard interpretation of the correspondence between this model and a biological neuron, the weights $w_1, \dots, w_n$ are identified with the "coupling strengths" of the presynaptic neurons (the efficiency of synapses impinging on j), and the activation of a unit is identified with the firing rate of the neuron. The theory behind this model is that, in the neural code, the "signal" is not carried by the specific times at which individual spikes occur, but rather by the mean rate at which they occur, that is, by the number of spikes generated in some relatively long window. It is important to emphasize that this principle, known as "rate coding" is a *hypothesis* about the neural code.

The rate coding hypothesis has come under significant scrutiny, and several alternatives have been proposed (e.g., Gerstner 2001, Gerstner & Kistler 2002, Maass 1998, Hopfield & Brody 2001, Shastri & Ajjanagadde 1993). In most such models, action potentials are considered to be "stereotyped" events (i.e., they are all equivalent – a spike is a spike is a spike) and are therefore modeled simply as formal events that mark points in time. Coding is hypothesized to take place by the specific timing of spikes, by their phase difference with respect to some periodic signal, or by their temporal cooccurrence or synchrony. The best-known model is called the "leaky integrate-and-fire" model, because it models a neuron as summing its postsynaptic potentials over time (integrating) with some decay (leaking) and "firing" or spiking when its membrane potential exceeds some threshold. In a generalization of the leaky integrate-and-fire model known as the "spike response model" (Gerstner 2001), the membrane potential of neuron j is modeled as:

$$u_j(t) = \sum_{k=1}^{m} \eta \left(t - t_j^k\right) + \sum_{i=1}^{n} \sum_{l=1}^{o} w_{ij} \varepsilon \left(t - t_i^l\right) \qquad (8)$$

where $t_j^1, \dots, t_j^m \in \mathbb{R}$ are the previous firing times of neuron j; $w_{1j}, \dots, w_{nj} \in \mathbb{R}$ are measures of the efficiency of the synapses impinging on j; and $t_i^1, \dots, t_i^o \in \mathbb{R}$ are the previous firing times of the n presynaptic neurons $i_1, \dots, i_n$. The function $\eta(t - t_j^k) : \mathbb{R} \to \mathbb{R}$ determines the voltage contribution to the membrane potential of j at time t that is due to the previous spike of j at time t_j^k. It characterizes the reset of the membrane potential to a resting level immediately after each spike, causing a period of "refractoriness" after each spike that tends to prevent another spike from happening for some time. The function $\varepsilon(t - t_i^l) : \mathbb{R} \to \mathbb{R}$ determines the voltage contribution to the membrane potential of j at time t that is due to

the presynaptic spike on neuron i at time t_i^l. It therefore characterizes the response of j to incoming spikes – the postsynaptic potential.

The neuron j emits a spike when its membrane potential reaches some threshold ϑ Hence, its spiking history $\vec{t}_j$ is updated as follows:

$$\vec{t}_j = \begin{cases} [\vec{t}_j, t] & \text{if } u_j(t) = \vartheta \\ \vec{t}_j & \text{otherwise} \end{cases} \tag{9}$$

In other words, whenever a neuron spikes, the time is added to the neuron's spiking history.

The difference in complexity between Equation (4), on the one hand, and Equations (8) and (9) on the other, is obvious. However, the complexity of the mathematics is not itself an issue. The question is: does the difference really matter with respect to Churchland's position? We believe that it does. In a connectionist network consisting of sigmoidal gating units acting according to Equation (4), the information that a unit contributes to the network is accurately and completely characterized by its current state – its activation z_j. Hence, in a large network consisting of many such neurons (i.e., a connectionist network), it is fair to say that the informational state of the network consists of a vector of the current activations of each of the units. By contrast, in a neural network consisting of spiking neurons acting according to Equation (8), the information that a neuron contributes to the network is *not* accurately and completely characterized by its current state – its membrane potential $u_j(t)$. Rather, the information that the neuron contributes to the network is characterized by its spiking history $\vec{t}_j$, the vector of times at which it emitted an action potential. One could consider the spiking history to be part of the state of a neuron, for example by characterizing it by set of differential equations. However, this entails attributing a greater complexity to the unit itself (the state of the unit consists of the values of at least two equations instead of the one that characterizes a PDP neuron), and therefore jeopardizes the idea that the state of a network can be captured by a single vector.

It is possible to calculate a firing rate from the spiking history by counting the number of spikes in some time window. Hence, the spiking response model embodied in Equation (8) is consistent with the rate coding hypothesis, and, therefore, consistent with Connectionism. However, the rate coding hypothesis itself has come under attack. One of the main reasons for this is that a code based on an average over time is slow, because it requires a sufficiently long period of time to accumulate enough spikes for the average of their intervals to be meaningful. While firing rates in peripheral neurons are relatively fast, the typical firing rates of cortical neurons are

under 100 Hz. If we assume that the rate is 100 Hz, then we would need 50 ms to sample 5 spikes, a reasonable lower bound for a meaningful average. If we assume that classifying a visual stimulus requires 10 processing steps, then it would require 500 ms. However, empirical data shows that human beings can actually classify complex visual stimuli in about 200 ms (Thorpe, Fize, & Marlot 1996). There is also evidence that stochastic fluctuations in the timing of primate cortical action potentials are not simply due to random noise within individual cells, and that the cortical circuitry preserves the fine temporal structure of those fluctuations across many synapses (Abeles, Bergman, Margalit, & Vaadia 1993; Bair & Koch 1996). It has been established physiologically that connections between cortical neurons are modified according to a spike-timing dependent temporally asymmetric Hebbian learning rule (synapses that are active a few milliseconds before the cell fires are strengthened, whereas those that are active a few milliseconds after the cell fires are weakened), and modeling studies have established that this mechanism implements a form of temporal difference learning that could be used for predicting temporal sequences and detecting motion (Rao & Sejnowski 2001, Shon, Rao, & Sejnowski 2004). Finally, there have been theoretical arguments that temporal synchrony is required for binding representations of features together when appropriate (Malsburg 1995), and experimental evidence supports this hypothesis (Singer & Gray 1995). All in all, it seems unlikely that human visual cortex uses rate coding exclusively.

It might be possible to salvage rate coding by interpreting the rate not as the rate at which a single neuron spikes but as the mean rate at which a population of neurons spike. This is sometimes called *population rate coding*. The idea is to determine the mean rate of firing of all the neurons in a small spatial neighborhood over a short time interval. By increasing the number of spikes per unit of time, this approach solves the "slowness" problem with the naive rate coding approach that brings it into conflict with empirical results. In fact, it turns out that neurons coding by a population rate can respond nearly instantaneously to changes in their inputs, under reasonable assumptions (Gerstner 2001).

However, even if population rate coding is a reasonable hypothesis about networks of biological neurons, it is by no means clear how it should be mapped onto connectionist models. If unit activations model population firing rates, then it is no longer reasonable to assume that units correspond to neurons; instead, units must correspond to populations of neurons. Then, connections between units cannot correspond to synapses between (individual) neurons. Perhaps we could consider connections between units to

correspond to overall average connection strength between two populations. However, the populations of the population rate coding hypothesis are defined spatially (by their physical proximity to each other) not topologically (by their connections to each other). Hence, the "connection strength between two populations" must be considered merely a statistical regularity, not a causal mechanism. (Of course, in some cases – namely, when two populations under consideration are not only spatially proximal but also topologically connected – there will be a causal mechanism; our point is only that there need not be one in every case.) As a consequence, it is no longer reasonable to assume (in general) that weights correspond to synaptic efficiencies. It is possible that the analogy between connectionist networks and neural networks could be reconstructed along these lines, but this would require giving a detailed account of how the key elements of connectionist networks (minimally: units, connections, activations, and weights) *do* correspond to features of biological networks using population rate coding. This is an interesting problem, but not one that we can take up here.

In the absence of a defensible mapping between connectionist networks and biological neural networks, it seems only fair to say that connectionist networks are simply *not* biologically plausible. This can be difficult to accept, on two counts: first, the intuitive plausibility of a network of interconnected units modeling a network of neurons; and second, the enormous success that connectionist networks have displayed in modeling, generating, and predicting complex cognitive skills. With respect to the first objection (the intuitive plausibility of connectionism), we can only remind the reader that – as discussed in detail in a previous section – our best current understanding of biological neural networks is potentially inconsistent with some fundamental principles of connectionist modeling. At least in cortex, it appears that neurons may not use rate coding. Instead, their operation may be crucially dependent on the timing of individual spikes, and the history of such timings. With respect to the second objection (the success of connectionism), it is important to note that connectionism's success in modeling complex cognitive abilities is consistent with its biological implausibility. It might be, for example, that both connectionist networks and biological neural networks are calculating statistical properties of their inputs and performing complex probabilistic inferences on them. The fact that connectionist networks perform such calculations in a way that is biologically implausible does not hinder their abilities to do so.

It is still tempting to say that connectionist networks are somehow *more* biologically plausible than Computationalist models. Computationalist

models are primarily declarative and procedural programs executed on von Neumann digital computers. These intuitively *seem* further removed from the brain than connectionist networks. It was possible for a long time to tell a plausible and consistent story about exactly how connectionist networks map to biological neural networks. Even now it is possible to tell a somewhat implausible but still consistent story about how connectionist networks map to biological neural networks. Computationalist models, on the other hand, have never had such a story, and show little promise of generating one anytime soon. So perhaps there is something to the view that, while connectionist networks may not be the most biologically plausible models available, at least they are more biologically plausible than Computational models. It seems not only unreasonable but unnecessary to divide models of cognition into two categories, plausible and implausible, and assert that a model is either one or the other. Rather, biological plausibility is a spectrum, with a wide range between the most plausible models and the most implausible. Computationalism, we might say, is closer to the implausible end of things. While we might once have thought that connectionism was quite clearly on the plausible end of things, we might say that it is somewhere in the middle. Phenomenological models from computational neuroscience, such as the leaky integrate-and-fire model and the spiking response model, are more on the plausible side. Finally, detailed biophysical models, such as the Hodgkin-Huxley model of the giant axon of the squid and more recent models in the same vein, are as plausible as we can be right now.

Churchland's more abstract thesis, which we have dubbed Vectorialism, may fare somewhat better. Recall from Table 5.1 that Vectorialism is the theory that thoughts are vectors that beliefs are matrices, that thinking is transformation of thought-vectors, and that learning is transformation of belief-matrices. In the spiking response model, the state of a neuron is a vector consisting of the times of its previous firings. On a simplistic model of "spike coding," these times themselves would carry information. On one interpretation, the time between the stimulus and the first spike encodes information; on another ("phase coding"), information is carried by the phase of a spike with respect to some other periodic signal; on a third ("correlation coding"), information is encoded by the intervals between the firings of two or more neurons (Gerstner & Kistler 2002). It is not yet known whether or when biological nervous systems use these or other possible coding schemes; deciphering the neural code is an ongoing research project. It is entirely possible that biological neural networks use all of these codes and others not mentioned here and even not yet imagined, and that

the code used might vary from one organism to another, from one system to another within the same organism, and even from one task to another within the same system (Maass 1998). What is important here is that *all* of these coding schemes carry information by quantities that can be represented by numeric vectors. The actual history of spike times for a neuron is a vector, as we saw in Equation (9), so the state of the system could be captured by a matrix of such vectors. Time-to-first-spike is a real number, so the state of the system could be captured by a vector of time-to-first-spikes for all the relevant neurons. Likewise for phase. The correlation coding hypothesis is particularly interesting, because it surmises that information is not encoded in the spikes themselves but rather in their relations. Presynaptic neurons that fire simultaneously communicate to a postsynaptic neuron that they belong together. Of course, this too could be encoded numerically and expressed as a vector. Indeed, Paul Smolensky has shown how to encode Lokendra Shastri's temporal synchrony model of variable binding as a tensor (Tesar & Smolensky 1994).

Reducing the information content of all of these diverse coding schemes to a raw description as "vectors" obscures their important differences and unique properties. There is first of all the fundamental difference between all of the various spike coding hypotheses and the rate coding hypotheses: that the phenomena of interest are points in time rather than rates. Then there is the difference between population coding hypotheses, single-neuron coding hypotheses, and correlational hypotheses, with respect to "how many" neurons are relevant to determining the signal. Finally, there are all of the fine differences between the various spike coding hypotheses. These are important differences, not only for neuroscience *per se* but also for any theory of cognition that wants to make a claim of biological plausibility. Surely, it matters whether the computational units are single neurons, topologically connected combinations (groups) of neurons, or whole (spatially contiguous) populations of neurons. Similarly, it matters whether the phenomena of interest are time series, rates, or some other kind of quantity. Finally, whatever the computational units are, whatever the quanta of information are, it matters how the information is encoded.

Since Vectorialism encompasses all these alternatives, we must ask whether it is too general a view to be of much value. After all, nearly anything can be reduced to a vector or a matrix by a suitable interpretation, and any vector or matrix can be transformed by many mathematical operations. It is even possible to construe Computationalism as Vectorialism, by taking bits in the machine's memory to be vectors of truth values and the logical operations of the CPU to be transformations between such

vectors. Churchland's particular brand of Vectorialism got its teeth from its association with Connectionism. The only charitable way to interpret Churchland's flavor of Vectorialism without making it vacuous is to suppose that the core hypothesis is not that thoughts are vectors *per se*, but that thoughts are vectors-of-activations. Likewise for beliefs (not matrices *per se*, but matrices-of-connection-weights), thinking (not vector transformations *per se*, but transformations-of-activation-vectors) and learning (not matrix transformations *per se*, but transformations-of-weight-matrices). We could call this view "Connectionist-Vectorialism."

The problem with interpreting Vectorialism as Connectionist-Vectorialism is that it then becomes subject to the fate of Connectionism. Specifically, since Connectionism is not all that biologically plausible, neither is Connectionist-Vectorialism. So, if biological plausibility is a desideratum of our theory of mind, Connectionist-Vectorialism does not fit the bill all that well.

How should we proceed from here? One possible approach would be to wait for a single clear victor in the neural coding debate currently being waged in computational neuroscience. One could then build a theory of mind around that hypothesis, much as Churchland has built one around connectionism, and go on to explore its ramifications in other areas like epistemology, philosophy of science, and ethics, again much as Churchland has done with Connectionism. One problem with this sort of "winner take all" strategy for determining our ultimate theory of mind is that it is entirely possible, even likely, that there will be no single winner in the neural coding debate. As we mentioned earlier, it is possible, if not likely, that it will turn out that different organisms, different systems, and even different tasks evoke fundamentally different neural coding schemes.

An even larger problem with the "winner take all" strategy for determining our ultimate theory of mind is that there are all kinds of other models of cognition besides computational neuroscience. Connectionism, for one, shows no signs of going away – despite widespread acknowledgment within the modeling community that it is not very realistic biologically, it continues to be widely used to model cognitive phenomena. Connectionism has also contributed to the explosion of interest in machine learning, statistical learning theory, and information theory in recent years. Machine learning is the study of algorithms for programming machines (computers) to learn to solve pattern recognition, categorization, classification, approximation, estimation, regression, and generalization problems, without concern for biological plausibility. Statistical learning theory is the study of the mathematical nature of those learning problems, in the absence even of any specific

concern for machine implementation. Information theory is the study of the fundamental properties of information, and turns out to have important and interesting links with statistical learning theory. Each of these fields (computational neuroscience, connectionism, machine learning, statistical learning theory, information theory) has made important contributions to the others, and these contributions have flowed in all directions.

To take a single example, Bayesian belief networks (also known as graphical models) arose out of probabilistic extensions to the binary or three-valued logics commonly used in early expert systems. They have since received penetrating statistical analysis, resulting in a solid mathematical foundation. They have also engendered an intense interest in the machine learning community, resulting in efficient exact inference algorithms for special cases and relatively fast approximation algorithms for the general case (see Russell & Norvig 2003 for a review). Psychologists have used them to model many cognitive phenomena. At least one respectable philosopher, Clark Glymour (2001), has asserted that they are a fundamental part of our cognitive architecture and are the mechanism behind our grasp of scientific explanations. These are, of course, much the same claims that Churchland has made about connectionism.

We do not believe that there is any reason to think that Glymour is any more right *or any more wrong* about graphical models than Churchland was about connectionism. Connectionist networks are an instance of graphical models, and *both* frameworks provide useful models of significant domains of cognition. So too do many other models from many other areas, including many other models from psychology, computational neuroscience, connectionism, machine learning, statistical learning theory, and information theory that we have not discussed. The obsession with finding a single theory that "explains the mind" seems to be a peculiarly philosophical affliction. Other fields – including those that are arguably the most centrally engaged in "explaining the mind," cognitive psychology and computational neuroscience – seem quite at home with having multiple models. The driving force behind the philosophers' affliction seems to be a fondness for unity, specifically the unity of science and the unity of explanation. None of the models that we have discussed are nonmaterialistic, nor do they challenge the unity of science as a whole in any other way. Considering them together, in all of their diversity, it is tempting to say that they do not provide a unified explanation of "mind." Taking this to mean that they do not provide a *single* explanation, it is clearly true. Rather than concluding that they must all be subsumed into some higher-level, "more unified" explanation, however, we would argue that the proper response is to conclude that "the mind"

is not a unitary phenomenon. Not only are multiple levels of explanation required to explain cognitive phenomena, so too (at least in some cases) are multiple models required at the same (e.g., computational) level. There does not seem be a single "privileged" perspective (those worshipping at the Church of Bayes notwithstanding). Although the resulting explanation is not unitary, it is consistent.

This is not to say that any and all models are equally good. Freud's model of cognition, for example, was quite bad (even though, like Fortran, it is surprisingly resilient). Models can and must be evaluated on the basis of many criteria, including precision, accuracy, falsifiability, consistency, simplicity, comprehensibility, plausibility, and utility. Some will certainly fare better than others. New and better models will be developed and older and worse models will fall out of use. In the end, there is no particular reason to think that just one will triumph. We think that Churchland would be happy with this conclusion, since it is consistent with both his scientific realism (his view that science is converging on the truth) and his pragmatism (his view that the truth is what works best).

CONCLUSIONS

Churchland's use of connectionism to support novel theories in the philosophy of mind, epistemology, the philosophy of science, and ethics is highly original and thought provoking. It has also had a lasting effect on the field of philosophy, generating many intense exchanges between parties of all philosophical persuasions. In this chapter, we outlined how Churchland has applied connectionism in a variety of philosophical areas, and then discussed several empirical issues with Churchland's interpretation of connectionism. Specifically, we showed that: (1) Churchland's claim that semantic similarity corresponds to proximity in activation space is contradicted by some experimental findings in psychology; (2) Churchland's claim that ordinary connectionist networks trained by backpropagation represent categories by prototype vectors is ill founded, although there are other sorts of connectionist networks that can be interpreted as representing categories by prototypes; and (3) in light of recent developments in computational neuroscience that call the rate coding hypothesis into question, it may turn out that connectionist networks are not very biologically plausible after all.

While making an effort to present Churchland's use of connectionism in context, we have avoided making too much of the more specifically philosophical issues that Churchland addresses. There is certainly

enough criticism of Churchland's philosophy around. At the same time, it is possible that someone will try to use the empirical results we have discussed to advance a more philosophical criticism against Churchland. While Churchland might have his own reservations – either about the empirical conclusions we have drawn here, or about any philosophical uses to which they might be put – we would expect that he would be more excited than dismayed by a philosophical criticism based on empirical data. After all, Churchland is the first truly natural epistemologist. Quine (1951) opened the doors by arguing that natural science *does* matter to philosophy (and vice versa). Churchland was the first to boldly step through those doors and demonstrate how naturalized epistemology could, and should, be done. Even if everything Churchland ever wrote about connectionism and neuroscience should turn out to be utterly wrong, which is unlikely, that legacy will remain.

EPILOGUE

Although this chapter has been critical of some aspects of Churchland's position, the authors would like to end on a personal note.

GWC

Paul Churchland and I found ourselves at the same AI workshop in Austria in 1990. At that conference, Paul gave a talk about how one might model the notion of someone changing their mind. He saw current beliefs as a pattern of activation, and that new activation entering the network would change where the network settled to. He pointed out that it would have to be a recurrent network or this wouldn't work. This is one time when I got an idea from a philosopher that I could act upon, and this idea led to one of my students, Dave Noelle, doing his thesis on learning by being told. I would like to acknowledge Paul for seeding that thesis in my mind!

AL

Paul's work on Connectionism and the reaction to it from other quarters in the philosophical community first grabbed my interest while I was an undergraduate, and they have held it ever since. Paul also inspired my thesis, which was in large part a defense of Churchland's Connectionism against certain objections that Fodor and Lepore had raised.

Notes

1. In this chapter, we use "Connectionism" (with a capital *C*) to refer to the philosophical position that the fundamental architecture of cognition is something like a connectionist network. We continue to use "connectionism" (with a lower-case *c*) to refer to the practice of using such networks in general, where the practitioners are agnostic about the philosophical claim. This distinction parallels our use of the term "Computationalism" to refer to the philosophical position that the fundamental architecture of cognition is something like a digital computer.

2. The question we ask here (a) "How many points can have the same nearest neighbor?" is different from the question (b) "How many points can be each other's nearest neighbors?" to which the answer is 2 points on a line in 1*D*, the 3 vertices of an equilateral triangle in 2*D*, the 4 apexes of a tetrahedron in 3*D*, and so on. It is also different from the question (c) "How many points can be the nearest neighbor of a given point?" to which the answer is 2 points in 1*D* and an infinite number in any higher dimension, arrayed around a circle, a sphere, or a hypersphere. The reason that (a) and (c) are different is that the nearest neighbor relation is not symmetric: the fact that *i* is the nearest neighbor of *j* does not entail that *j* is the nearest neighbor of *i*.

References

Abeles, M., Bergman, H., Margalit, E., & Vaadia, E. (1993). "Spatiotemporal firing patterns in the frontal cortex of behaving monkeys." *Journal of Neurophysiology* **70**(4): 1629–38.

Ashby, F. G., Boynton, G., & Lee, W. W. (1994). "Categorization response time with multidimensional stimuli." *Perception & Psychophysics* **55**(1): 11–27.

Bair, W., & Koch, C. (1996). "Temporal precision of spike trains in extrastriate cortex of the behaving macaque monkey." *Neural Computation* **8**(6): 1185–202.

Ballard, D. H. (1986). Parallel logical inference and energy minimization. *Proceedings of the 5th National Conference on Artificial Intelligence (AAAI-86)* (Vol. 1, pp. 203–9). Philadelphia, Morgan Kaufmann.

Ballard, D. H. (1999). *An introduction to natural computation*. Cambridge, MA, MIT Press.

Barsalou, L. W. (1985). "Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories." *Journal of Experimental Psychology: Learning, Memory, & Cognition* **11**(1–4): 629–54.

Barsalou, L. W (1991). Deriving categories to acheive goals. G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory*, (Vol. 27, pp. 1–64). San Diego, Academic Press.

Bickerton, D. (1995). *Language and human behavior*. Seattle, University of Washington Press.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford, Clarendon Press.

Churchland, P. M. (1986). "Some reductive strategies in cognitive neurobiology." *Mind* **95**(379): 279–309.

Churchland, P. M. (1988). Folk psychology and the explanation of human behavior. *A neurocomputational perspective: The nature of mind and the structure of science* (pp. 111–35). Cambridge, MA, MIT Press/Bradford Books.

Churchland, P. M. (1989a). Learning and conceptual change. *A neurocomputational perspective: The nature of mind and the structure of science* (pp. 231–53). Cambridge, MA, MIT Press/Bradford Books.

Churchland, P. M. (1989b). Moral facts and moral knowledge. *A neurocomputational perspective: The nature of mind and the structure of science* (pp. 297–303). Cambridge, MA, MIT Press/Bradford Books.

Churchland, P. M. (1989c). *A neurocomputational perspective: The nature of mind and the structure of science*. Cambridge, MA, MIT Press/Bradford Books.

Churchland, P. M. (1989d). On the nature of explanation: A PDP approach. *A neurocomputational perspective: The nature of mind and the structure of science* (pp. 197–230). Cambridge, MA, MIT Press/Bradford Books.

Churchland, P. M. (1989e). Preface. *A neurocomputational perspective: The nature of mind and the structure of science* (pp. xi–xvii). Cambridge, MA, MIT Press/Bradford Books.

Churchland, P. M. (1990). On the nature of theories: A neurocomputational perspective. C. W. Savage (Ed.), *Scientific theories* (Vol. 14). Minneapolis, University of Minneapolis Press.

Churchland, P. M. (1995). *The engine of reason, the seat of the soul: A philosophical journey into the brain*. Cambridge, MA, MIT Press/Bradford Books.

Cottrell, G. (1985). Parallelism in inheritance hierarchies with exceptions. *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, Los Angeles, CA.

Cottrell, G. W., Bartell, B., & Haupt, C. (1990). Grounding meaning in perception. H. Marburger (Ed.), *Proceedings of the German Workshop on Artificial Intelligence (GWAI)* (pp. 307–21). Berlin, Springer-Verlag.

Derthick, M. A. (1987). A connectionist architecture for representing and reasoning about structured knowledge. *Proceedings of the Ninth Annual Conference of the Cognitive Science Society* (pp. 131–42). Hillsdale, NJ, Lawrence Erlbaum Associates.

Elman, J. L. (1991). "Distributed representations, simple recurrent networks, and grammatical structure." *Machine Learning* 7: 195–225.

Fodor, J. A. (1975). *The language of thought*. Cambridge, MA, Harvard University Press.

Gerstner, W. (2001). What's different with spiking neurons? H. Mastebroek & H. Vos (Eds.), *Plausible neural networks for biological modeling* (pp. 23–48). Boston: Kluwer.

Gerstner, W., & Kistler, W. M. (2002). *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge, UK, Cambridge University Press.

Gilmore, G. C., Hersh, H., Caramazza, A., & Griffin, J. (1979). "Multidimensional letter similarity derived from recognition errors." *Perception and Psychophysics*, **25**: 425–31.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge, MA, MIT Press.

Goldstone, R. L. (1994). "The role of similarity in categorization: providing a groundwork." *Cognition* **52**: 125–57.

Goldstone, R. L., & Son, J. (2005). Similarity. K. Holyoak & R. Morrison (Eds.), *Cambridge handbook of thinking and reasoning*. Cambridge, UK, Cambridge University Press.

Gorman, R. P., & Sejnowski, T. J. (1988). "Analysis of hidden units in a layered network trained to classify sonar targets." *Neural Networks* **1**: 75–89.

Hahn, U., Chater, N., & Richardson, L. B. (2002). "Similarity as transformation." *Cognition*, **87**: 1–32.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the theory of neural computation*. New York, Addison-Wesley.

Holyoak, K. J., & Gordon, P. C. (1983). "Social reference points." *Journal of Personality and Social Psychology* **44**: 881–7.

Hopfield, J. J., & Brody, C. D. (2001). "What is a moment? Transient synchrony as a collective mechanism for spatiotemporal integration." *Proceedings of the National Academy of Sciences* **98**(3).

Krumhansl, C. L. (1978). "Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density." *Psychological Review* **85**: 450–63.

Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago, University of Chicago Press.

Laakso, A., & Cottrell, G. W. (2000). "Content and cluster analysis: Assessing representational similarity in neural systems." *Philosophical Psychology* **13**(1): 77–95.

Maass, W. (1998). On the role of time and space in neural computation. *Proceedings of the Federated Conference of CLS'98 and MFCS'98* (Vol. 1450, pp. 72–83). Berlin: Springer.

Malsburg, C. von der. (1995). "Binding in models of perception and brain function." *Current Opinion in Neurobiology* **5**: 520–6.

Nosofsky, R. M. (1991). "Stimulus bias, asymmetric similarity, and classification." *Cognitive Psychology* **23**: 94–140.

Palmeria, T. J., & Nosofsky, R. M. (2001). "Central tendencies, extreme points, and prototype enhancment effects in ill-defined perceptual categorization." *Quarterly Journal of Experimental Psychology: Human Experimental Psychology* **54A**(1): 197–235.

Pellionisz, A., & Llinas, R. (1979). "Brain modeling by tensor network theory and computer simulation. The cerebellum: Distributed processor for predictive coordination." *Neuroscience* **4**: 323–48.

Plate, T. A. (1995). "Holographic reduced representations." *IEEE Transactions on Neural Networks* **6**(3): 623.

Podgorny, P., & Garner, W. R. (1979). "Reaction time as a measure of inter-intraobject visual similarity: Letters of the alphabet." *Perception and Pyschophysics* **26**(1): 37–52.

Pollack, J. B. (1990). "Recursive distributed representations." *Artificial Intelligence* **46**(1–2): 77–105.

Quine, W. V. O. (1951). "Two dogmas of empiricism." *Philosophical Review* **60**: 20–43.

Rao, R. P. N., & Sejnowski, T. J. (2001). "Spike-timing-dependent Hebbian placticity as temporal difference learning." *Neural Computation* **13**(10): 2221–37.

Rosch, E., & Mervis, C. B. (1975). "Family resemblances: Studies in the internal structure of categories." *Cognitive Psychology* **7**(4): 573–605.

Russell, S. J., & Norvig, P. (2003). *Artificial intelligence: A modern approach* (2nd ed.). Upper Saddle River, NJ, Prentice-Hall.

Schlimmer, J. S. (1987a). *Concept acquisition through representational adjustment.* Unpublished doctoral dissertation, University of California, Irvine.

Schlimmer, J. S (1987b). Mushrooms dataset. The UCI Machine Learning Repository. (Retrieved August 11, 2004, from ftp://ftp.ics.uci.edu/pub/machine-learning-databases/mushroom).

Sejnowski, T. J., & Rosenberg, C. R. (1987). "NETtalk: Parellel networks that learn to pronounce english text." *Complex Systems* **1**, 145–68.

Shastri, L., & Ajjanagadde, V. (1993). "From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings using temporal synchrony." *Behavioral and Brain Sciences* **16**: 417–94.

Shon, A. P., Rao, R. P. N., & Sejnowski, T. J. (2004). "Motion detection and prediction through spike-timing dependent plasticity." *Network: Computation in Neural Systems* **15**: 179–98.

Singer, W., & Gray, C. M. (1995). "Visual feature integration and the temporal correlation hypothesis." *Annual Review of Neuroscience* **18**: 555–86.

Smolensky, P. (1990). "Tensor product variable binding and the representation of symbolic structures in connectionist systems." *Artificial Intelligence* **46**(1–2): 159–216.

Tesar, B., & Smolensky, P. (1994). Synchronous-firing variable binding is spatio-temporal tensor product representation. A. Ram & K. Eiselt (Eds.), *Proceedings of the 16th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ, Lawrence Erlbaum Associates.

Thorpe, S., Fize, D., & Marlot, C. (1996). "Speed of processing in the human visual system." *Nature* **381**: 520–2.

Touretzky, D. S. (1990). "BoltzCONS: dynamic symbol structures in a connectionist network." *Artificial Intelligence* **46**: 5–46.

Touretzky, D. S., & Hinton, G. E. (1985). Symbols among the neurons: Details of a connectionist inference architecture. *Proceedings of the 9th International Joint Conference on Artificial Intelligence (IJCAI 85)* (pp. 238–43). San Mateo, CA, Morgan Kaufmann.

Tversky, A. (1977). "Features of similarity." *Psychological Review* **84**(4): 327–52.

Tversky, A., & Gati, I. (1978). Studies of similarity. E. Rosch & B. Lloyd (Eds.). *Cognition and categorization* (pp. 79–98). Hillsdale, NJ, Lawrence Erlbaum Associates.

Tversky, A., & Hutchinson, J. W. (1986). "Nearest-neighbor analysis of psychological spaces." *Psychological Review* **93**: 3–22.