General Article

NETWORKS AND THEORIES: The Place of Connectionism in Cognitive Science

by Michael McCloskey

This article considers how connectionist modeling can contribute to understanding of human cognition. I argue that connectionist networks should not be thought of as theories or simulations of theories, but may nevertheless contribute to the development of theories.

It is no exaggeration to say that connectionism has exploded into prominence within cognitive science over the last several years. The enormous upsurge of interest is attested by conferences and symposia too numerous to list, new journals (e.g., *Neural Networks, Neural Computation*), special issues of existing journals (e.g., *Cognition, Journal of Memory and Language, The Southern Journal of Philosophy*), and a virtual avalanche of monographs and edited volumes (e.g., Anderson & Rosenfeld, 1988; Barnden & Pollack, 1991; Bechtel & Abrahamsen, 1991; Clark, 1990; McClelland & Rumelhart, 1986; Morris, 1989; Pinker & Mehler, 1988; Rumelhart & McClelland, 1986).

In this article I discuss an issue that has sparked considerable debate (Bechtel, 1987; Estes, 1988; Fodor & Pylyshyn, 1988; Massaro, 1988; McClelland, 1988; Minsky & Papert, 1988; Pinker & Prince, 1988; Smolensky, 1987, 1988): What is the proper role of connectionism in cognitive science? This issue is complex and multifaceted, and I will not endeavor to explore all of its ramifications (nor am I capable of doing so). Instead, I attempt to develop two specific points. First, I suggest that connectionist networks should not be viewed as theories of human cognitive functions, or as simulations of theories, or even as demonstrations of specific theoretical points. Second, I argue that these networks nevertheless hold considerable promise as tools for development of cognitive theories.

My remarks are directed specifically at the class of connectionist models that has generated the greatest interest among cognitive scientists: the class involving networks in which (1) concepts have distributed network representations; (2) the network includes hidden as well as visible units; and (3) the connection weights encoding the network's "knowledge" are not specified directly by the modeler but rather are established through application of a learning algorithm (e.g., back propagation, or the Boltzmann Machine algorithm). The most well-developed model of this class, and the one I will take as an example, is the Seidenberg and McClelland (1989) model of word recognition and naming (see also Seidenberg, 1989; Patterson, Seidenberg, & McClelland, 1989). This model addresses the ability of literate humans to categorize a letter string (e.g., *house, boke*) as a word or nonword, as well as the ability to read both words and nonwords aloud.

Seidenberg and McClelland (1989) trained a three-layer network with the back-propagation learning algorithm. The network maps a distributed orthographic representation of a word or nonword onto a distributed phonological representation, which is presumed to provide the basis for a spoken naming response. The network also generates an orthographic output representation, and the extent to which this representation duplicates the initial network input is taken as a basis for classifying the stimulus as a word or nonword (i.e., lexical decision). Seidenberg and McClelland (1989) argue that the network performs well at lexical decision and naming, and also exhibits several specific phenomena obtained in studies with human subjects (e.g., an interaction of word frequency and regularity in reaction time for naming words). The question to be considered here is the following: In what way can networks of this sort contribute to our understanding of human cognition?

SIMULATION VERSUS EXPLANATION

In exploring this question, it is worthwhile first to distinguish between simulating and explaining human cognitive performance. Suppose I present you with a black box connected to a keyboard and video monitor. When you enter a letter sequence at the keyboard, the device displays on the monitor a word/nonword classification, a sequence of phonetic symbols, and a reaction time for each of these outputs. Thus, imagine that given the input LAUGH, the device displayed the following output:

Address correspondence to: Michael McCloskey, Cognitive Science Department, Johns Hopkins University, Baltimore, MD 21218; email: m_mcclos@jhuvms.bitnet.

Word/Nonword Classification: word (487 msec)
Phonological Representation: /læf/ (654 msec)

Suppose further that you test the device extensively by presenting many word and nonword stimuli, and find that its performance matches reasonably well, although not perfectly, the performance of human subjects. That is, it usually though not always generates correct word/nonword classifications and pronunciations; it shows the interaction of frequency and regularity, and so forth.

Certainly this would be an interesting and impressive device. But would it, as a black box, constitute a theory of human word recognition and naming? Would it explain the ability of humans to recognize and name words, or specific phenomena such as the frequency–regularity interaction? Obviously, the answer is no. Before you would credit me with having offered an explanatory theory, I would presumably have to describe how the black box worked, and what aspects of its structure and functioning were to be counted as relevant to modeling human performance. And, of course, this description would have to be in a form that made clear to you how the device was able to generate pronunciations and word/nonword decisions, and how its functioning gave rise to particular phenomena. You might then consider the description a theory of word recognition and naming, and the device a simulation of the theory.

The point here is simple and presumably obvious: Although the ability of a connectionist network (or other computational device) to reproduce certain aspects of human performance is interesting and impressive, this ability alone does not qualify the network as a theory, and does not amount to explaining the performance (see also Cummins & Schwarz, 1987; Humphreys, 1989).

SIMULATION IN SEARCH OF THEORY

Suppose I now tell you that the black box encloses a connectionist network of the form described by Seidenberg and McClelland (1989): 400 orthographic units that send activation to 200 hidden units, which in turn send activation to 460 phonological units, and also feed activation back to the orthographic units. Suppose further that I specify the activation rule for the units, the way in which letter strings and pronunciations are represented across the orthographic and phonological units, the learning algorithm used in training the network, and the training regimen (e.g., the corpus of words presented to the network, the procedures for determining frequency and order of presentation). Suppose finally that I specify how the pattern of activation generated by the network across the phonological and orthographic units in response to a stimulus is used to determine a word/nonword decision, a pronunciation, and a latency for each of these outputs.

Does this description of the network now constitute a specific theory of word recognition and naming, a theory simulated by the network itself? Certainly the temptation is to say yes. After all, I have provided (let us say) a precise and detailed description of network architecture, input and output representations, the functioning of individual units, and the learning procedures. Upon reflection, however, it becomes clear that my description falls short of being a theory in two critical respects.

Separating Wheat from Chaff

First, I have not specified what aspects of the network are to be considered relevant, and what aspects are to be considered irrelevant, for modeling human lexical processing. For example, is the specific form of orthographic representation implemented in the network integral to the theory, or is this representation an instance of a more general class of representations, such that any member of the class could be considered equivalent from the standpoint of the theory? And, if the latter, how is the class to be defined? In other words, what features of the representation are relevant to the theory, and what features are irrelevant?

Similarly, which aspects of network architecture are crucial to the theory, and which aspects are incidental features of the particular simulation? Is it significant, for instance, that the same set of hidden units is involved in generating the orthographic and phonological output? Further, what features of the learning procedures are relevant and irrelevant to the theory? Is it important that the network was trained with the back-propagation algorithm? Should any significance be attached to the particular settings of parameters such as the learning rate and momentum? And so forth and so on. A description of a particular network cannot provide answers to questions of this sort, and hence cannot be considered a theory. In describing the network, then, I have at best characterized a simulation, without specifying what theory it simulates.

Elucidating Cognitive Processes

My description of the network also falls short as a theory of human word recognition and naming in that it fails to provide a specific account of how word–nonword discrimination and orthography–phonology conversion are carried out. In describing network architecture, input and output representations, the functioning of individual units, and the learning procedures, I have presented important information. However, this information is not

sufficient to explicate the structure and functioning of the network at a level relevant for understanding human word recognition and naming (Grossberg, 1987; see also Chandrasekaran, Goel, & Allemang, 1988). What knowledge is encoded in the network's connection weights? How is this knowledge distributed among the weights, and how is the knowledge used in converting inputs to outputs? For example, what transformation is accomplished in the mapping from orthographic to hidden units, or from hidden to phonological units? My description of the network does not answer these sorts of questions, and hence cannot be said to explain how the network (or a person) discriminates words from nonwords, or reads words and nonwords aloud.[1]

THEORIES AND SIMULATIONS

If, then, one's goal is to present a connectionist theory of a human cognitive function, along with a network simulating the theory, one must offer more than a detailed description of a particular network.

Stating a Theory

First, one must formulate the theory at a level more abstract than that of a particular network simulation, identifying the theory's claims about how the cognitive function is carried out, and how specific phenomena arise. This is not to say that the theory must be stated in terms of rules operating on symbolic representations; the appropriate theoretical vocabulary for connectionist theories remains an open issue. What is critical, however, is that the formulation be adequate to fulfill the functions we expect a theory to fulfill. For example, a theory should organize and make sense of the available observations concerning the cognitive process in question by allowing generalizations to be stated (e.g., because orthography–phonology conversion is accomplished in such-and-such a way, variable x should affect naming time and accuracy as follows, whereas variable y should be irrelevant). Similarly, a theory should support clear-cut credit and blame assignment. That is, when the theory successfully predicts some phenomenon, it should be possible to identify the aspects of the theory important for generating the prediction; and when the theory makes an incorrect prediction, it should be possible to relate the failure to specific assumptions of the theory, and hence to assess the extent to which the failure is fundamental. Further, the statement of a theory should provide a basis for discerning its important similarities to, and differences from, alternative theories in the same domain. Only if a theory fulfills these sorts of functions can it serve as a foundation for further theoretical and empirical progress.

Tying Simulation to Theory

In presenting a connectionist theory-plus-simulation, it is important not only to provide an appropriate statement of the theory, but also to establish that the network simulation is relevant to the theory. In particular, one must demonstrate that the network actually instantiates the theory's assumptions about the cognitive function in question. Further, the characterization of network structure and functioning should provide a basis for assessing the extent to which successes or failures in simulating particular phenomena reflect theory-relevant or theory-irrelevant features of the network.

CONNECTIONIST THEORIZING AND SIMULATION

If we now consider current connectionist work in light of the preceding discussion, two points become apparent. First, although explicit theories of human cognitive processes could conceivably be developed within the connectionist framework, this potential remains unrealized at present. Second, attempts to tie theoretical claims to network implementations face a serious obstacle in the form of limited understanding of complex connectionist networks.

Theoretical Proposals

Connectionist work on human cognition has led to some interesting and innovative theoretical proposals. However, these proposals (as distinguished from descriptions of particular networks) are in most instances rather

1. These difficulties cannot be resolved simply by providing more details about the network. Suppose that in addition to the above-mentioned information, I also present all of the connection weights established by the training procedure. Although I have now provided all of the information needed to duplicate the network exactly, my description would still not constitute a theory of word recognition and naming. First, a listing of connection weights would not identify the theory-relevant features of the network, and in fact would introduce additional problems of separating crucial from incidental properties. Presumably, the configuration of weights generated in a particular network training run is merely an exemplar of a more general class, any member of which would count as an implementation of the same theory of word recognition and naming. However, a listing of weights does not specify what properties of a weight configuration are crucial for instantiating a theory, and what properties are irrelevant. Furthermore, merely listing the large number of weights would not illuminate how the network performed orthography–phonology conversion, or word/nonword discrimination.

vague and general, and do not amount to explicit theories of human cognitive functions.[2] This point may be illustrated by referring once again to the Seidenberg and McClelland (1989) model, widely regarded as one of the most fully developed connectionist cognitive models.

Seidenberg and McClelland offer several important theoretical claims. They assume that word recognition and naming involve distributed orthographic and phonological representations, such that orthographically similar words have overlapping orthographic representations, and phonologically similar words have overlapping phonological representations. Further, they argue that contrary to the claims of "dual-route" theories, a single mechanism suffices to map orthographic to phonological representations for regular words (e.g., *dog*), irregular words (e.g., *yacht*), and nonwords (e.g., *mave*). In particular, they argue that the mapping is accomplished without representations of specific lexical items, via connection weights encoding statistical regularities as well as idiosyncracies in relationships between the orthography and phonology of English words.

As these examples indicate, Seidenberg and McClelland's theoretical proposals are novel and substantive. However, these proposals do not add up to an explicit theory of human word recognition and naming. For one thing, Seidenberg and McClelland do not specify just what regularities and idiosyncracies are encoded through experience with words, how the acquired knowledge is distributed over a set of connection weights, or how the appropriate knowledge is brought into play in just the appropriate circumstances. For example, exactly what do people learn about the phonological correspondences of the letter *a* in various contexts? How is the knowledge represented in a network of simple processing units? And how does propagation of activation through the network compute the appropriate phonological instantiation of *a* in the case of regular words like *hat* or *hate*, regular inconsistent words like *gave*, exception words like *have*, and nonwords like *mab* or *mave?* The Seidenberg and McClelland theory is not sufficiently well-developed to provide specific answers. Similarly, at the level of theory (as opposed to simulation) Seidenberg and McClelland's claims about the form of orthographic and phonological representations are limited to the rather vague and general assumptions described above (i.e., representations are distributed, and similar words have similar representations).

Consult the simulation?

It might be objected that Seidenberg and McClelland neither intended nor needed to offer a complete verbal formulation of their theory, because the details are provided by the implemented simulation. However, we cannot look to the network simulation to fill gaps in Seidenberg and McClelland's verbal statement of their theory. In the first place, any simulation includes theory-irrelevant as well as theory-relevant details; hence, the details of a simulation cannot be identified straightforwardly with the details of the corresponding theory. For example, Seidenberg and McClelland (1989, p. 563) are careful to emphasize that they are not committed at the level of theory to the specific orthographic and phonological representations implemented in the simulation.

A second reason we cannot use the simulation to flesh out the verbal formulation of the theory is that our understanding of the network's "knowledge" and functioning is quite limited. We do not know what regularities and idiosyncracies are captured in the network, how this information is reflected in the weights, and so forth. In other words, our understanding of how the network accomplishes word–nonword discrimination and orthography–phonology mapping is no more detailed than the description of these processes in Seidenberg and McClelland's statement of their theory.

The difficulty is not simply that Seidenberg and McClelland failed to describe in sufficient detail the network's encoding of knowledge and its functioning as a whole. Rather, the problem is that connectionist networks of any significant size are complex nonlinear systems, the dynamics of which are extremely difficult to analyze and apprehend (Dyer, 1988; Grossberg, 1987; Minsky & Papert, 1988; Pavel, 1990; Rager, 1990). At present, understanding of these systems is simply inadequate to support a detailed description of a network's knowledge and functioning. In several recent studies attempts have been made to analyze the internal representations established through training of a complex network—for example, by applying cluster analysis techniques to patterns of activation across hidden units elicited by different inputs (see, e.g., Gorman & Sejnowski, 1988; Hanson & Burr, 1990; Sejnowski & Rosenberg, 1987). These important studies reflect a recognition of the need to understand connectionist networks more thoroughly. In fact, in a later section I suggest that the promise of connectionism lies in work of this genre. For present purposes, however, the relevant point is that techniques for network analysis are currently rather crude. Although these techniques have yielded some interesting insights, they are not adequate to provide a de-

2. Further, it is not clear to what extent the theoretical proposals are dependent upon the connectionist view that cognition involves propagation of activation within networks of simple processing units. That is, the contributions of connectionist theorists may have more to do with their insights into the abstract nature of cognitive processes (e.g., that many such processes involve satisfaction of multiple soft constraints) than with the specific connectionist conceptions of the computational machinery instantiating these processes.

tailed understanding of a network's knowledge or functioning.[3] Thus, although we may understand in detail the architecture, input and output representations, individual unit functioning, and learning procedures for the Seidenberg and McClelland (1989) network, we can achieve at best a vague and general understanding of how the network accomplishes word–nonword discrimination, or orthography–phonology conversion. This point may perhaps be brought home by noting that the Seidenberg and McClelland network incorporates more than 1,000 units, and more than 250,000 weighted connections between units. Seidenberg and McClelland's (1989) network simulation cannot, then, remedy the vagueness in the verbal formulation of their theory.

The nature of connectionist simulations

It may seem odd or even obviously incorrect to describe the Seidenberg and McClelland (1989) theory as vague, given that the ability to instantiate a theory in a computer simulation is generally taken to indicate that the theory has been specified explicitly. However, connectionist modeling is not simulation in the traditional sense. A modeler developing a traditional computer simulation must build in each of the crucial features of an independently specified theory. If the theory is not explicitly formulated, the simulation cannot be built. In connectionist modeling, on the other hand, the modeler may be able to proceed on the basis of vague and fragmentary theoretical notions, because much of the work is left to the learning algorithm. In a sense the modeler "grows" the network rather than building it. And, just as a gardener's ability to grow a plant from a seed does not imply that the gardener has an explicit theory of plant physiology, the ability to grow a network that mimics in some respects a human cognitive function does not demonstrate that the modeler has an explicit theory of that function. In essence, the learning algorithms constitute procedures for creating complex systems we do not adequately understand.[4]

Development of a connectionist simulation might be taken as evidence of an explicit theory if the modeler built the simulation by specifying the connection weights directly, rather than growing the simulation by delegating this work to the learning algorithm. However, neither the Seidenberg and McClelland theory nor our understanding of complex connectionist networks is sufficiently detailed to sustain such an endeavor.

Relating Simulations to Theories

Limited understanding of complex connectionist networks also represents a serious problem for attempts to establish that a network simulation is germane to a set of theoretical claims. Even given an explicit theory, it may be difficult or impossible to (1) determine whether a particular network actually instantiates the theory's assumptions, or (2) assess the extent to which theory-relevant as opposed to theory-irrelevant features of the network are responsible for its successes and failures in reproducing phenomena.

Does the network simulate the theory?

Consider as an example Seidenberg and McClelland's (1989) theoretical claim that a single mechanism accomplishes orthography-to-phonology mapping for words and nonwords. Given this claim, it is important to consider whether the simulation actually performs orthography–phonology conversion for words and nonwords with a single mechanism. In one sense it certainly does, because processing of both words and nonwords is accomplished by a single network. However, one could argue by a similar logic that all human cognitive processes involve a single mechanism, because all are implemented by a single organ (i.e., the brain).

Thus, the issue is not whether there is some level of description at which the Seidenberg and McClelland (1989) network handles words and nonwords with a single mechanism. Rather, the question is whether the network, when characterized at levels relevant for cognitive theorizing (i.e., levels appropriate for stating generalizations about word recognition and naming, for establishing credit and blame, etc.), accomplishes orthography–phonology conversion in the same way for words and nonwords. As I have repeatedly emphasized, however, our understanding of the network at such levels is extremely limited. Hence, Seidenberg and McClelland's (1989) assertions notwithstanding (see, e.g., p. 549), it is unclear whether the network in any interesting sense employs a single mechanism for naming words and nonwords.

Evaluating successes and failures

Similar difficulties arise in assessing the implications of a simulation's successes and failures in reproducing specific phenomena. In developing any simulation one must make some choices that are arbitrary from the standpoint of the theory, and introduce some simplifica-

3. Seidenberg and McClelland (1989, pp. 540–543) describe efforts to analyze their network's processing of a few specific words. However, these efforts were less formal and systematic than those of the above-cited researchers, and yielded little more than a few vague and fragmentary glimpses of internal network representations.

4. A related point may be made with respect to the learning process itself. The ability to grow a network that reproduces some human cognitive phenomena does not imply that the modeler has an explicit theory of how the cognitive function in question develops. Although we have procedures for training a network, we do not fully understand what the network has learned at the completion of, or at any time during, training. Hence, we can hardly be said to understand the learning process.

tions and idealizations. For example, in the Seidenberg and McClelland (1989) simulation some aspects of the orthographic and phonological representations were essentially arbitrary, as were the settings of several parameters. Further, the phonological representations collapsed some distinctions among phonemes (e.g., the initial sounds in *shin* and *chin* were not distinguished); the network was trained on only 2,897 words, all of which were monosyllabic and most of which were monomorphemic; the procedures for training the network greatly compressed the range of word frequencies actually experienced by a human reader; and so forth.

These arbitrary choices and simplifications do not in and of themselves constitute a weakness of the Seidenberg and McClelland simulation; decisions of this sort must be made in developing any simulation. The difficulties arise from the fact that the functioning of the network is not sufficiently well-understood to assess the consequences of the decisions (although see Seidenberg and McClelland, 1989, for some discussion). Did the network perform as well as it did only, for example, because it was trained on a constrained set of words, or because the frequency range was compressed (Bever, in press; McCloskey & Cohen, 1989)? Similarly, do the network's shortcomings reflect incidental properties of the simulation, or fundamental problems with the Seidenberg and McClelland (1989) theory? Questions of this sort are difficult for any simulation, but they are made doubly difficult by our limited understanding of connectionist networks.

For example, Besner, Twilley, McCann, and Seergobin (1990) argue that the Seidenberg and McClelland network performs much more poorly than human subjects at lexical decision (i.e., word/nonword discrimination). But what are the implications of this deficiency? Does it reflect some arbitrary feature(s) of the network's representations, architecture, or parameter settings, such that theory-neutral modifications could remedy the problem? Or might the responsibility lie in the limited set of words to which the network was exposed, so that training with a larger set would improve its lexical decision performance? Or is the problem more fundamental, reflecting feature(s) of the simulation that are central to the underlying theory (e.g., the putative absence of lexical representations)? The unfortunate fact is that our understanding of the network is insufficient to answer these questions, or even to suggest systematic approaches toward finding answers. We might undertake an empirical exploration of various network modifications chosen on largely pretheoretical grounds—we could, for example, try more (or fewer) hidden units or a different form of orthographic representation. However, if we succeed thereby in improving the network's lexical decision performance the success will be largely due to chance, and if we fail we will not know how to interpret the failure.

CONNECTIONIST NETWORKS AS DEMONSTRATIONS

In response to these arguments it might be suggested that connectionist networks, at least at present, are properly viewed not as simulations of fully developed theories, but rather as more limited demonstrations of specific theoretical points concerning human cognition. For example, even if Seidenberg and McClelland (1989) have not presented an explicit theory, perhaps their network at least demonstrates that naming of words and nonwords can be accomplished without lexical representations, and hence that a theory of human lexical processing need not necessarily postulate such representations.

Unfortunately, this view of connectionist modeling does not escape the above-noted problems of relating network to theory. For example, we do not know whether the Seidenberg and McClelland network, characterized at appropriate levels, in fact lacks lexical representations. Also, whatever the nature of the mechanism(s) implemented in the network, we do not know whether these mechanisms are capable of achieving word-naming performance comparable to that of human subjects under the conditions humans typically encounter (e.g., exposure to at least tens of thousands of different words with widely varying frequencies). Finally, we do not know whether the implemented mechanisms are adequate to reproduce other human phenomena that have been thought to require lexical representations. (Indeed, the current network's poor lexical decision performance provides some grounds for skepticism on this point.) Thus, the Seidenberg and McClelland results do not demonstrate (for lexical processing in general, or even for naming in particular) that performance comparable to that of humans can be achieved in the absence of lexical representations.

A ROLE FOR CONNECTIONIST MODELING

Am I suggesting, then, that connectionist modeling has no role to play in cognitive science? Definitely not. In my view connectionist models hold substantial promise as tools for developing cognitive theories, if viewed from an appropriate perspective. Specifically, it may prove fruitful to think of connectionist models as akin to animal models of human functions or disorders (e.g., an animal model of working memory, or an animal model of attention deficit disorder).

Animal Models

In work with an animal model, some animal system thought to share critical features with a human system of interest is studied with the aim of shedding light on the

human system. By studying the animal model rather than working directly with the human system, one may be able to carry out manipulations that could not be performed on human subjects (e.g., lesions to particular brain structures, histological examination of brain tissue). The model system may also be somewhat simpler than the human system, and therefore more amenable to analysis.

Thus, an animal model is not a theory (or a simulation of a theory), but rather an object of study. In work with an animal model the goal is to elucidate the structure and functioning of the animal system, and on this basis to formulate a theory of the corresponding human system. Of course, one does not assume that insights gained through study of the animal system will necessarily apply without modification to the human system. Instead, it is simply assumed that because the animal system may be similar in relevant respects to the human system, studying the former may aid in developing a theory of the latter.

Connectionist Networks as Analogues to Animal Models

Connectionist networks should, I suggest, be viewed from a similar perspective. A network that exhibits some of the phenomena observed for a cognitive process such as naming may perhaps resemble in relevant respects the mechanisms underlying the process in humans. If by studying the network we can gain some understanding of its structure and functioning at a level relevant for cognitive theorizing, this understanding might aid in developing a theory of the human cognitive process (see Cummins & Schwarz, 1987, for a similar suggestion). For example, if one could determine how the Seidenberg and McClelland (1989) network accomplishes word–nonword discrimination and orthography–phonology mapping, this information might contribute to development of an explicit theory of human word recognition and naming.[5]

Connectionist networks share both the advantages and disadvantages of animal models. On the positive side, networks can be subjected to manipulations that are not possible with human subjects. For example, one can observe the effects of varying the network architecture while holding the training process constant; one can subject a network to several different forms of damage, and restore it to its "premorbid" state; one can inspect connection weights and activation patterns across hidden units; and so forth. On the other side of the ledger, a network that reproduces human phenomena in some cognitive domain may not actually resemble in critical respects the mechanisms mediating human performance (Hanson & Burr, 1990; Lamberts & d'Ydewalle, 1990). Perhaps, for example, the Seidenberg and McClelland (1989) network does not accomplish word recognition or naming in anything like the manner of humans. If this were the case, then analyses of the network might be uninformative or even misleading with respect to human lexical processing. Thus, careful attention must be paid to issues concerning the extent to which a network shares critical features with the human mechanisms of interest.[6]

It must also be emphasized that connectionist modeling does not represent an atheoretical procedure for discovering cognitive theories. In the absence of at least some general theoretical premises, a modeler would be reduced to random exploration of the huge model space defined by possible network architectures, representations, training procedures, and so forth. Similarly, analyses of network functioning taking the form of atheoretical fishing expeditions are unlikely to prove fruitful (Hanson & Burr, 1990). Modeling is an aid to, but not a substitute for, theoretical work (Olson & Caramazza, 1991).

The Practice of Connectionist Modeling

To a large extent, connectionist cognitive modeling efforts have focused on demonstrating that a network can recapitulate certain human phenomena in the domain of interest. Analysis of network functioning has been considered interesting, but perhaps not essential to the enterprise (although see Hanson & Burr, 1990). The analogy to animal models suggests that some refocusing of effort is needed. Demonstrations that a network reproduces human phenomena are certainly important, as such demonstrations may contribute to assessing whether a network is similar in relevant respects to the human cognitive mechanisms under investigation. However, attention must also be directed toward elucidating the structure and functioning of the network, and applying the insights gained thereby in developing an explicit theory of the human mechanisms. On this view, analyses of the network at levels relevant for cognitive theorizing are

5. It is important to emphasize that in speaking here of theories, I have in mind functional as opposed to neural theories. It is not at all clear that connectionist networks resemble neural circuits in ways relevant to understanding the implementation of cognitive processes in the brain (e.g., Douglas & Martin, 1991; Olson & Caramazza, 1991).

6. It is worth noting that even a network exhibiting performance clearly discrepant from that of human subjects might nevertheless contribute to understanding of the human system. If the inability of a network to reproduce certain human phenomena could be tied to particular features of the network, then the failure might be interpreted as evidence that the human system differs from the network in these features. For example, if the poor lexical decision performance of the Seidenberg and McClelland (1989) network turned out to reflect the network's (putative) lack of lexical representations, the network's performance might be taken to suggest that an adequate theory of human lexical processing will need to postulate such representations (see Besner et al., 1990).

crucial. Further, the ultimate aim of the endeavor is not merely to understand the network, but rather to develop a theory that characterizes human cognitive mechanisms independent of any particular network.

Prospects for Connectionism

It remains to be seen how fruitful connectionist modeling will prove in advancing our understanding of human cognition. The capabilities of connectionist networks are impressive, and study of these networks appears to hold promise for shedding light on such features of human cognition as content-addressable memory, learning, categorization (Estes, 1988, 1991; Gluck & Bower, 1988), and effects of brain damage (Hinton & Shallice, in press; McCloskey & Lindemann, in press; Patterson et al., 1989). However, in assessing the prospects for connectionism there are at least two imponderables. First, it is not clear how fast and how far we will progress in attempting to analyze connectionist networks at levels relevant for cognitive theorizing. Indeed, Suppes (1990) offers the pessimistic conjecture that connectionist networks may be computationally irreducible, such that "no analysis in terms smaller than the nets themselves will give anything like a really detailed and accurate picture of how they work" (p. 508).

The second, and closely related, imponderable concerns the terms in which a connectionist theory should be stated. As we have seen, a description of network architecture, input and output representations, individual unit functioning, and training procedures does not suffice to delineate a network's functioning at a level relevant for cognitive theorizing, and certainly does not constitute a theory of a human cognitive process. Thus far, however, no alternative framework or vocabulary has emerged for characterizing networks or formulating connectionist cognitive theories. It may turn out that connectionist networks can be aptly characterized in terms of computations carried out on symbols (although perhaps symbols representing information at a level below that of whole concepts such as *dog* or *tree;* see, e.g., Fodor & Pylyshyn, 1988). In this case connectionist modeling might lead to the development of detailed and explicit cognitive theories that were not, however, different in kind from traditional symbol-processing theories. Alternatively, connectionist work may eventuate in development of a new form of cognitive theory that does not employ symbols and rules as explanatory concepts. At present, however, this new form of theory is a possibility and not a reality.

The appeal of connectionism within cognitive science stems in part from a (not entirely unwarranted) dissatisfaction with traditional theoretical frameworks (see, e.g., Bechtel & Abrahamsen, 1991; McClelland, Rumelhart, & Hinton, 1986; Seidenberg, 1988; Smolensky, 1987, 1988). Hence, if connectionist modeling led either to more detailed, explicit, and widely applicable symbol-processing theories, or to a new form of theory that was more satisfactory for at least some purposes, the contribution of the connectionist approach would be important and positive. In any event, the coming years are certain to be interesting and productive for cognitive scientists, as we grapple with the fascinating and difficult questions raised by connectionism.

Acknowledgments—Preparation of this article was supported by NIH grant NS21047. I thank Bill Badecker, Alfonso Caramazza, Margrethe Lindemann, and Brenda Rapp for their helpful comments.

REFERENCES

Anderson, J.A., & Rosenfeld, E. (Eds.). (1988). *Neurocomputing*. Cambridge, MA: MIT Press.

Barnden, J.A., & Pollack, J.B. (Eds.). (1991). *Advances in connectionist and neural computation theory. Vol. 1: High-level connectionist models*. Norwood, NJ: Ablex.

Bechtel, W. (1987). Connectionism and the philosophy of mind: An overview. *The Southern Journal of Philosophy, 26* (Suppl.), 17–41.

Bechtel, W., & Abrahamsen, A. (1991). *Connectionism and the mind*. Cambridge, MA: Basil Blackwell.

Besner, D., Twilley, L., McCann, R.S., & Seergobin, K. (1990). On the association between connectionism and data: Are a few words necessary? *Psychological Review, 97,* 432–446.

Bever, T.G. (in press). The demons and the beast—Modular and nodular kinds of knowledge. In R. Ronan & N. Sharkey (Eds.), *Connectionist approaches to natural language processing*. Hillsdale, NJ: Erlbaum.

Clark, A. (1990). *Microcognition: Philosophy, cognitive science, and parallel distributed processing*. Cambridge, MA: MIT Press.

Chandrasekaran, B., Goel, A., & Allemang, D. (1988). Information processing abstractions: The message still counts more than the medium. *Behavioral and Brain Sciences, 11,* 26–27.

Cummins, R., & Schwarz, G. (1987). Radical connectionism. *The Southern Journal of Philosophy, 26* (Suppl.), 43–61.

Douglas, R.J., & Martin, K.A.C. (1991). Opening the grey box. *Trends in Neuroscience, 14,* 286–293.

Dyer, M.G. (1988). The promise and problems of connectionism. *Behavioral and Brain Sciences, 11,* 32–33.

Estes, W.K. (1988). Toward a framework for combining connectionist and symbol-processing models. *Journal of Memory and Language, 27,* 196–212.

Estes, W.K. (1991). Cognitive architectures from the standpoint of an experimental psychologist. *Annual Review of Psychology, 42,* 1–28.

Fodor, J.A., & Pylyshyn, Z.W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28,* 3–71.

Gluck, M.A., & Bower, G.H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117,* 227–247.

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

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science, 11,* 23–63.

Hanson, S.J., & Burr, D.J. (1990). What connectionist models learn: Learning and representation in connectionist networks. *Behavioral and Brain Sciences, 13,* 471–518.

Hinton, G.E., & Shallice, T. (in press). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*.

Humphreys, G.W. (1989). Introduction: Parallel distributed processing and psychology. In R.G.M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology* (pp. 65–75). Oxford: Clarendon Press.

Lamberts, K., & d'Ydewalle, G. (1990). What can psychologists learn from hidden-unit nets? *Behavioral and Brain Sciences, 13,* 499–500.

Massaro, D.W. (1988). Some criticisms of connectionist models of human performance. *Journal of Memory and Language, 27,* 213–234.

McClelland, J.L. (1988). Connectionist models and psychological evidence. *Journal of Memory and Language, 27,* 107–123.

McClelland, J.L., & Rumelhart, D.E. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 2: Psychological and biological models.* Cambridge, MA: MIT Press.

McClelland, J.L., Rumelhart, D.E., & Hinton, G.E. (1986). The appeal of parallel distributed processing. In D.E. Rumelhart & J.L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations* (pp. 3–44). Cambridge, MA: MIT Press.

McCloskey, M., & Cohen, N.J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G.H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory, 24* (pp. 109–165). San Diego: Academic Press.

McCloskey, M., & Lindemann, A.M. (in press). MATHNET: Preliminary results from a distributed model of arithmetic fact retrieval. In J.I.D. Campbell (Ed.), *The nature and origin of mathematical skills.* New York: Elsevier.

Minsky, M.L., & Papert, S.A. (1988). *Perceptrons: An introduction to computational geometry* (Expanded ed.). Cambridge, MA: MIT Press.

Morris, R.G.M. (Ed). (1989). *Parallel distributed processing: Implications for psychology and neurobiology.* Oxford: Clarendon Press.

Olson, A., & Caramazza, A. (1991). The role of cognitive theory in neuropsychological research. In S. Corkin, J. Grafman, & F. Boller (Eds.), *Handbook of neuropsychology* (pp. 287–309). Amsterdam: Elsevier.

Patterson, K., Seidenberg, M.S., & McClelland, J.L. (1989). Connections and disconnections: Acquired dyslexia in a computational model of reading processes. In R.G.M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology* (pp. 132–181). Oxford: Clarendon Press.

Pavel, M. (1990). Learning from learned networks. *Behavioral and Brain Sciences, 13,* 503–504.

Pinker, S., & Mehler, J. (Eds.). (1988). *Connections and symbols.* Cambridge, MA: MIT Press.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition, 28,* 73–193.

Rager, J.E. (1990). The analysis of learning needs to be deeper. *Behavioral and Brain Sciences, 13,* 505–506.

Rumelhart, D.E., & McClelland, J.L. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations.* Cambridge, MA: MIT Press.

Seidenberg, M.S. (1988). Cognitive neuropsychology and language: The state of the art. *Cognitive Neuropsychology, 5,* 403–426.

Seidenberg, M.S. (1989). Visual word recognition and pronunciation: A computational model and its implications. In W. Marslen-Wilson (Ed.), *Lexical representation and process* (pp. 25–74). Cambridge, MA: MIT Press.

Seidenberg, M.S., & McClelland, J.L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96,* 523–568.

Sejnowski, T.J., & Rosenberg, C.R. (1987). Parallel networks that learn to pronounce English text. *Complex Systems, 1,* 145–168.

Smolensky, P. (1987). The constituent structure of connectionist mental states: A reply to Fodor and Pylyshyn. *The Southern Journal of Philosophy, 26* (Suppl.), 137–161.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences, 11,* 1–74.

Suppes, P. (1990). Problems of extension, representation, and computational irreducibility. *Behavioral and Brain Sciences, 13,* 507–508.