Paul M. Churchland

On the Nature of Theories: A Neurocomputational Perspective

I. The Classical View of Theories

Not long ago, we all knew what a theory was: it was a set of sentences or propositions, expressible in the first-order predicate calculus. And we had what seemed to be excellent reasons for that view. Surely any theory had to be stat*able*. And after it had been fully stated, as a set of sentences, what residue remained? Furthermore, the sentential view made systematic sense of how theories could perform the primary business of theories, namely, prediction, explanation, and intertheoretic reduction. It was basically a matter of first-order deduction from the sentences of the theory conjoined with relevant premises about the domain at hand.

Equally important, the sentential view promised an account of the nature of learning, and of rationality. Required was a set of formal rules to dictate appropriate changes or updates in the overall set of believed sentences as a function of new beliefs supplied by observation. Of course there was substantial disagreement about which rules were appropriate. Inductivists, falsificationists, hypothetico-deductivists, and Bayesian subjectivists each proposed a different account of them. But the general approach seemed clearly correct. Rationality would be captured as the proper set of formal rules emerged from logical investigation.

Finally, if theories are just sentences, then the ultimate virtue of a theory is truth. And it was widely expected that an adequate account of rational methodol-

Several pages in section 4 are based on material from an earlier paper, "Reductionism, Connectionism, and the Plasticity of Human Consciousness," *Cultural Dynamics*, 1 (1): 1988. Three pages in section 5 are drawn from "Simplicity: The View from the Neuronal Level." In *Aesthetic Factors in Natural Science*, ed. N. Rescher (forthcoming, 1990). My thanks to the editors for permission to use that material here. For many useful discussions, thanks also to Terry Sejnowski, Patricia Churchland, David Zipser, Dave Rumelhart, Francis Crick, Stephen Stich, and Philip Kitcher.

ogy would reveal why humans must tend, in the long run, toward theories that are true.

Hardly anyone will now deny that there are serious problems with every element of the preceding picture—difficulties we shall discuss below. Yet the majority of the profession is not yet willing to regard them as fatal. I profess myself among the minority that does so regard them. In urging the poverty of 'sentential epistemologies' for over a decade now (Churchland 1975, 1979, 1981, 1986), I have been motivated primarily by the *pattern* of the failures displayed by that approach. Those failures suggest to me that what is defective in the classical approach is its fundamental assumption that languagelike structures of some kind constitute the basic or most important form of representation in cognitive creatures, and the correlative assumption that cognition consists in the manipulation of those representations by means of structure-sensitive rules.

To be sure, not everyone saw the same pattern of failure, nor were they prepared to draw such a strong conclusion even if they did. For any research program has difficulties, and so long as we lack a comparably compelling *alternative* conception of representation and computation, it may be best to stick with the familiar research program of sentences and rules for their manipulation.

However, it is no longer true that we lack a comparably compelling alternative approach. Within the last five years, there have been some striking theoretical developments and experimental results within cognitive neurobiology and 'connectionist' AI (artificial intelligence). These have provided us with a powerful and fertile framework with which to address problems of cognition, a framework that owes nothing to the sentential paradigm of the classical view. My main purpose in this essay is to make the rudiments of that framework available to a wider audience, and to explore its far-reaching consequences for traditional issues in the philosophy of science. Before turning to this task, let me prepare the stage by briefly summarizing the principal failures of the classical view, and the most prominent responses to them.

II. Problems and Alternative Approaches

The depiction of learning as the rule-governed updating of a system of sentences or propositional attitudes encountered a wide range of failures. For starters, even the best of the rules proposed failed to reproduce reliably our preanalytic judgments of credibility, even in the artificially restricted or 'toy' situations in which they were asked to function. Paradoxes of confirmation plagued the H-D accounts (Hempel 1965; Scheffler 1963). The indeterminacy of falsification plagued the Popperian accounts (Lakatos 1970; Feyerabend 1970; Churchland 1975). Laws were assigned negligible credibility on Carnapian accounts (Salmon, 1966). Bayesian accounts, like Carnapian ones, presupposed a given probability space as the epistemic playground within which learning takes place, and they could not account for the rationality of major shifts from one probability space to another, which is what the most interesting and important cases of learning amount to. The rationality of large-scale *conceptual change*, accordingly, seemed beyond the reach of such approaches. Furthermore, simplicity emerged as a major determinant of theoretical credibility on most accounts, but none of them could provide an adequate definition of simplicity in syntactic terms, or give a convincing explanation of why it was relevant to truth or credibility in any case. One could begin to question whether the basic factors relevant to learning were to be found at the linguistic level at all.

Beyond these annoyances, the initial resources ascribed to a learning subject by the sentential approach plainly presupposed the successful completion of a good deal of sophisticated learning on the part of that subject already. For example, reliable observation judgments do not just appear out of nowhere. Living subjects have to *learn* to make the complex perceptual discriminations that make perceptual judgments possible. And they also have to *learn* the linguistic or propositional system within which their beliefs are to be constituted. Plainly, both cases of learning will have to involve some procedure quite distinct from that of the classical account. For that account presupposes antecedent possession of both a determinate propositional system and a capacity for determinate perceptual judgment, which is precisely what, prior to extensive learning, the human infant lacks. Accordingly, the classical story cannot possibly account for all cases of learning. There must exist a type of learning that is prior to and more basic than the process of sentence manipulation at issue.

Thus are we led rather swiftly to the idea that there is a level of representation *beneath* the level of the sentential or propositional attitudes, and to the correlative idea that there is a learning dynamic that operates primarily on sublinguistic factors. This idea is reinforced by reflection on the problem of cognition and learning in nonhuman animals, none of which appear to have the benefit of language, either the external speech or the internal structures, but all of which engage in sophisticated cognition. Perhaps their cognition proceeds entirely without benefit of any system for processing sentencelike representations.

Even in the human case, the depiction of one's knowledge as an immense set of individually stored 'sentences' raises a severe problem concerning the relevant retrieval or application of those internal representations. How is it one is able to retrieve, from the millions of sentences stored, exactly the handful that is relevant to one's current predictive or explanatory problem, and how is it one is generally able to do this in a few tenths of a second? This is known as the "Frame Problem" in AI, and it arises because, from the point of view of fast and relevant retrieval, a long list of sentences is an appallingly inefficient way to store information. And the more information a creature has, the worse its application problem becomes.

A further problem with the classical view of learning is that it finds no essential connection whatever between the learning of *facts* and the learning of *skills*. This is a problem in itself, since one might have hoped for a unified account of learning, but it is doubly a problem when one realizes that so much of the business of understanding a theory and being a scientist is a matter of the skills one has acquired. Memorizing a set of sentences is not remotely sufficient: one must learn to *recognize* the often quite various instances of the terms they contain; one must learn to *manipulate* the peculiar formalism in which they may be embedded; one must learn to *apply* the formalism to novel situations; one must learn to *control* the instruments that typically produce or monitor the phenomena at issue. As T. S. Kuhn first made clear (Kuhn 1962), these dimensions of the scientific trade are only artificially separable from one's understanding of its current theories. It begins to appear that even if we do harbor internal sentences, they capture only a small part of human knowledge.

These failures of the classical view over the full range of learning, both in humans and in nonhuman animals, are the more suspicious given the classical view's total disconnection from any theory concerning the structure of the biological *brain*, and the manner in which it might *implement* the kind of representations and computations proposed. Making acceptable contact with neurophysiological theory is a long-term constraint on any epistemology: a scheme of representation and computation that cannot be implemented in the machinery of the human brain cannot be an adequate account of human cognitive activities.

The situation on this score used to be much better than it now is: it was clear that the classical account of representation and learning could easily be realized in typical digital computers, and it was thought that the human brain would turn out to be relevantly like a digital computer. But quite aside from the fact that computer implementations of sentential learning chronically produced disappointing results, it has become increasingly clear that the brain is organized along computational lines radically different from those employed in conventional digital computers. The brain, as we shall see below, is a massively parallel processor, and it performs computational tasks of the classical kind at issue only very slowly and comparatively badly. To speak loosely, it does not appear to be designed to perform the tasks the classical view assigns to it.

I conclude this survey by returning to specifically philosophical matters. A final problem with the classical approach has been the failure of all attempts to explain why the learning process must tend, at least in the long run, to lead us toward *true* theories. Surprisingly, and perhaps distressingly, this Panglossean hope has proved very resistant to vindication (Van Fraassen 1980; Laudan 1981). Although the history of human intellectual endeavor does support the view that, over the centuries, our theories have become dramatically *better* in many dimensions, it is quite problematic whether they are successively 'closer' to 'truth'. Indeed, the notion of truth itself has recently come in for critical scrutiny (Putnam 1981; Churchland 1985; Stich 1990). It is no longer clear that there *is* any unique and unitary relation that virtuous belief systems must bear to the nonlinguistic world. Which leaves us free to reconsider the great many different dimensions of epistemic and pragmatic virtue that a cognitive system can display.

The problems of the preceding pages have not usually been presented in concert, and they are not usually regarded as conveying a unitary lesson. A few philosophers, however, have been moved by them, or by some subset of them, to suggest significant modifications in the classical framework. One approach that has captured some adherents is the 'semantic view' of theories (Suppe 1974; Van Fraassen 1980; Giere 1988). This approach attempts to drive a wedge between a theory and its possibly quite various linguistic formulations by characterizing a theory as a *set of models*, those that will make a first-order linguistic statement of the theory come out *true* under the relevant assignments. The models in the set all share a common abstract structure, and that structure is what is important about any theory, according to the semantic view, not any of its idiosyncratic linguistic expressions. A theory is true, on this view, just in case it includes the actual world, or some part of it, as one of the models in the set.

This view buys us some advantages, perhaps, but I find it to be a relatively narrow response to the panoply of problems addressed above. In particular, I think it strange that we should be asked, at this stage of the debate, to embrace an account of theories that has absolutely nothing to do with the question of how real physical systems might embody representations of the world, and how they might execute principled computations on those representations in such a fashion as to learn. Prima facie, at least, the semantic approach takes theories even farther into Plato's Heaven, and away from the buzzing brains that use them, than did the view that a theory is a set of sentences. This complaint does not do justice to the positive virtues of the semantic approach (see especially Giere, whose version does make some contact with current cognitive psychology). But it is clear that the semantic approach is a response to only a small subset of the extant difficulties.

A more celebrated response is embodied in Kuhn's *The Structure of Scientific Revolutions* (1962). Kuhn centers our attention not on sets of sentences, nor on sets of models, but on what he calls paradigms or exemplars, which are specific *applications* of our conceptual, mathematical, and instrumental resources. Mastering a theory, on this view, is more a matter of being able to perform in various ways, of being able to solve a certain class of problems, of being able to recognize diverse situations as relevantly similar to that of the original or paradigmatic application. Kuhn's view brings to the fore the historical, the sociological, and the psychological factors that structure our theoretical cognition. Of central importance is the manner in which one comes to perceive the world as one internalizes a theory. The perceptual world is redivided into new categories, and while the theory may be able to provide necessary and sufficient conditions for being an instance of any of its categories, the perceptual recognition of any instance of a category does not generally proceed by reference to those conditions, which often transcend perceptual experience. Rather, perceptual recognition proceeds by some inarticulable process that registers *similarity* to one or more perceptual *prototypes* of the category at issue. The recognition of new applications of the apparatus of the entire theory displays a similar dynamic. In all, a successful theory provides a prototypical beachhead that one attempts to expand by analogical extensions to new domains.

Reaction to this view has been deeply divided. Some applaud Kuhn's move toward naturalism, toward a performance conception of knowledge, and away from the notion of truth as the guiding compass of cognitive activity (Munevar 1981; Stich 1990). Others deplore his neglect of normative issues, his instrumentalism and relativism, and his alleged exaggeration of certain lessons from perceptual and developmental psychology (Fodor 1984). We shall address these issues later in the paper.

A third and less visible reaction to the classical difficulties has simply rejected the sentential or propositional attitudes as the most important form of representation used by cognitive creatures, and has insisted on the necessity of empirical and theoretical research into *brain* function in order to answer the question of what *are* the most important forms of representation and computation within cognitive creatures. Early statements can be found in Churchland 1975 and Hooker 1975; extended arguments appear in Churchland 1979 and 1981; and further arguments appear in Churchland, P.S., 1980 and 1986, and in Hooker 1987.

While the antisentential diagnosis could be given some considerable support, as the opening summary of this section illustrates, neuroscience as the recommended cure was always more difficult to sell, given the functional opacity of the biological brain. Recently, however, this has changed dramatically. We now have some provisional insight into the functional significance of the brain's microstructure, and some idea of how it represents and computes. What has been discovered so far appears to vindicate the claims of philosophical relevance and the expectations of fertility in this area, and it appears to vindicate some central elements in Kuhn's perspective as well. This neurofunctional framework promises to sustain wholly new directions of cognitive research. In the sections to follow I shall try to outline the elements of this framework and its applications to some familiar problems in the philosophy of science. I begin with the physical structure and the basic activities of the brainlike systems at issue.

III. Elementary Brainlike Networks

The functional atoms of the brain are cells called neurons (figure 1). These have a natural or default level of activity that can, however, be modulated up or down by external influences. From each neuron there extends a long, thin output fiber called an *axon*, which typically branches at the far end so as to make a large number of *synaptic connections* with either the central cell body or the bushy *den-*

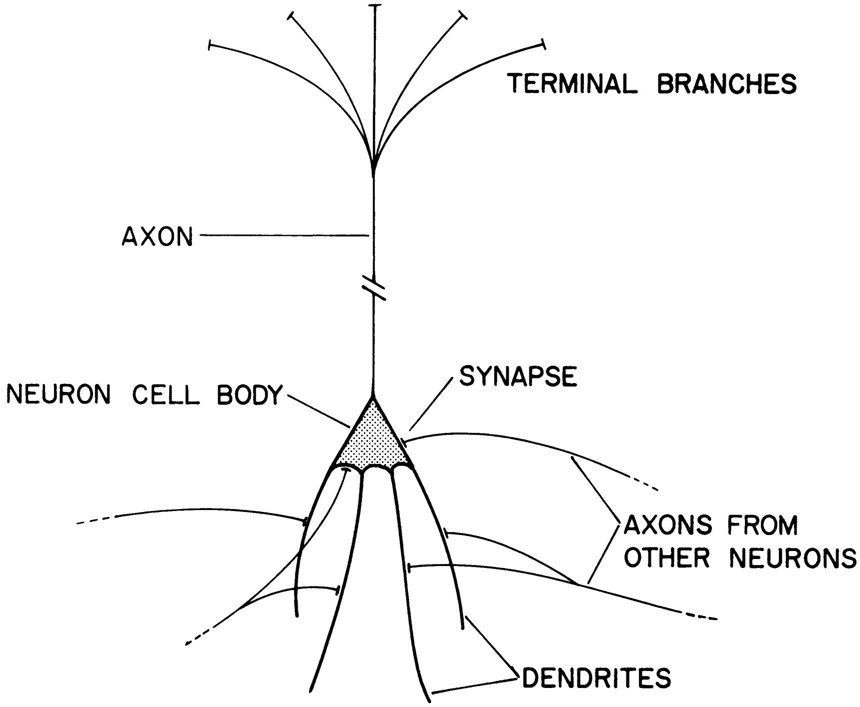


Figure 1.

drites of other neurons. Each neuron thus receives inputs from a great many other neurons, which inputs tend to excite (or to inhibit, depending on the type of synaptic connection) its normal or default level of activation. The level of activation induced is a function of the *number* of connections, of their size or *weight*, of their *polarity* (stimulatory or inhibitory), and of the *strength* of the incoming signals. Furthermore, each neuron is constantly emitting an output signal along its own axon, a signal whose strength is a direct function of the overall level of activation in the originating cell body. That signal is a train of pulses or *spikes*, as they are called, which are propagated swiftly along the axon. A typical cell can emit spikes along its axon at anything between zero and perhaps 200 Hz. Neurons, if you like, are humming to one another, in basso notes of varying frequency.

The networks to be explored attempt to simulate natural neurons with artifical units of the kind depicted in figure 2. These units admit of various levels of activation, which we shall assume to vary between 0 and 1. Each unit receives input signals from other units via 'synaptic' connections of various weights and polari-

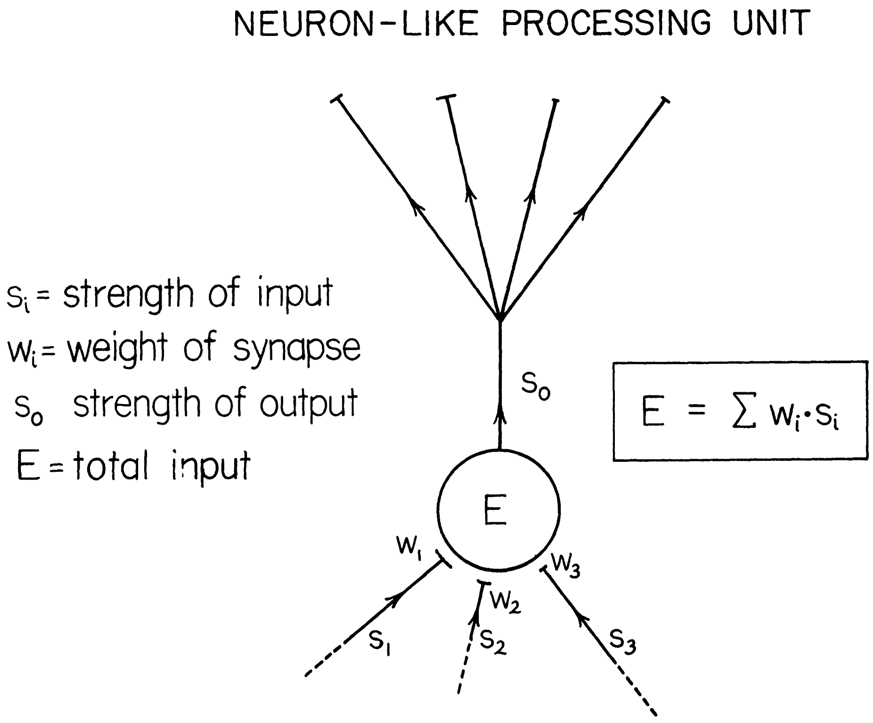
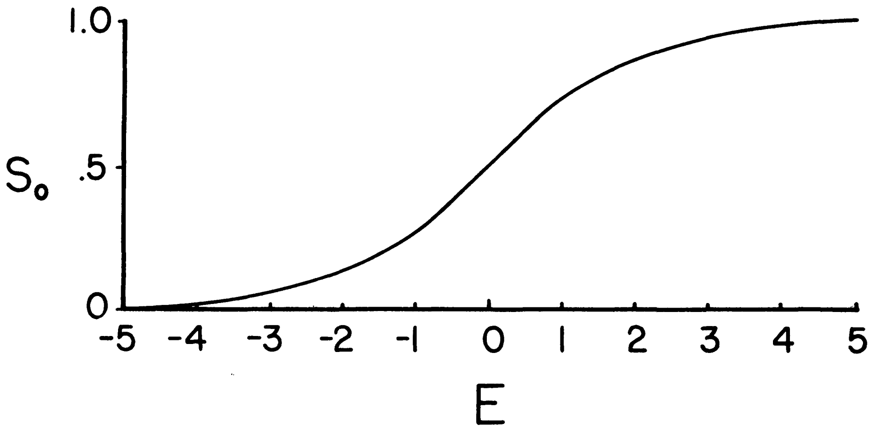


Figure 2.

ties. These are represented in the diagram as small end-plates of various sizes. For simplicity's sake, we dispense with dendritic trees: the axonal end branches from other units all make connections directly to the 'cell body' of the receiving unit. The total modulating effect E impacting on that unit is just the sum of the contributions made by each of the connections. The contribution of a single connection is just the product of its weight w_i times the strength s_i of the signal arriving at that connection. Let me emphasize that if for some reason the connection weights were to change over time, then the unit would receive a quite different level of overall excitation or inhibition in response to the very same configuration of input signals.

Turn now to the output side of things. As a function of the total input E, the unit modulates its activity level and emits an output signal of a certain strength s_o along its 'axonal' output fiber. But s_o is not a direct or *linear* function of E. Rather, it is an S-shaped function as in figure 3. The reasons for this small wrinkle will emerge later. I mention it here because its inclusion completes the



$$S_o = \frac{1}{1 + e^{-E}}$$

Figure 3.

story of the elementary units. Of their intrinsic properties, there is nothing left to tell. They are very simple indeed.

It remains to arrange them into networks. In the brain, neurons frequently consitute a population, all of which send their axons to the site of a second population of neurons, where each arriving axon divides into terminal end branches in order to make synaptic connections with many different cells within the target population. Axons from cells in this second population can then project to a third population of cells, and so on. This is the inspiration for the arrangement of figure 4.

The units in the bottom or input layer of the network may be thought of as 'sensory' units, since the level of activation in each is directly determined by aspects of the environment (or perhaps by the experimenter, in the process of simulating some environmental input). The activation level of a given input unit is designed to be a response to a specific aspect or dimension of the overall input stimulus that strikes the bottom layer. The assembled set of simultaneous activation levels in all of the input units is the network's *representation* of the input stimulus. We may refer to that configuration of stimulation levels as the *input vector*, since it is just an ordered set of numbers or magnitudes. For example, a given stimulus might produce the vector ⟨.5, .3, .9, .2⟩.

These input activation levels are then propagated upwards, via the output sig-

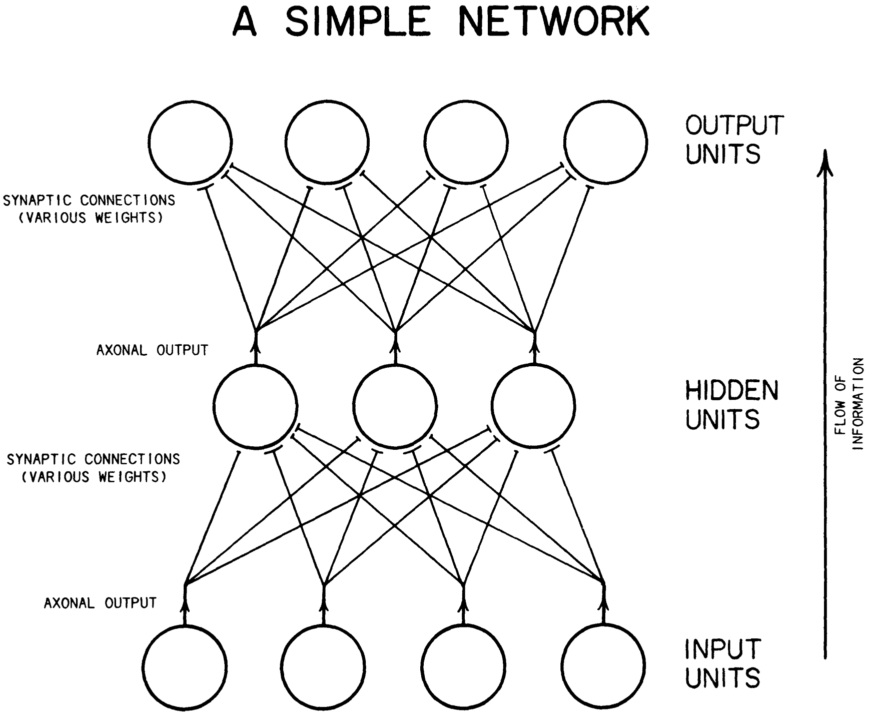


Figure 4.

nal in each unit's axon, to the middle layer of the network, to what are called the *hidden units*. As can be seen in figure 4, any unit in the input layer makes a synaptic connection of some weight or other with every unit at this intermediate layer. Each hidden unit is thus the target of several inputs, one for each cell at the input layer. The resulting activation level of a given hidden unit is essentially just the sum of all of the influences reaching it from the cells in the lower layer.

The result of this upward propagation of the input vector is a set of activation levels across the three units in the hidden layer, called the *hidden unit activation vector*. The values of that three-element vector are strictly determined by

(a) the makeup of the *input vector* at the input layer, and
(b) the various values of the *connection weights* at the ends of the terminal branches of the input units.

What this bottom half of the network does, evidently, is convert or transform one activation vector into another.

The top half of the network does exactly the same thing, in exactly the same way. The activation vector at the hidden layer is propagated upward to the output (topmost) layer of units, where an *output vector* is produced, whose character is determined by

(a) the makeup of the activation vector at the hidden layer, and
(b) the various values of the connection weights at the ends of the terminal branches of the hidden units.

Looking now at the whole network, we can see that it is just a device for transforming any given input-level activation vector into a uniquely corresponding output-level activation vector. And what determines the character of the global transformation effected is the peculiar set of values possessed by the many connection weights. This much is easy to grasp. What is not so easy to grasp, prior to exploring examples, is just how very powerful and useful those transformations can be. So let us explore some real examples.

IV. Representation and Learning in Brainlike Networks

A great many of the environmental features to which humans respond are difficult to define or characterize in terms of their purely physical properties. Even something as mundane as being the vowel sound *ā*, as in "rain," resists such characterization, for the range of acoustical variation among acceptable and recognizable *ā*s is enormous. A female child at two years and a basso male at fifty will produce quite different sorts of atmospheric excitations in pronouncing this vowel, but each sound will be easily recognized as an *ā* by other members of the same linguistic culture.

I do not mean to suggest that the matter is utterly intractable from a physical point of view, for an examination of the acoustical power spectrum of voiced vowels begins to reveal some of the similarities that unite *ā*s. And yet the analysis continues to resist a simple list of necessary and sufficient physical conditions on being an *ā*. Instead, being an *ā* seems to be a matter of being *close enough* to a *typical* *ā* sound along a *sufficient* number of distinct *dimensions of relevance*, where each notion in italics remains difficult to characterize in a nonarbitrary way. Moreover, some of those dimensions are highly contextual. A sound type that would not normally be counted or recognized as an *ā* when voiced in isolation may be unproblematically so counted if it regularly occurs, in someone's modestly accented speech, in all of the phonetic places that would normally be occupied by *ā*s. Evidently, what makes something an *ā* is in part a matter of the entire linguistic surround. In this way do we very quickly ascend to the abstract and holistic level, for even the simplest of culturally embedded properties.

What holds for phonemes holds also for a great many other important features recognizable by us—colors, faces, flowers, trees, animals, voices, smells, feelings, songs, words, meanings, and even metaphorical meanings. At the outset, the categories and resources of physics, and even neuroscience, look puny and impotent in the face of such subtlety.

And yet it is a purely physical system that recognizes such intricacies. Short of appealing to magic, or of simply refusing to confront the problem at all, we must assume that some configuration of purely physical elements is capable of grasping and manipulating these features, and by means of purely physical principles. Surprisingly, networks of the kind described in the preceding section have many of the properties needed to address precisely this problem. Let me explain.

Suppose we are submarine engineers confronted with the problem of designing a sonar system that will distinguish between the sonar echoes returned from explosive mines, such as might lie on the bottom of sensitive waterways during wartime, and the sonar echoes returned from rocks of comparable sizes that dot the same underwater landscapes. The difficulty is twofold: echoes from both objects sound indistinguishable to the casual ear, and echoes from each type show wide variation in sonic character, since both rocks and mines come in various sizes, shapes, and orientations relative to the probing sonar pulse.

Enter the network of figure 5. This one has thirteen units at the input layer, since we need to code a fairly complex stimulus. A given sonar echo is run through a frequency analyzer, and is sampled for its relative energy levels at thirteen frequencies. These thirteen values, expressed as fractions of 1, are then entered as activation levels in the respective units of the input layer, as indicated in figure 5. From here they are propagated through the network, being transformed as they go, as explained earlier. The result is a pair of activation levels in the two units at the output layer. We need only two units here, for we want the network eventually to produce an output activation vector at or near $\langle 1, 0 \rangle$ when a mine echo is entered as input, and an output activation vector at or near $\langle 0, 1 \rangle$ when a rock echo is entered as input. In a word, we want it to *distinguish* mines from rocks.

It would of course be a miracle if the network made the desired discrimination immediately, since the connection weights that determine its transformational activity are initially set at random values. At the beginning of this experiment then, the output vectors are sure to disappoint us. But we proceed to *teach* the network by means of the following procedure.

We procure a large set of recorded samples of various (genuine) mine echoes, from mines of various sizes and orientations, and a comparable set of genuine rock echoes, keeping careful track of which is which. We then feed these echoes into the network, one by one, and observe the output vector produced in each case. What interests us in each case is the amount by which the actual output vector *differs* from what would have been the 'correct' vector, given the identity of the specific echo that produced it. The details of that error, for each element of the output vector, are then fed into a special rule that computes a set of small

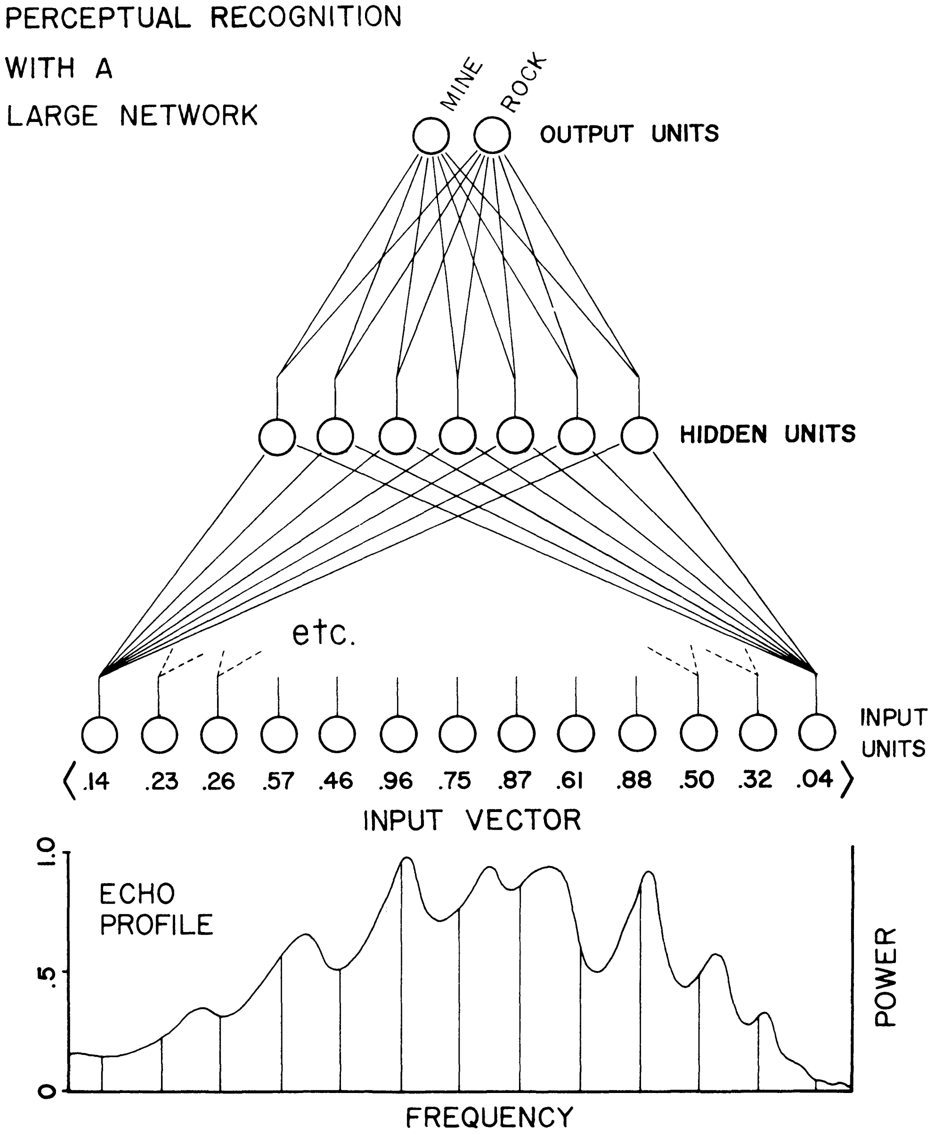

PERCEPTUAL RECOGNITION
WITH A
LARGE NETWORK
MINE
ROCK
OUTPUT UNITS
HIDDEN UNITS
etc.
INPUT
UNITS
⟨ .14 .23 .26 .57 .46 .96 .75 .87 .61 .88 .50 .32 .04 ⟩
INPUT VECTOR
ECHO
PROFILE
1.0
.5
0
POWER
FREQUENCY

Figure 5.

changes in the values of the various synaptic weights in the system. The idea is to identify those weights most responsible for the error, and then to nudge their values in a direction that would at least reduce the amount by which the output vector is in error. The slighty modified system is then fed another echo from the training set, and the entire procedure is repeated.

This provides the network with a 'teacher'. The process is called "training up the network," and it is standardly executed by an auxiliary computer programmed to feed samples from the training set into the network, monitor its responses, and adjust the weights according to the special rule after each trial. Under the pressure of such repeated corrections, the behavior of the network slowly converges on the behavior we desire. That is to say, after several thousands of presentations of recorded echoes and subsequent adjustments, the network starts to give the right answer close to 90 percent of the time. When fed a mine echo, it generally gives something close to a $\langle 1, 0\rangle$ output. And when fed a rock echo, it generally gives something close to a $\langle 0, 1\rangle$.

A useful way to think of this is captured in figure 6. Think of an abstract space of many dimensions, one for each weight in the network (105 in this case), plus one dimension for representing the overall error of the output vector on any given trial. Any point in that space represents a unique configuration of weights, plus the performance error that that configuration produces. What the learning rule does is steadily nudge that configuration away from erroneous positions and toward positions that are less erroneous. The system inches its way down an 'error gradient' toward a global error minimum. Once there, it responds reliably to the relevant kinds of echoes. It even responds well to echoes that are 'similar' to mine echoes, by giving output vectors that are closer to $\langle 1, 0\rangle$ than to $\langle 0, 1\rangle$.

There was no guarantee the network would succeed in learning to discriminate the two kinds of echoes, because there was no guarantee that rock echoes and mine echoes would differ in any systematic or detectable way. But it turns out that mine echoes do indeed have some complex of relational or structural features that distinguishes them from rock echoes, and under the pressure of repeated error corrections, the network manages to lock onto, or become 'tuned' to, that subtle but distinctive weave of features.

We can test whether it has truly succeeded in this by now feeding the network some mine and rock echoes not included in the training set, echoes it has never encountered before. In fact, the network does almost as well classifying the new echoes as it does with the samples in its training set. The 'knowledge' it has acquired generalizes quite successfully to new cases. (This example is a highly simplified account of some striking results from Gorman and Sejnowski 1988.)

All of this is modestly amazing, because the problem is quite a difficult one, at least as difficult as learning to discriminate the phoneme $\bar{a}$. Human sonar operators, during a long tour of submarine duty, eventually learn to distinguish the two kinds of echoes with some uncertain but nontrivial regularity. But they never per-

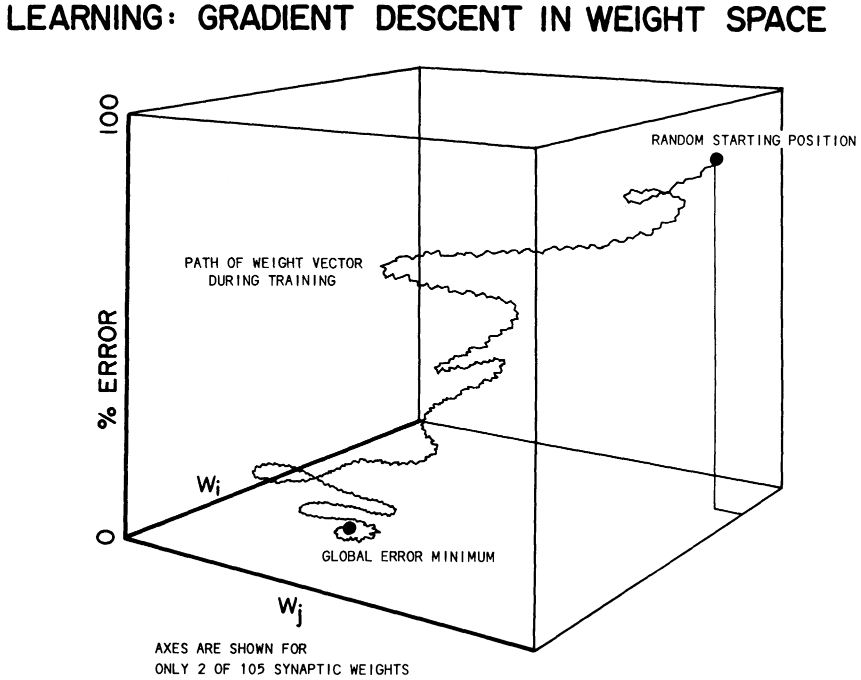

LEARNING: GRADIENT DESCENT IN WEIGHT SPACE
100
% ERROR
0
RANDOM STARTING POSITION
PATH OF WEIGHT VECTOR DURING TRAINING
W_i
GLOBAL ERROR MINIMUM
W_j
AXES ARE SHOWN FOR ONLY 2 OF 105 SYNAPTIC WEIGHTS

Figure 6.

form at the level of the artificial network. Spurred on by this success, work is currently underway to train up a network to distinguish the various phonemes characteristic of English speech (Zipser and Elman 1988). The idea is to produce a speech-recognition system that will not be troubled by the acoustic idiosyncracies of diverse speakers, as existing speech-recognition systems are.

The success of the mine/rock network is further intriguing because the 'knowledge' the network has acquired, concerning the distinctive character of mine echoes, consists of nothing more than a carefully orchestrated set of connection weights. And it is finally intriguing because there exists a learning algorithm – the rule for adjusting the weights as a function of the error displayed in the output vector – that will eventually produce the required set of weights, given sufficient examples on which to train the network (Rumelhart et al. 1986).

How can a set of connection weights possibly embody knowledge of the desired distinction? Think of it in the following way. Each of the thirteen input units represents one aspect or dimension of the incoming stimulus. Collectively, they give a simultaneous profile of the input echo along thirteen distinct dimensions. Now perhaps there is only one profile that is roughly characteristic of mine echoes; or perhaps there are many different profiles, united by a common relational feature (e.g., that the activation value of unit #6 is always three times the value of unit #12); or perhaps there is a disjunctive set of such relational features; and so forth. In each case, it is possible to rig the weights so that the system will respond in a typical fashion, at the output layer, to all and only the relevant profiles.

The units at the hidden layer are very important in this. If we consider the abstract space whose seven axes represent the possible activation levels of each of the seven hidden units, then what the system is searching for during the training period is a set of weights that *partitions* this space so that any mine input produces an activation vector across the hidden units that falls somewhere within one large subvolume of this abstract space, while any rock input produces a vector that falls somewhere into the complement of that subvolume (figure 7). The job of the top half of the network is then the relatively easy one of distinguishing these two subvolumes into which the abstract space has been divided.

Vectors near the center of (or along a certain path in) the mine-vector subvolume represent *prototypical* mine echoes, and these will produce an output vector very close to the desired $\langle 1, 0 \rangle$. Vectors nearer to the surface (strictly speaking, the *hyper*surface) that partitions the abstract space represent atypical or problematic mine echoes, and these produce more ambiguous output vectors such as $\langle .6, .4 \rangle$. The network's discriminative responses are thus graded responses: the system is sensitive to *similarities* along all of the relevant dimensions, and especially to rough conjunctions of these subordinate similarities.

So we have a system that learns to discriminate hard-to-define perceptual features, and to be sensitive to similarities of a comparably diffuse but highly relevant character. And once the network is trained up, the recognitional task takes only a split second, since the system processes the input stimulus in parallel. It finally gives us a discriminatory system that performs something like a living creature, both in its speed and in its overall character.

I have explained this system in some detail, so that the reader will have a clear idea of how things work in at least one case. But the network described is only one instance of a general technique that works well in a large variety of cases. Networks can be constructed with a larger number of units at the output layer, so as to be able to express not just two, but a large number of distinct discriminations.

One network, aptly called NETtalk by its authors (Rosenberg and Sejnowski 1987), takes vector codings for seven-letter segments of printed words as inputs, and gives vector codings for phonemes as outputs. These output vectors can be fed directly into a sound synthesizer as they occur, to produce audible sounds. What this network learns to do is to transform printed words into audible speech. Though it involves no understanding of the words that it 'reads', the network's feat

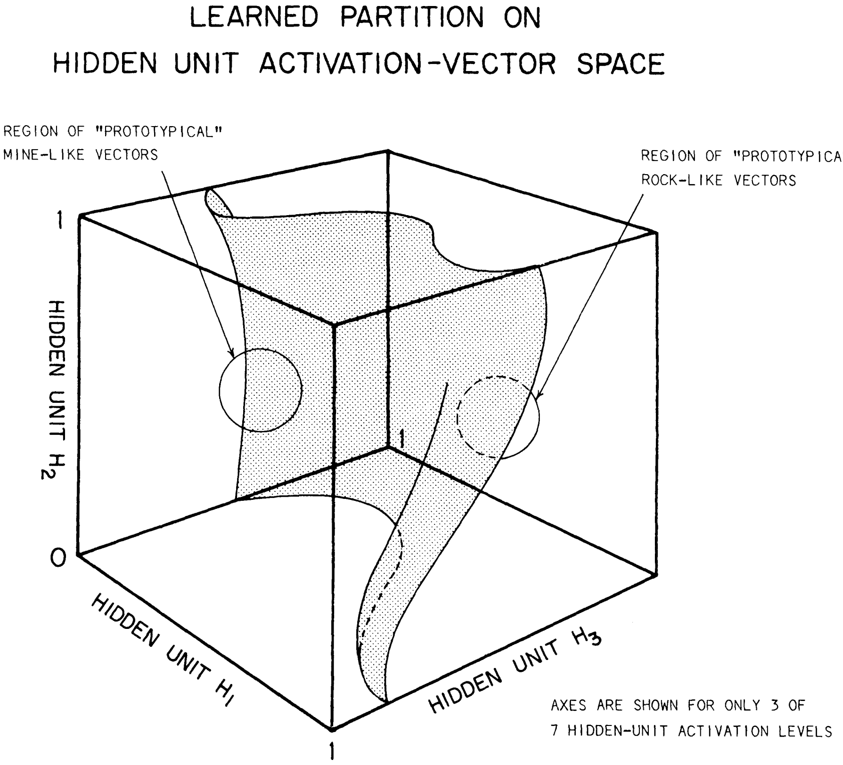

LEARNED PARTITION ON
HIDDEN UNIT ACTIVATION-VECTOR SPACE
REGION OF "PROTOTYPICAL" MINE-LIKE VECTORS
REGION OF "PROTOTYPICAL" ROCK-LIKE VECTORS
1
HIDDEN UNIT H₂
0
1
1
HIDDEN UNIT H₁
HIDDEN UNIT H₃
AXES ARE SHOWN FOR ONLY 3 OF 7 HIDDEN-UNIT ACTIVATION LEVELS

Figure 7.

is still very impressive, because it was given no rules whatever concerning the phonetic significance of standard English spelling. It began its training period by producing a stream of unintelligible babble in response to text entered as input. But in the course of many thousands of word presentations, and under the steady pressure of the weight-nudging algorithm, the set of weights slowly meanders its way to a configuration that reduces the measured error close to zero. After such training it will then produce as output, given arbitrary English text as input, perfectly intelligible speech with only rare and minor errors.

This case is significant for a number of reasons. First, the trained network makes a large number of discriminations (79, in fact), not just a binary one. Second, it contains no explicit representation of any *rules*, however much it might seem to be following a set of rules. Third, it has mastered an input/output transformation that is notoriously irregular, and it must be sensitive to lexical context in order to do so. (Specifically, the phoneme it assigns to the center or focal letter of its seven-letter input is in large part a function of the identity of the three letters on either side.) And fourth, it portrays some aspects of a 'sensori*motor*' skill, rather than a purely sensory skill: it is producing highly complex behavior.

NETtalk has some limitations, of course. Pronunciations that depend on specifically semantical or grammatical distinctions will generally elude its grasp (unless they happen to be reflected in some way in the corpus of its training words, as occasionally they are), since NETtalk knows neither meanings nor syntax. But such dependencies affect only a very small percentage of the transformations appropriate to any text, and they are in any case to be expected. To overcome them completely would require a network that actually understands the text being read. And even then mistakes would occur, for even humans occasionally misread words as a result of grammatical or semantical confusion. What is arresting about NETtalk is just how very much of the complex and irregular business of text-based pronunciation can be mastered by a simple network with only a few hundred neuronlike units.

Another rather large network (by Lehky and Sejnowski 1988a, 1988b) addresses problems in vision. It takes codings for smoothly varying gray-scale pictures as input, and after training it yields as outputs surprisingly accurate codings for the curvatures and orientations of the physical objects portrayed in the pictures. It solves a form of the 'shape from shading' problem long familiar to theorists in the field of vision. This network is of special interest because a subsequent examination of the 'receptive fields' of the trained hidden units shows them to have acquired some of the same response properties as are displayed by cells in the visual cortex of mature animals. Specifically, they show a maximum sensitivity to spots, edges, and bars in specific orientations. This finding echoes the seminal work of Hubel and Wiesel (1962), in which cells in the visual cortex were discovered to have receptive fields of this same character. Results of this kind are very important, for if we are to take these artificial networks as models for how the brain works, then they must display realistic behavior not just at the macrolevel: they must also display realistic behavior at the microlevel.

Enough examples. You have seen something of what networks of this kind can do, and of how they do it. In both respects they contrast sharply with the kinds of representational and processing strategies that philosophers of science, inductive logicians, cognitive psychologists, and AI workers have traditionally ascribed to us (namely, sentencelike representations manipulated by formal rules). You can see also why this theoretical and experimental approach has captured the interest of those who seek to understand how the microarchitecture of the biological brain produces the phenomena displayed in human and animal cognition. Let us now explore the functional properties of these networks in more detail, and see how they bear on some of the traditional issues in epistemology and the philosophy of science.

V. Some Functional Properties of Brainlike Networks

The networks described above are descended from a device called the *Perceptron* (Rosenblatt 1959), which was essentially just a two-layer as opposed to a three-layer network. Devices of this configuration could and did learn to discriminate a considerable variety of input patterns. Unfortunately, having the input layer connected directly to the output layer imposes very severe limitations on the range of possible transformations a network can perform (Minsky and Papert 1969), and interest in Perceptron-like devices was soon eclipsed by the much faster-moving developments in standard 'program-writing' AI, which exploited the high-speed general-purpose digital machines that were then starting to become widely available. Throughout the seventies, research in artificial 'neural nets' was an underground program by comparison.

It has emerged from the shadows for a number of reasons. One important factor is just the troubled doldrums into which mainstream or program-writing AI has fallen. In many respects, these doldrums parallel the infertility of the classical approach to theories and learning within the philosophy of science. This is not surprising, since mainstream AI was proceeding on many of the same basic assumptions about cognition, and many of its attempts were just machine implementations of learning algorithms proposed earlier by philosophers of science and inductive logicians (Glymour 1987). The failures of mainstream AI—unrealistic learning, poor performance in complex perceptual and motor tasks, weak handling of analogies, and snaillike cognitive performance despite the use of very large and fast machines—teach us even more dramatically than do the failures of mainstream philosophy that we need to rethink the style of representation and computation we have been ascribing to cognitive creatures.

Other reasons for the resurgence of interest in networks are more positive. The introduction of additional layers of intervening or 'hidden' units produced a dramatic increase in the range of possible transformations that the network could effect. As Sejnowski et al. (1986) describe it:

> . . . only the first-order statistics of the input pattern can be captured by direct connections between input and output units. The role of the hidden units is to capture higher-order statistical relationships and this can be accomplished if significant underlying features can be found that have strong, regular relationships with the patterns on the visible units. The hard part of learning is to find the set of weights which turn the hidden units into useful feature detectors.

Equally important is the S-shaped, nonlinear response profile (figure 3) now assigned to every unit in the network. So long as this response profile remains linear, any network will be limited to computing purely linear transformations. (A transformation $f(x)$ is *linear* just in case $f(n \bullet x) = n \bullet f(x)$, and $f(x + y) = f(x) + f(y)$.) But a nonlinear response profile for each unit brings the entire range of possible nonlinear transformations within reach of three-layer networks, a dramatic expansion of their computational potential. Now there are *no* transformations that lie beyond the computational power of a large enough and suitably weighted network.

A third factor was the articulation, by Rumelhart, Hinton, and Williams (1986a), of the *generalized delta rule* (a generalization, to three-layer networks, of Rosenblatt's original teaching rule for adjusting the weights of the Perceptron), and the empirical discovery that this new rule very rarely got permanently stuck in inefficient 'local minima' on its way toward finding the best possible configuration of connection weights for a given network and a given problem. This was a major breakthrough, not so much because "learning by the back-propagation of error," as it has come to be called, was just like human learning, but because it provided us with an efficient technology for quickly training up various networks on various problems, so that we could study their properties and explore their potential.

The way the generalized delta rule works can be made fairly intuitive given the idea of an abstract weight space as represented in figure 6. Consider any output vector produced by a network with a specific configuration of weights, a configuration represented by a specific position in weight space. Suppose that this output vector is in error by various degrees in various of its elements. Consider now a single synapse at the ouput layer, and consider the effect on the output vector that a small positive or negative change in its weight would have had. Since the output vector is a determinate function of the system's weights (assuming we hold the input vector fixed), we can calculate which of these two possible changes, if either, would have made the greater improvement in the output vector. The relevant change is made accordingly. (For more detail, see Rumelhart et al. 1986b.)

If a similar calculation is performed over every synapse in the network, and the change in its weight is then made accordingly, what the resulting shift in the position of the system's overall point in weight space amounts to is a small slide *down* the steepest face of the local 'error surface'. Note that there is no guarantee that this incremental shift moves the system directly towards the global position of zero error (that is why perfection cannot be achieved in a single jump). On the contrary, the descending path to a global error minimum may be highly circuitous. Nor is there any guarantee that the system must eventually reach such a global minimum. On the contrary, the downward path from a given starting point may well lead to a merely 'local' minimum, from which only a large change in the system's weights will afford escape, a change beyond the reach of the delta rule. But in fact this happens relatively rarely, for it turns out that the more dimensions (synapses) a system has, the smaller the probability of there being an intersecting local minimum in *every one* of the available dimensions. The global point is usually able to slide down some narrow cleft in the local topography. Empirically then, the back-propagation algorithm is surprisingly effective at driving the system to the global error minimum, at least where we can identify that global minimum effectively.

The advantage this algorithm provides is easily appreciated. The possible combinations of weights in a network increases exponentially with the size of the network. Assuming conservatively that each weight admits of only ten possible values, the number of distinct positions in 'weight space' (i.e., the number of possible weight configurations) for the simple rock/mine network of figure 5 is already 10^{105}! This space is far too large to explore efficiently without something like the generalized delta rule and the back-propagation of error to do it for us. But with the delta rule, administered by an auxiliary computer, researchers have shown that networks of the simple kind described are capable of learning some quite extraordinary skills, and of displaying some highly intriguing properties. Let me now return to an exploration of these.

An important exploratory technique in cognitive and behavioral neuroscience is to record, with an implanted microelectrode, the electrical activity of a single neuron during cognition or behavior in the intact animal. This is relatively easy to do, and it does give us tantalizing bits of information about the cognitive significance of neural activity (recall the results of Hubel and Wiesel mentioned earlier). Single-cell recordings give us only isolated bits of information, however, and what we would really like to monitor are the *patterns* of simultaneous neural activation across large numbers of cells in the same subsystem. Unfortunately, effective techniques for simultaneous recording from large numbers of adjacent cells are still in their infancy. The task is extremely difficult.

By contrast, this task is extremely easy with the artificial networks we have been describing. If the network is real hardware, its units are far more accessible than the fragile and microscopic units of a living brain. And if the network is merely being simulated within a standard computer (as is usually the case), one can write the program so that the activation levels of any unit, or set of units, can be read out on command. Accordingly, once a network has been successfully trained up on some skill or other, one can then examine the collective behavior of its units during the exercise of that skill.

We have already seen the results of one such analysis in the rock/mine network. Once the weights have reached their optimum configuration, the activation vectors (i.e., the patterns of activation) at the hidden layer fall into two disjoint classes: the vector space is partitioned in two, as depicted schematically in figure 7. But a mere binary discrimination is an atypically simple case. The reader NETtalk, for example, partitions its hidden-unit vector space into fully seventy nine subspaces. The reason is simple. For each of the twenty six letters in the alphabet, there is at least one phoneme assigned to it, and for many letters there are several phonemes that might be signified, depending on the lexical context. As it happens, there are seventy nine distinct letter-to-phoneme associations to be learned if one is to master the pronunciation of English spelling, and in the successfully trained network a distinct hidden-unit activation vector occurs when each of these seventy nine possible transformations is effected.

In the case of the rock/mine network, we noted a similarity metric within each of its two hidden-unit subspaces. In the case of NETtalk, we also find a similarity metric, this time across the seventy nine functional hidden-unit vectors (by 'functional vector', I mean a vector that corresponds to one of the seventy nine desired letter-to-phoneme transformations in the trained network). Rosenberg and Sejnowski (1987) did a 'cluster analysis' of these vectors in the trained network. Roughly, their procedure was as follows. They asked, for every functional vector in that space, what other such vector is closest to it? The answers yielded about thirty vector pairs. They then constructed a secondary vector for each such pair, by averaging the two original vectors, and asked, for every such secondary vector, what other secondary vector (or so far unpaired primary vector) is closest to it? This produced a smaller set of secondary-vector pairs, on which the averaging procedure was repeated to produce a set of tertiary vectors. These were then paired in turn, and so forth. This procedure produces a hierarchy of groupings among the original transformations, and it comes to an end with a grand division of the seventy nine original vectors into two disjoint classes.

As it happens, that deepest and most fundamental division within the hidden-unit vector space corresponds to the division between the consonants and the vowels! Looking further into this hierarchy, into the consonant branch, for example, we find that there are subdivisions into the principal consonant types, and that within these branches there are further subdivisions into the most similar consonants. All of this is depicted in the tree diagram of figure 8. What the network has managed to recover, from its training set of several thousand English words, is the highly irregular phonological significance of standard English spelling, plus the hierarchical organization of the phonetic structure of English speech.

Here we have a clear illustration of two things at once. The first lesson is the capacity of an activation-vector space to embody a rich and well-structured hierarchy of categories, complete with a similarity metric embracing everything within it. And the second lesson is the capacity of such networks to embody representations of factors and patterns that are only partially or implicitly reflected in the corpus of inputs. Though I did not mention it earlier, the rock/mine network provides another example of this, in that the final partition made on its hidden-unit vector space corresponds in fact to the objective distinction between sonar targets made of *metal* and sonar targets made of *nonmetal*. That is the true uniformity that lies behind the apparently chaotic variety displayed in the inputs.

It is briefly tempting to suggest that NETtalk has the concept of a 'hard *c*', for example, and that the rock/mine network has the concept of 'metal'. But this won't really do, since the vector-space representations at issue do not play a conceptual

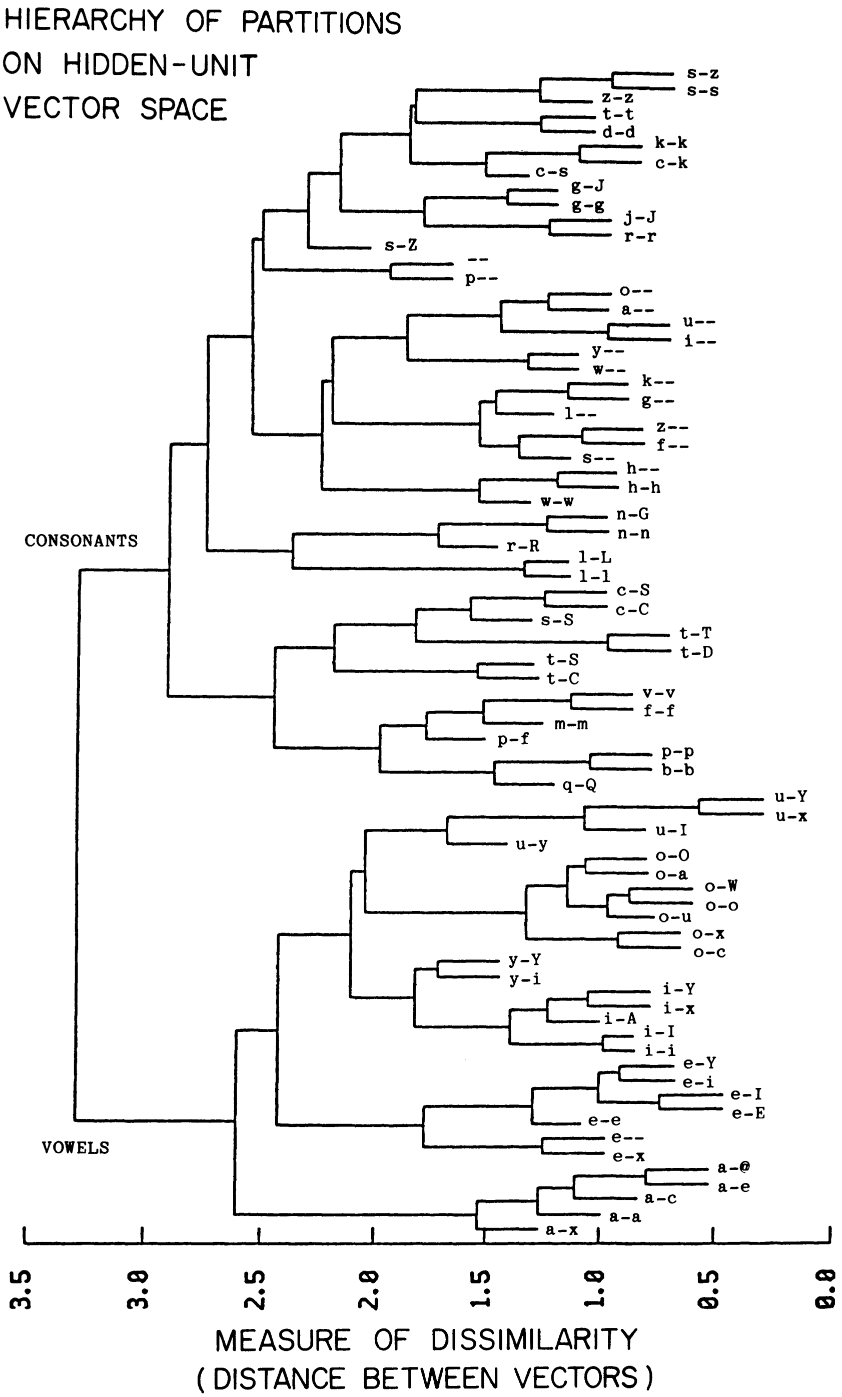

HIERARCHY OF PARTITIONS
ON HIDDEN-UNIT
VECTOR SPACE
CONSONANTS
VOWELS
s-z
s-s
z-z
t-t
d-d
k-k
c-k
c-s
g-J
g-g
j-J
r-r
s-Z
--
p--
o--
a--
u--
i--
y--
w--
k--
g--
l--
z--
f--
s--
h--
h-h
w-w
n-G
n-n
r-R
l-L
l-l
c-S
c-C
s-S
t-T
t-D
t-S
t-C
v-v
f-f
m-m
p-f
p-p
b-b
q-Q
u-Y
u-x
u-I
u-y
o-O
o-a
o-W
o-o
o-u
o-x
o-c
y-Y
y-i
i-Y
i-x
i-A
i-I
i-i
e-Y
e-i
e-I
e-E
e-e
e--
e-x
a-@
a-e
a-c
a-a
a-x
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
MEASURE OF DISSIMILARITY
(DISTANCE BETWEEN VECTORS)

Figure 8.

or computational role remotely rich enough to merit their assimilation to specifically human concepts. Nevertheless, it is plain that both networks have contrived a system of internal representations that truly corresponds to important distinctions and structures in the outside world, structures that are not explicitly represented in the corpus of their sensory inputs. The value of those representations is that they and only they allow the networks to 'make sense' of their variegated and often noisy input corpus, in the sense that they and only they allow the network to respond to those inputs in a fashion that systematically reduces the error messages to a trickle. These, I need hardly remind, are the functions typically ascribed to *theories*.

What we are confronting here is a possible conception of 'knowledge' or 'understanding' that owes nothing to the sentential categories of current common sense. An individual's overall theory-of-the-world, we might venture, is not a large collection or a long list of stored symbolic items. Rather, it is a specific point in that individual's synaptic weight space. It is a configuration of connection weights, a configuration that partitions the system's activation-vector space(s) into useful divisions and subdivisions relative to the inputs typically fed the system. 'Useful' here means 'tends to minimize the error messages'.

A possible objection here points to the fact that differently weighted systems can produce the same, or at least roughly the same, partitions on their activation-vector spaces. Accordingly, we might try to abstract from the idiosyncratic details of a system's connection weights, and identify its global theory directly with the set of partitions they produce within its activation-vector space. This would allow for differently weighted systems to have the same theory.

There is some virtue in this suggestion, but also some vice. While differently weighted systems can embody the same partitions and thus display the same output performance on any given input, they will still *learn* quite differently in the face of a protracted sequence of new and problematic inputs. This is because the learning algorithm that drives the system to new points in weight space does not care about the relatively global partitions that have been made in activation space. All it cares about are the individual *weights* and how they relate to apprehended error. The laws of cognitive evolution, therefore, do not operate primarily at the level of the partitions, at least on the view of things here being explored. Rather, they operate at the level of the weights. Accordingly, if we want our 'unit of cognition' to figure in the *laws* of cognitive development, the point in weight space seems the wiser choice of unit. We need only concede that different global theories can occasionally produce identical short-term behavior.

The level of the partitions certainly corresponds more closely to the 'conceptual' level, as understood in common sense and traditional theory, but the point is that this seems not to be the most important dynamical level, even when explicated in neurocomputational terms. Knowing a creature's vector-space partitions may suffice for the accurate short-term prediction of its behavior, but that knowledge is inadequate to predict or explain the evolution of those partitions over the course of time and cruel experience. Knowledge of the weights, by contrast, *is* sufficient for this task. This gives substance to the conviction, voiced back in section II, that to explain the phenomenon of *conceptual change*, we need to unearth a level of subconceptual combinatorial elements within which different concepts can be articulated, evaluated, and then modified according to their performance. The connection weights provide a level that meets all of these conditions.

This general view of how knowledge is embodied and accessed in the brain has some further appealing features. If we assume that the brains of the higher animals work in something like the fashion outlined, then we can explain a number of puzzling features of human and animal cognition. For one thing, the speed-of-relevant-access problem simply disappears. A network the size of a human brain—with 10^{11} neurons, 10^3 connections on each, 10^{14} total connections, and at least 10 distinct layers of 'hidden' units—can be expected, in the course of growing up, to partition its internal vector spaces into many billions of functionally relevant subdivisions, each responsive to a broad but proprietary range of highly complex stimuli. When the network receives a stimulus that falls into one of these classes, the network produces the appropriate activation vector in a matter of only tens or hundreds of milliseconds, because that is all the time it takes for the parallel-coded stimulus to make its way through only two or three or ten layers of the massively parallel network to the functionally relevant layer that drives the appropriate behavioral response. Since information is not stored in a long list that must somehow be searched, but rather in the myriad connection weights that configure the network, relevant aspects of the creature's total information are automatically accessed by the coded stimuli themselves.

A third advantage of this model is its explanation of the functional persistence of brains in the face of minor damage, disease, and the normal but steady loss of its cells with age. Human cognition degrades fairly gracefully as the physical plant deteriorates, in sharp contrast to the behavior of typical computers, which have a very low fault tolerance. The explanation of this persistence lies in the massively parallel character of the computations the brain performs, and in the very tiny contribution that each synapse or each cell makes to the overall computation. In a large network of 100,000 units, the loss or misbehavior of a single cell will not even be detectable. And in the more dramatic case of widespread cell loss, so long as the losses are more or less randomly distributed throughout the network, the gross character of the network's activity will remain unchanged: what happens is that the *quality* of its computations will be progressively degraded.

Turning now toward more specifically philosophical concerns, we may note an unexpected virtue of this approach concerning the matter of *simplicity*. This important notion has two problems. It is robustly resistant to attempts to define or measure it, and it is not clear why it should be counted an epistemic virtue in any case. There seems no obvious reason, either a priori or a posteriori, why the world should be simple rather than complex, and epistemic decisions based on the contrary assumption thus appear arbitrary and unjustified. Simplicity, conclude some (Van Fraassen 1980), is a merely pragmatic or aesthetic virtue, as opposed to a genuinely epistemic virtue. But consider the following story.

The rock/mine network of figure 5 displays a strong capacity for generalizing beyond the sample echoes in its training set: it can accurately discriminate entirely new samples of both kinds. But trained networks do not always generalize so well, and it is interesting what determines their success in this regard. How well the training generalizes is in part a function of *how many* hidden units the system possesses, or uses to solve the problem. There is, it turns out, an optimal number of units for any given problem. If the network to be trained is given more than the optimal number of hidden units, it will learn to respond appropriately to all of the various samples in its training set, but it will generalize to new samples only very poorly. On the other hand, with less than the optimal number, it never really learns to respond appropriately to all of the samples in its training set.

The reason is as follows. During the training period, the network gradually generates a set of internal representations at the level of the hidden units. One class of hidden-unit activation vectors is characteristic of rocklike input vectors; another class is characteristic of minelike input vectors. During this period, the system is *theorizing* at the level of the hidden units, exploring the space of possible activation vectors, in hopes of finding some partition or set of partitions on it that the output layer can then exploit in turn, so as to draw the needed distinctions and thus bring the process of error-induced synaptic adjustments to an end.

If there are far too many hidden units, then the learning process can be partially subverted in the following way. The lazy system cheats: it learns a set of *unrelated* representations at the level of the hidden units. It learns a distinct representation for each sample input (or for a small group of such inputs) drawn from the very finite training set, a representation that does indeed prompt the correct response at the output level. But since there is nothing common to all of the hidden-unit rock representations, or to all of the hidden-unit mine representations, an input vector from outside the training set produces a hidden-unit representation that bears no relation to the representations already formed. The system has not learned to see *what is common* within each of the two stimulus classes, which would allow it to generalize effortlessly to new cases that shared that common feature. It has just knocked together an *ad hoc* 'look-up table' that allows it to deal successfully with the limited samples in the training set, at which point the error messages cease, the weights stop evolving, and the system stops learning. (I am grateful to Terry Sejnowski for mentioning to me this wrinkle in the learning behavior of typical networks.)

There are two ways to avoid this *ad hoc*, unprojectible learning. One is to enlarge dramatically the size of the training set. This will overload the system's ability to just 'memorize' an adequate response for each of the training samples. But a more effective way is just to reduce the number of hidden units in the network, so that it lacks the resources to cobble together such wasteful and ungeneralizable internal representations. We must reduce them to the point where it has to find a *single* partition on the hidden-unit vector space, a partition that puts all of the sample rock representations on one side, and all of the sample mine representations on the other. A system constrained in this way will generalize far better, for the global partition it has been forced to find corresponds to something *common* to each member of the relevant stimulus class, even if it is only a unifying dimension of variation (or set of such dimensions) that unites them all by a similarity relation. It is the generation of that similarity relation that allows the system to respond appropriately to novel examples. They may be new to the system, but they fall on a spectrum for which the system now has an adequate representation.

Networks with only a few hidden units in excess of the optimal number will sometimes spontaneously achieve the maximally simple 'hypothesis' despite the excess units. The few unneeded units are slowly shut down by the learning algorithm during the course of training. They become zero-valued elements in all of the successful vectors. Networks will not always do this, however. The needed simplicity must generally be forced from the outside, by a progressive reduction in the available hidden units.

On the other hand, if the network has too few hidden units, then it lacks the resources even to express an activation vector that is adequate to characterize the underlying uniformity, and it will never master completely even the smallish corpus of samples in the training set. In other words, simplicity may be a virtue, but the system must command sufficient complexity at least to meet the task at hand.

We have just seen how forcing a neural network to generate a smaller number of distinct partitions on a hidden-unit vector space of fewer dimensions can produce a system whose learning achievements generalize more effectively to novel cases. *Ceteris paribus*, the simpler hypotheses generalize better. Getting by with fewer resources is of course a virtue in itself, though a pragmatic one, to be sure. But this is not the principal virtue here displayed. Superior generalization is a genuinely epistemic virtue, and it is regularly displayed by networks constrained, in the fashion described, to find the simplest hypothesis concerning whatever structures might be hidden in or behind their input vectors.

Of course, nothing guarantees successful generalization: a network is always hostage to the quality of its training set relative to the total population. And there may be equally simple alternative hypotheses that generalize differentially well. But from the perspective of the relevant microdynamics, we can see at least one clear reason why simplicity is more than a merely pragmatic virtue. It is an epistemic virtue, not principally because simple hypotheses avoid the vice of being complex, but because they avoid the vice of being *ad hoc*.

VI. How Faithfully Do These Networks Depict the Brain?

The functional properties so far observed in these model networks are an encouraging reward for the structural assumptions that went into them. But just how accurate are these models, as depictions of the brain's microstructure? A wholly appropriate answer here is uncertain, for we continue to be uncertain about what features of the brain's microstructure are and are not functionally relevant, and we are therefore uncertain about what is and is not a 'legitimate' simplifying assumption in the models we make. Even so, it is plain that the models are *in*accurate in a variety of respects, and it is the point of the present section to summarize and evaluate these failings. Let me begin by underscoring the basic respects in which the models appear to be correct.

It is true that real nervous systems display, as their principal organizing feature, layers or populations of neurons that project their axons *en masse* to some distinct layer or population of neurons, where each arriving axon divides into multiple branches whose end bulbs make synaptic connections of various weights onto many cells at the target location. This description captures all of the sensory modalities and their primary relations to the brain; it captures the character of the various areas of the central brain stem; and it captures the structure of the cerebral cortex, which in humans contains at least six distinct layers of neurons, where each layer is the source and/or the target of an orderly projection of axons to and/or from elsewhere.

It captures the character of the cerebellum as well (figure 9a), a structure discussed in an earlier paper (Churchland 1986) in connection with the problem of motor control. I there described the cerebellum as having the structure of a very large 'matrix multiplier', as schematized in figure 9b. Following Pellionisz and Llinas (1982), I ascribed to this neural matrix the function of performing sophisticated transformations on incoming activation vectors. This is in fact the same function performed between any two layers of the three-layered networks described earlier, and the two cases are distinct only in the superficial details of their wiring diagrams. A three-layered network of the kind discussed earlier is equivalent to a pair of neural matrices connected in series, as is illustrated in figures 10a and 10b. The only substantive difference is that in figure 10a the end branches synapse directly onto the receiving cell body itself, while in 10b they synapse onto some dendritic filaments extending out from the receiving cell body. The actual connectivity within the two networks is identical. The cerebellum and the motor end of natural systems, accordingly, seem further instances of the gross pattern at issue.

But the details present all manner of difficulties. To begin with small ones, note

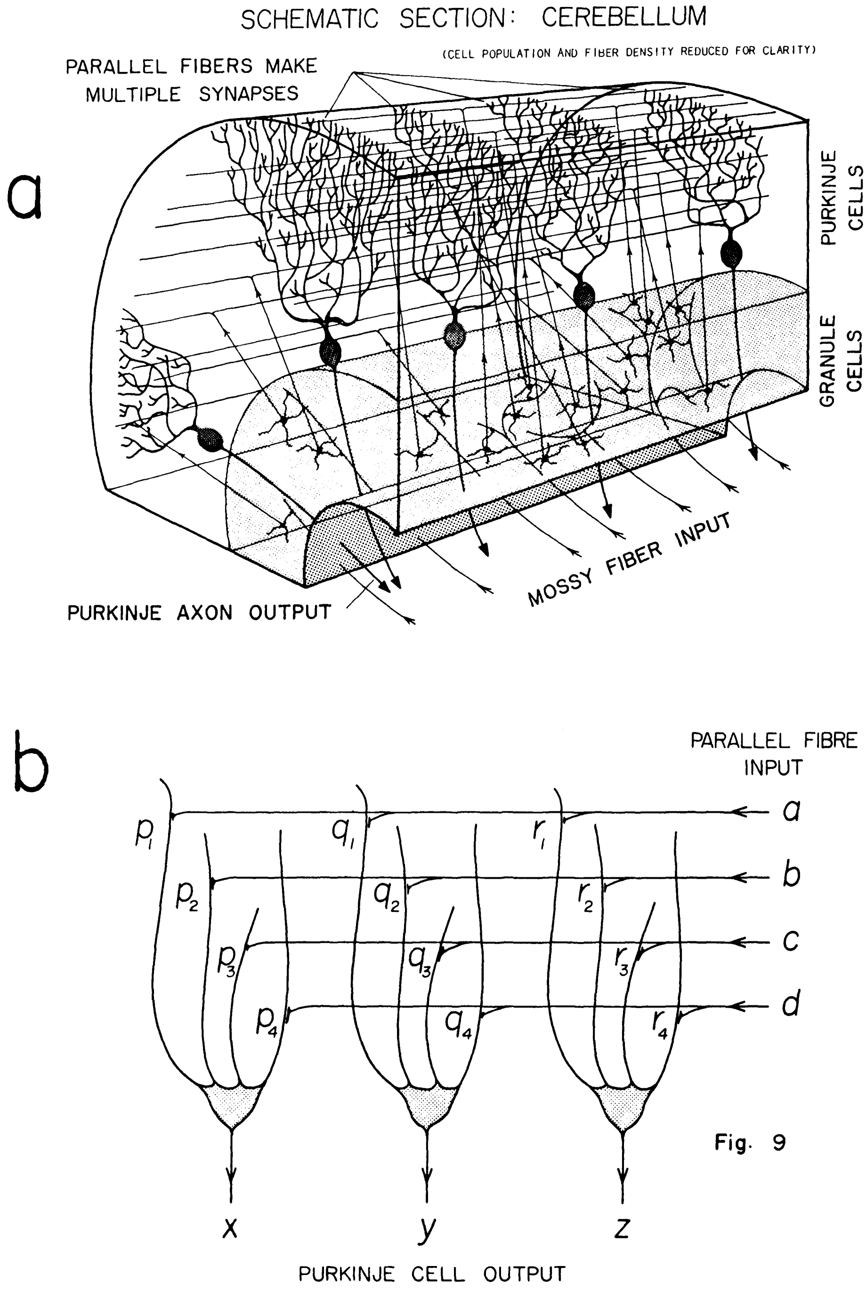

SCHEMATIC SECTION: CEREBELLUM
(CELL POPULATION AND FIBER DENSITY REDUCED FOR CLARITY)
PARALLEL FIBERS MAKE MULTIPLE SYNAPSES
a
PURKINJE CELLS
GRANULE CELLS
MOSSY FIBER INPUT
PURKINJE AXON OUTPUT
b
PARALLEL FIBRE INPUT
a
b
c
d
p_1
p_2
p_3
p_4
q_1
q_2
q_3
q_4
r_1
r_2
r_3
r_4
Fig. 9
x
y
z
PURKINJE CELL OUTPUT

Figure 9.

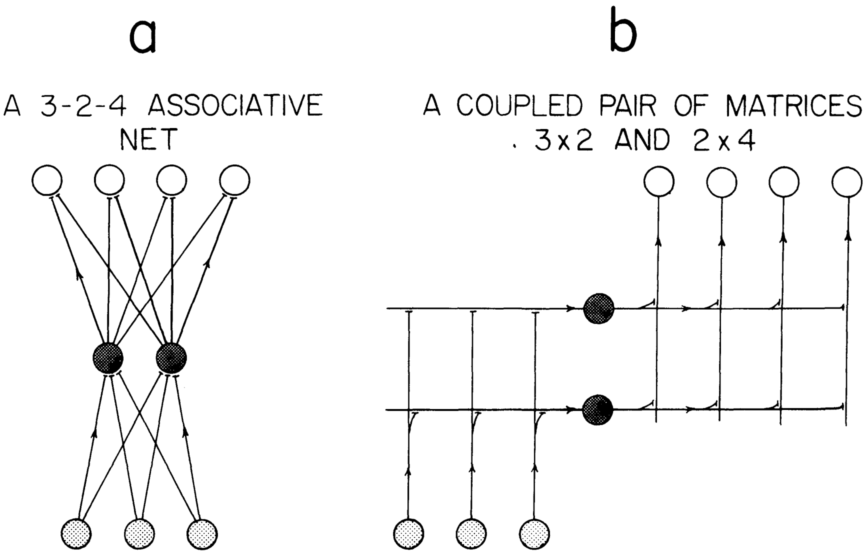


Figure 10.

that in real brains an arriving axon makes synaptic contact with only a relatively small percentage of the thousands or millions of cells in its target population, not with every last one of them as in the models. This is not a serious difficulty, since model networks with comparably pared connections still manage to learn the required transformations quite well, though perhaps not so well as a fully connected network.

More seriously, real axons, so far as is known, have terminal end bulbs that are uniformly inhibitory, or uniformly excitatory, depending on the type of neuron. We seem not to find a mixture of both kinds of connections radiating from the same neuron, nor do we find connections changing their sign during learning, as is the case in the models. Moreover, that mixture of positive and negative influences is essential to successful function in the models: the same input cell must be capable of inhibiting some cells down the line at the same time that it is busy exciting others. Further, cell populations in the brain typically show extensive 'horizontal' cell-to-cell connections *within* a given layer. In the models there are none at all (see, e.g., figure 4). Their connections join cells only to cells in distinct layers.

These last two difficulties might conceivably serve to cancel each other. One way in which an excitatory end bulb might serve to *inhibit* a cell in its target population is first to make an excitatory connection onto one of the many small *inter-*

neurons typically scattered throughout the target population of main neurons, which interneuron has made an inhibitory synaptic connection onto the target main neuron. Exciting the inhibitory interneuron would then have the effect of inhibiting the main neuron, as desired. And such a system would display a large number of short 'horizontal' intralayer connections, as is observed. This is just a suggestion, however, since it is far from clear that the elements mentioned are predominantly connected in the manner required.

More seriously still, there are several major problems with the idea that networks in the brain learn by means of the learning algorithm so effective in the models : the procedure of back-propagating apprehended errors according to the generalized delta rule. That procedure requires two things: 1) a computation of the partial correction needed for each unit in the output layer, and via these a computation of a partial correction for each unit in the earlier layers, and 2) a method of causally conveying these correction messages back through the network to the sites of the relevant synaptic connections in such a fashion that each weight gets nudged up or down accordingly. In a computer simulation of the networks at issue (which is currently the standard technique for exploring their properties), both the computation and the subsequent weight adjustments are easily done: the computation is done *outside* the network by the host computer, which has direct access to and control over every element of the network being simulated. But in the self-contained biological brain, we have to find some real source of adjustment signals, and some real pathways to convey them back to the relevant units. Unfortunately, the empirical brain displays little that answers to exactly these requirements.

Not that it contains nothing along these lines: the primary ascending pathways already described are typically matched by reciprocal or 'descending' pathways of comparable density. These allow higher layers to have an influence on affairs at lower layers. Yet the influence appears to be on the activity levels of the lower cells themselves, rather than on the myriad synaptic connections whose weights need adjusting during learning. There may be indirect effects on the synapses, of course, but it is far from clear that the brain's wiring diagram answers to the demands of the back-propagation algorithm.

The case is a little more promising in the cerebellum (figure 9a), which contains a second major input system in the aptly-named *climbing fibers* (not shown in the diagram for reasons of clarity). These fibers envelop each of the large Purkinje cells from below in the same fashion that a climbing ivy envelops a giant oak, with its filamentary tendrils reaching well up into the bushy dendritic tree of the Purkinje cell, which tree is the locus of all of the synaptic connections made by the incoming parallel fibers. The climbing fibers are thus at least roughly positioned to do the job that the back-propagation algorithm requires of them, and they are distributed one to each Purkinje cell, as consistent delivery of the error message requires. Equally, they might serve some other quite different learning algorithm, as advocated by Pellionisz and Llinas (1985). Unfortunately, there is as yet no compelling reason to believe that the modification of the weights of the parallel-fiber-to-Purkinje-dendrite synapses is even within the causal power of the climbing fibers. Nor is there any clear reason to see either the climbing fibers in the cerebellum, or the descending pathways elsewhere in the brain, as the bearers of any appropriately computed error-correction messages appropriate to needed synaptic change.

On the hardware side, therefore, the situation does not support the idea that the specific back-propagation procedure of Rumelhart et al. is the brain's central mechanism for learning. (Neither, it should be mentioned, did they claim that it is.) And it is implausible on some functional grounds as well. First, in the process of learning a recognition task, living brains typically show a progressive reduction in the reaction time required for the recognitional output response. With the delta rule, however, learning involves a progressive reduction in error, but reaction times are constant throughout. A second difficulty with the delta rule is as follows. A necessary element in its calculated apportionment of error is a representation of what would have been the *correct* vector in the output layer. That is why back-propagation is said to involve a global *teacher*, an information source that always knows the correct answers and can therefore provide a perfect measure of output error. Real creatures generally lack any such perfect information. They must struggle along in the absence of any sure compass toward the truth, and their synaptic adjustments must be based on much poorer information.

And yet their brains learn. Which means that somehow the configuration of their synaptic weights must undergo change, change steered in some way by error or related dissatisfaction, change that carves a path toward a regime of decreased error. Knowing this much, and knowing something about the microstructure and microdynamics of the brain, we can explore the space of possible learning procedures with some idea of what features to look for. If the generalized delta rule is not the brain's procedure, as it seems not to be, there remain other possible strategies for back-propagating sundry error measures, strategies that may find more detailed reflection in the brain. If these prove unrealizable, there are other procedures that do not require the organized distribution of any global error measures at all; they depend primarily on local constraints (Hinton and Sejnowski 1986; Hopfield and Tank 1985; Barto 1985; Bear et al. 1987).

One of these is worthy of mention, since something along these lines does appear to be displayed in biological brains. *Hebbian* learning (so-called after D. O. Hebb, who first proposed the mechanism) is a process of weight adjustment that exploits the temporal coincidence, on either side of a given synaptic junction, of a strong signal in the incoming axon and a high level of excitation in the receiving cell. When such conjunctions occur, Hebb proposed, some physical or chemical change is induced in the synapse, a change that increases its 'weight'. Of course, high activation in the receiving cell is typically caused by excitatory stimulation from many other incoming axons, and so the important temporal coincidence here is really between high activation among certain of the incoming axons. Those whose high activation coincides with the activation of many others have their subsequent influence on the cell increased. Crudely, those who vote with winners become winners.

A Hebbian weight-adjusting procedure can indeed produce learning in artificial networks (Linsker, 1986), although it does not seem to be as general in its effectiveness as is back-propagation. On the other hand, it has a major functional advantage over back-propagation. The latter has scaling problems, in that the process of calculating and distributing the relevant adjustments expands geometrically with the number of units in the network. But Hebbian adjustments are locally driven; they are independent of one another and of the overall size of the network. A large network will thus learn just as quickly as a small one. Indeed, a large network may even show a slight advantage over a smaller, since the temporal coincidence of incoming stimulations at a given cell will be better and better defined with increasing numbers of incoming axons.

We may also postulate 'anti-Hebbian' processes, as a means of reducing synaptic weights instead of increasing them. And we need to explore various possible flavors of each. We still have very little understanding of the functional properties of these alternative learning strategies. Nor are we at all sure that Hebbian learning, as described above, is really how the brain typically adjusts its weights. There does seem to be a good deal of activity-sensitive synaptic modification occurring in the brain, but whether its profile is specifically Hebbian is not yet established. Nor should we expect the brain to confine itself to only one learning strategy, for even at the behavioral level we can discern distinct types of learning. In sum, the problem of what mechanisms actually produce synaptic change during learning is an unsolved problem. But the functional success of the generalized delta rule assures us that the problem is solvable in principle, and other more plausible procedures are currently under active exploration.

While the matter of how real neural networks generate the right configuration of weights remains obscure, the matter of how they perform their various cognitive tasks once configured is a good deal clearer. If even small artifical networks can perform the sophisticated cognitive tasks illustrated earlier in this paper, there is no mystery that real networks should do the same or better. What the brain displays in the way of hardware is not radically different from what the models contain, and the differences invite exploration rather than disappointment. The brain is of course very much larger and denser than the models so far constructed. It has many layers rather than just two or three. It boasts perhaps a hundred distinct and highly specialized cell types, rather than just one. It is not a single *n*-layer network, but rather a large committee of distinct but parallel networks, interacting in sundry ways. It plainly commands many spaces of stunning complexity, and many skills in consequence. It stands as a glowing invitation to make our humble models yet more and more realistic, in hopes of unlocking the many secrets remaining.

VII. Computational Neuroscience: The Naturalization of Epistemology

One test of a new framework is its ability to throw a new and unifying light on a variety of old phenomena. I will close this essay with an exploration of several classic issues in the philosophy of science. The aim is to reconstruct them within the framework of the computational neuroscience outlined above. In section 5 we saw how this could be done for the case of theoretical simplicity. We there saw a new way of conceiving of this feature, and found a new perspective on why it is a genuine epistemic virtue. The hope in what follows is that we may do the same for other problematic notions and issues.

A good place to begin is with the issue of foundationalism. Here the central bone of contention is whether our observation judgments must always be theory laden. The traditional discussion endures largely for the good reason that a great deal hangs on the outcome, but also for the less momentous reason that there is ambiguity in what one might wish to count as an 'observation judgment' (an explicitly uttered sentence? a covert assertion? a propositional attitude? a conscious experience? a sensation?), and a slightly different issue emerges depending on where the debate is located.

But from the perspective of this essay, it makes no difference at what level the issue might be located. If our cognitive activities arise from a weave of networks of the kind discussed above, and if we construe a global theory as a global configuration of synaptic weights, as outlined in section 5, then it is clear that no cognitive activity whatever takes place in the absence of vectors being processed by some specific configuration of weights. That is, no cognitive activity whatever takes place in the absence of some theory or other.

This perspective bids us see even the simplest of animals and the youngest of infants as possessing theories, since they too process their activation vectors with some configuration of weights or other. The difference between us and them is not that they lack theories. Rather, their theories are just a good deal simpler than ours, in the case of animals. And their theories are much less coherent and organized and informed than ours, in the case of human infants. Which is to say, they have yet to achieve points in overall weight space that partition their activation-vector spaces into useful and well-structured subdivisions. But insofar as there is cognitive activity at all, it exploits whatever theory the creature embodies, however useless or incoherent it might be.

The only place in the network where the weights need play no role is at the absolute sensory periphery of the system, where the external stimulus is transduced into a coded input vector, for subsequent delivery to the transforming layers of weights. However, at the first occasion on which these preconceptual states have any effect at all on the downstream cognitive system, it is through a changeable configuration of synaptic weights, a configuration that produces one set of partitions on the activation-vector space of the relevant layer of neurons, one set out of millions of alternative possible sets. In other words, the very first thing that happens to the input signal is that it gets conceptualized in one of many different possible ways. At subsequent layers of processing, the same process is repeated, and the message that finally arrives at the linguistic centers, for example, has been shaped at least as much by the partitional constraints of the embedded conceptual system(s) through which it has passed as by the distant sensory input that started things off.

From the perspective of computational neuroscience, therefore, cognition is constitutionally theory laden. Presumptive processing is not a blight on what would otherwise be an unblemished activity; it is just the natural signature of a cognitive system doing what it is supposed to be doing. It is just possible that some theories are endogenously specified, of course, but this will change the present issue not at all. Innateness promises no escape from theory ladenness, for an endogenous theory is still a *theory*.

In any case, the idea is not in general a plausible one. The visual system, for example, consists of something in the neighborhood of 10^{10} neurons, each of which enjoys better than 10^3 synaptic connections, for a total of at least 10^{13} weights, each wanting specific genetic determination. That is an implausibly heavy load to place on the coding capacity of our DNA molecules. (The entire human genome contains only about 10^9 nucleotides.) It would be much more efficient to specify endogenously only the general structural principles of a type of learning network that is then likely to learn in certain standard directions, given the standard sorts of inputs and error messages that a typical human upbringing provides. This places the burden of steering our conceptual development where it belongs—on the external world, an information source far larger and more reliable than the genes.

It is a commonplace that we can construct endlessly different theories with which to explain the familiar facts of the observable world. But it is an immediate consequence of the perspective here adopted that that we can also apprehend the 'observable world' itself in a similarly endless variety of ways. For there is no 'preferred' set of partitions into which our sensory spaces must inevitably fall. It all depends on how the relevant networks are *taught*. If we systematically change the pattern of the error messages delivered to the developing network, then even the very same history of sensory stimulations will produce a quite differently weighted network, one that partitions the world into classes that cross-classify those of current 'common sense', one that finds perceptual similarities along dimensions quite alien to the ones we currently recognize, one that feeds its outputs into a very differently configured network at the higher cognitive levels as well.

In relatively small ways, this phenomenon is already familiar to us. Specialists in various fields, people required to spend years mastering the intricacies of some domain of perception and manipulation, regularly end up being able to perceive facts and to anticipate behaviors that are wholly opaque to the rest of us. But there is no reason why such variation should be confined to isolated skills and specialized understanding. In principle, the human cognitive system should be capable of sustaining any one of an enormous variety of decidedly global theories concerning the character of its commonsense *Lebenswelt* as a whole. (This possibility, defended in Feyerabend 1965, is explored at some length via examples in Churchland 1979. For extended criticism of this general suggestion see Fodor 1984. For a rebuttal and counterrebuttal see Churchland 1988 and Fodor 1988.)

To appreciate just how great is the conceptual variety that awaits us, consider the following numbers. With a total of perhaps 10^{11} neurons with an average of at least 10^3 connections each, the human brain has something like 10^{14} weights to play with. Supposing, conservatively, that each weight admits of only ten possible values, the total number of distinct possible configurations of synaptic weights (= distinct possible positions in weight space) is 10 for the first weight, times 10 for the second weight, times 10 for the third weight, etc., for a total of $10^{10^{14}}$, or $10^{100,000,000,000,000}$!! This is the total number of (just barely) distinguishable theories embraceable by humans, given the cognitive resources we currently command. To put this number into perspective, recall that the total number of elementary particles in the entire universe is only about 10^{87}.

In this way does a neurocomputational approach to perception allow us to reconstruct an old issue, and to provide novel reasons for the view that our perceptual knowledge is both theory laden and highly plastic. And it will do more. Notice that the activation-vector spaces that a matured brain has generated, and the prototypes they embody, can encompass far more than the simple sensory types such as phonemes, colors, smells, tastes, faces, and so forth. Given high-dimensional spaces, which the brain has in abundance, those spaces and the prototypes they embody can encompass categories of great complexity, generality, and abstraction, including those with a temporal dimension, such as harmonic oscillator, projectile, traveling wave, Samba, twelve-bar blues, democratic election, six-course dinner, courtship, elephant hunt, civil disobedience, and stellar collapse. It may be that the input dimensions that feed into such abstract spaces will themselves often have to be the expression of some earlier level of processing, but that is no problem. The networks under discussion are hierarchically arranged to do precisely this as a matter of course. In principle then, it is no harder for such a system to represent types of *processes*, *procedures*, and *techniques* than to represent the 'simple' sensory qualities. From the point of view of the brain, these are just more high-dimensional vectors.

This offers us a possible means for explicating the notion of a *paradigm*, as used by T. S. Kuhn in his arresting characterization of the nature of scientific understanding and development (Kuhn 1962). A paradigm, for Kuhn, is a prototypical *application* of some set of mathematical, conceptual, or instrumental resources; an application expected to have distinct but similar instances, which it is the job of normal science to discover or construct. Becoming a scientist is less a matter of learning a set of laws than it is a matter of mastering the details of the prototypical applications of the relevant resources in such a way that one can recognize and generate further applications of a relevantly similar kind.

Kuhn was criticized for the vagueness of the notion of a paradigm, and for the unexplicated criterion of similarity that clustered further applications around it. But from the perspective of the neurocomputational approach at issue, he can be vindicated on both counts. For a brain to command a paradigm is for it to have settled into a weight configuration that produces some well-structured similarity space whose central hypervolume locates the prototypical application(s). And it is only to be expected that even the most reflective subject will be incompletely articulate on what dimensions constitute this highly complex and abstract space, and even less articulate on what metric distributes examples along each dimension. A complete answer to these questions would require a microscopic examination of the subject's brain. That is one reason why exposure to a wealth of examples is so much more effective in teaching the techniques of any science than is exposure to any attempt at listing all the relevant factors. We are seldom able to articulate them all, and even if we were able, listing them is not the best way to help a brain construct the relevant internal similarity space.

Kuhn makes much of the resistance typically shown by scientific communities to change or displacement of the current paradigm. This stubbornness here emerges as a natural expression of the way in which networks learn, or occasionally fail to learn. The process of learning by gradient descent is always threatened by the prospect of a purely *local* minimum in the global error gradient. This is a position where the error messages are not yet zero, but where every *small* change in the system produces even larger errors than those currently encountered. With a very high-dimensional space, the probability of there being a simultaneous local minimum in every dimension of the weight space is small: there is usually some narrow cleft in the canyon out which the configuration point can eventually trickle, thence to continue its wandering slide down the error gradient and toward some truly global minimum. But genuine local minima do occur, and the only way to escape them once caught is to introduce some sort of random noise into the system in hopes of bouncing the system's configuration point out of such tempting cul-de-sacs. Furthermore, even if a local quasi-minimum does have an escape path along one or more dimensions, the error gradient along them may there be quite shallow, and the system may take a very long time to find its way out of the local impasse.

Finally, and just as importantly, the system can be victimized by a highly biased 'training set'. Suppose the system has reached a weight configuration that allows it to respond successfully to all of the examples in the (narrow and biased) set it has so far encountered. Subsequent exposure to the larger domain of more diverse examples will not necessarily result in the system's moving any significant distance away from its earlier configuration, unless the relative frequency with which it encounters those new and anomalous examples is quite high. For if the encounter frequency is low, the impact of those examples will be insufficient to overcome the gravity of the false minimum that captured the initial training set. The system may require 'blitzing' by new examples if their collective lesson is ever to 'sink in'.

Even if we do present an abundance of the new and diverse examples, it is quite likely that the delta rule discussed earlier will force the system through a sequence of new configurations that perform very poorly indeed when re-fed examples from the original training set. This temporary loss of performance on certain previously 'understood' cases is the price the system pays for the chance at achieving a broader payoff later, when the system finds a new and deeper error minimum. In the case of an artificial system chugging coolly away at the behest of the delta rule, such temporary losses need not impede the learning process, at least if their frequency is sufficiently high. But with humans the impact of such a loss is often more keenly felt. The new examples that confound the old configuration may simply be ignored or rejected in some fashion, or they may be quarantined and made the target of a distinçt and disconnected learning process in some adjacent network. Recall the example of sublunary and superlunary physics.

This raises the issue of explanatory unity. A creature thrown unprepared into a complex and unforgiving world must take its understanding wherever it can find it, even if this means generating a disconnected set of distinct similarity spaces, each providing the creature with a roughly appropriate response to some of the more pressing types of situation it typically encounters. But far better if it then manages to generate a single similarity space that unifies and replaces the variation that used to reside in two entirely distinct and smaller spaces. This provides the creature with an effective grasp on the phenomena that lay *between* the two classes already dealt with, but which were successfully comprehended by neither of the two old spaces. These are phenomena that the creature had to ignore, or avoid, or simply endure. With a new and more comprehensive similarity space now generating systematic responses to a wider range of phenomena, the creature has succeeded in a small piece of conceptual unification.

The payoff here recalls the virtue earlier discovered for simplicity. Indeed, it is the same virtue, namely, superior generalization to cases beyond those already encountered. This result was achieved, in the case described in section 5, by reducing the number of hidden units, thus forcing the system to make more efficient use of the representational resources remaining. This more efficient use is realized when the system partitions its activation-vector space into the minimal number of distinct similarity subspaces consistent with reducing the error messages to a minimum. When completed, this process also produces the maximal *organization* within and among those subspaces, for the system has found those enduring dimensions of variation that successfully unite the diversity confronting it.

Tradition speaks of developing a single 'theory' to explain everything. Kuhn (1962) speaks of extending and articulating a 'paradigm' into novel domains. Kitcher (1981, 1989) speaks of expanding the range of application of a given 'pattern of argument'. It seems to me that we might unify and illuminate all of these notions by thinking in terms of the evolving structure of a hidden-unit activation-vector space, and its development in the direction of representing all input vectors somewhere within a single similarity space.

This might seem to offer some hope for a Convergent Realist position within the philosophy of science, but I fear that exactly the opposite is the case. For one thing, nothing guarantees that we humans will avoid getting permanently stuck in some very deep but relatively local error minimum. For another, nothing guarantees that there exists a possible configuration of weights that would reduce the error messages to *zero*. A unique global error minimum relative to the human neural network there may be, but for us and for any other finite system interacting with the real world, it may always be nonzero. And for a third thing, nothing guarantees that there is only *one* global minimum. Perhaps there will in general be many quite different minima, all of them equally low in error, all of them carving up the world in quite different ways. Which one a given thinker reaches may be a function of the idiosyncratic details of its learning history. These considerations seem to remove the goal itself—a unique truth—as well as any sure means of getting there. Which suggests that the proper course to pursue in epistemology lies in the direction of a highly naturalistic and pluralistic form of pragmatism. For a running start on precisely these themes, see Munevar 1981 and Stich 1989.

VIII. Concluding Remarks

This essay opened with a survey of the problems plaguing the classical or 'sentential' approach to epistemology and the philosophy of science. I have tried to sketch an alternative approach that is free of all or most of those problems, and has some novel virtues of its own. The following points are worth noting. Simple and relatively small networks of the sort described above have already demonstrated the capacity to learn a wide range of quite remarkable cognitive skills and capacities, some of which lie beyond the reach of the older approach to the nature of cognition (e.g., the instantaneous discrimination of subtle perceptual qualities, the effective recognition of similarities, and the real-time administration of complex motor activity). While the specific learning algorithm currently used to achieve these results is unlikely to be the brain's algorithm, it does provide an existence proof: by procedures of this general sort, networks can indeed learn with fierce efficiency. And there are many other procedures awaiting exploration.

The picture of learning and cognitive activity here painted encompasses the entire animal kingdom: cognition in human brains is fundamentally the same as cognition in brains generally. We are all of us processing activation vectors through artfully weighted networks. This broad conception of cognition 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.

Conceptual change is no longer a problem: it happens continuously in the normal course of all cognitive development. It is sustained by many small changes in the underlying hardware of synaptic weights, which changes gradually repartition the activation-vector spaces of the affected population of cells. Conceptual *simplicity* is also rather clearer when viewed from a neurocomputational perspective, both in its nature and in its epistemological significance.

The old problem of how to retrieve relevant information is transformed by the realization that it does not need to be 'retrieved'. Information is stored in brainlike networks in the global pattern of their synaptic weights. An incoming vector activates the relevant portions, dimensions, and subspaces of the trained network by virtue of its own vectorial makeup. Even an incomplete version of a given vector (i.e., one with several elements missing) will often provoke essentially the same response as the complete vector by reason of its relevant similarity. For example, the badly whistled first few bars of a familiar tune will generally evoke both its name and the rest of the entire piece. And it can do this in a matter of milliseconds, because even if the subject knows thousands of tunes, there are still no lists to be searched.

It remains for this approach to comprehend the highly discursive and linguistic dimensions of human cognition, those that motivated the classical view of cognition. We need not pretend that this will be easy, but we can see how to start. We can start by exploring the capacity of networks to manipulate the structure of existing language, its syntax, its semantics, its pragmatics, and so forth. But we might also try some novel approaches, such as allowing each of two distinct networks, whose principal concerns and activities are nonlinguistic, to try to learn from scratch some systematic means of manipulating, through a proprietary dimension of input, the cognitive activities of the other network. What system of mutual manipulation—what *language*—might they develop?

The preceding pages illustrate some of the systematic insights that await us if we adopt a more naturalistic approach to traditional issues in epistemology, an approach that is grounded in computational neuroscience. However, a recurring theme in contemporary philosophy is that normative epistemology *cannot* be 'naturalized' or reconstructed within the framework of any purely descriptive scientific theory. Notions such as 'justified belief' and 'rationality', it is said, cannot be adequately defined in terms of the nonnormative categories to which any natural science is restricted, since "oughts" cannot be derived from "ises". Conclusions are then drawn from this to the principled autonomy of epistemology from any natural science.

While it may be true that normative discourse cannot be replaced without remainder by descriptive discourse, it would be a distortion to represent this as the aim of those who would naturalize epistemology. The aim is rather to enlighten our normative endeavors by reconstructing them within a more adequate conception of what cognitive activity consists in, and thus to free ourselves from the burden of factual misconceptions and tunnel vision. It is only the *autonomy* of epistemology that must be denied.

Autonomy must be denied because normative issues are never independent of factual matters. This is easily seen for our judgments of instrumental value, as these always depend on factual premises about causal sufficiencies and dependencies. But it is also true of our most basic normative concepts and our judgments of intrinsic value, for these have factual presuppositions as well. We speak of *justification*, but we think of it as a feature of *belief*, and whether or not there are any beliefs and what properties they have is a robustly factual matter. We speak of *rationality*, but we think of it as a feature of *thinkers*, and it is a substantive factual matter what thinkers are and what cognitive kinematics they harbor. Normative concepts and normative convictions are thus always hostage to some background factual presuppositions, and these can always prove to be superficial, confused, or just plain wrong. If they are, then we may have to rethink whatever normative framework has been erected upon them. The lesson of the preceding pages is that the time for this has already come.

References

Barto, A. G. 1985. Learning by Statistical Cooperation of Self-Interested Neuronlike Computing Elements. *Human Neurobiology* 4:229–56.

Bear, M. F., Cooper, L. N., and Ebner, F. F. 1987. A Physiological Basis for a Theory of Synapse Modification. *Science* 237 (no. 4810).

Churchland, P. M. 1975. Karl Popper's Philosophy of Science. *Canadian Journal of Philosophy* 5 (no. 1).

—— 1979. *Scientific Realism and the Plasticity of Mind.* Cambridge: Cambridge University Press.

——. 1981. Eliminative Materialism and the Propositional Attitudes. *Journal of Philosophy. 78 (no. 2).*

——. 1985. The Ontological Status of Observables: In Praise of the Superempirical Virtues. In *Images of Science*, ed. P. M. Churchland and C. A. Hooker. Chicago, University of Chicago Press.

——. 1986. Some Reductive Strategies in Cognitive Neurobiology. *Mind* 95 (no. 379).

—— 1988. Perceptual Plasticity and Theoretical Neutrality: A Reply to Jerry Fodor. *Philosophy of Science* 55 (no. 2).

Churchland, P. S. 1980. A Perspective on Mind-Brain Research. *Journal of Philosophy* 77 (no. 4).

——. 1986. *Neurophilosophy: Toward a Unified Understanding of the Mind-Brain*. Cambridge, MIT Press.

Feyerabend, P. K. 1965. Reply to Criticism: Comments on Smart, Sellars, and Putnam. In *Boston*

Studies in the Philosophy of Science, ed. M. Wartofsky. Dordrecht: Reidel. Reprinted in *Realism, Rationalism & Scientific Method, Philosophical Papers, vol. 1, Feyerabend, P. K. 1981. Cambridge: Cambridge University Press, 1981.*

——. *1980. Consolations for the Specialist. In Criticism and the Growth of Knowledge*, eds. I. Lakatos and A. Musgrave. Cambridge: Cambridge University Press.

Fodor, J. A. 1984. Observation Reconsidered. *Philosophy of Science* 51 (no. 1).

——. 1988. A Reply to Churchland's "Perceptual Plasticity and Theoretical Neutrality." *Philosophy of Science* 55 (no. 2).

Giere, R. 1988. *Explaining Science: A Cognitive Approach.* Chicago: University of Chicago Press.

Glymour, C. 1987. "Artificial Intelligence is Philosophy". In *Aspects of Artificial Intelligence*, ed. J. Fetzer. Dordrecht: Reidel.

Gorman, R. P., and Sejnowski, T. J. 1988. Learned Classification of Sonar Targets Using a Massively-Parallel Network. *IEEE Transactions: Acoustics, Speech, and Signal Processing.* Forthcoming.

Hempel, K. 1965. "Studies in the Logic of Confirmation". In *Aspects of Scientific Explanation.* New York: The Free Press.

Hinton, G. E., and Sejnowski, T. J. 1986. "Learning and Relearning in Boltzmann Machines". In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, eds. D. E. Rumelhart and J. L. McClelland. Cambridge: MIT Press. 1986.

Hooker, C. A. 1975. The Philosophical Ramifications of the Information-Processing Approach to the Mind-Brain. *Philosophy and Phenomenological Research* 36.

——. 1987. *A Realistic Theory of Science.* Albany: State University of New York Press.

Hopfield, J. J., and Tank, D. 1985. "Neural" Computation of Decisions in Optimization Problems. *Biological Cybernetics* 52:141–52.

Hubel, D. H., and Wiesel, T. N. 1962. Receptive Fields, Binocular Interactions, and Functional Architecture in the Cat's Visual Cortex. *Journal of Physiology* 160.

Kitcher, P. 1981. Explanatory Unification. *Philosophy of Science* 48 (no. 4).

——. 1989. "Explanatory Unification and the Causal Structure of the World". In *Minnesota Studies in the Philosophy of Science, vol. 13, Scientific Explanation*, ed. P. Kitcher. Minneapolis: University of Minnesota Press.

Kuhn, T. S. 1962. *The Structure of Scientific Revolutions.* Chicago: University of Chicago Press.

Lakatos, I. 1970. "Falsification and the Methodology of Scientific Research Programmes. In *Criticism and the Growth of Knowledge*, I. Lakatos and A. Musgrave. Cambridge University Press.

Laudan, L. 1981. A Confutation of Convergent Realism. *Philosophy of Science* 48 (no. 1).

Lehky, S., and Sejnowski, T. J. 1988a. "Computing Shape from Shading with a Neural Network Model". In *Computational Neuroscience*, ed. E. Schwartz. Cambridge: MIT Press.

——. 1988b. Network Model of Shape-From-Shading: Neural Function Arises from Both Receptive and Projective Fields. *Nature* 333 (June 2).

Linsker, R. 1986. From Basic Network Principles to Neural Architecture: Emergence of Orientation Columns. *Proceedings of the National Academy of Sciences, USA*, 83:8779–83.

Minsky, M., and Papert, S. 1969. *Perceptrons.* Cambridge: MIT Press.

Munevar, G. 1981. *Radical Knowledge.* Indianapolis: Hackett.

Pellionisz, A., and Llinas, R. 1982. Space-Time Representation in the Brain: The Cerebellum as a Predictive Space-Time Metric Tensor. *Neuroscience* 7 (no. 12):2949–70.

——. 1985. Tensor Network Theory of the Metaorganization of Functional Geometries in the Central Nervous System. *Neuroscience.* 16 (no. 2):245–74.

Putnam, H. 1981. *Reason, Truth, and History.* Cambridge: Cambridge University Press.

Rosenberg, C. R., and Sejnowski, T. J. 1987. Parallel Networks That Learn To Pronounce English Text. *Complex Systems, 1:145*–68.

Rosenblatt, F. 1959. *Principles of Neurodynamics.* New York: Spartan Books.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. 1986a. Learning Representations by Back-Propagating Errors. *Nature*, 323.

——. 1986b. "Learning Internal Representations by Error Propagation". In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, ed. D. E. Rumelhart and J. L. McClelland. Cambridge: MIT Press.

Salmon, W. 1966. *The Foundations of Scientific Inference*. Pittsburgh: University of Pittsburgh Press.

Scheffler, I. 1963. *The Anatomy of Inquiry*. New York: Knopf.

Sejnowski, T. J., Kienker, P. K., and Hinton, G. E. 1986. Learning Symmetry Groups with Hidden Units: Beyond the Perceptron. *Physica D*: 22.

Stich, S. P. 1990. *The Fragmentation of Reason*. Cambridge: MIT Press.

Suppe, F. 1974. *The Structure of Scientific Theories*. Chicago: University of Illinois Press.

Van Fraassen, Bas 1980. *The Scientific Image*. Oxford: Oxford University Press.

Zipser, D., and Elman, J. D. 1988. Learning the Hidden Structure of Speech. *Journal of the Acoustical Society of America* 83(4):1615–25.