psycoloquy.98.9.18.connectionist-explanation.15.goldsmith Fri May 8 1998
ISSN 1055-0143 (12 paragraphs, 33 references, 4 notes, 406 lines)
PSYCOLOQUY is sponsored by the American Psychological Association (APA)
Copyright 1998 Morris Goldsmith

CONNECTIONIST MODELING AND THEORIZING: WHO DOES THE EXPLAINING AND HOW?
Commentary on Green on Connectionist-Explanation

Morris Goldsmith
Department of Psychology
University of Haifa
Haifa, 31905, Israel
mgold@psy.haifa.ac.il

ABSTRACT: Green's (1998) criticism that connectionist models are devoid of theoretical substance rests on a simplistic view of the nature of connectionist models and a failure to acknowledge the division of labor between the model and the modeller in the enterprise of connectionist modelling. The "theoretical terms" of connectionist theory are not to be found in processing units or in connections but in more abstract characterizations of the functional properties of networks. Moreover, these properties are -- and at present should be -- only loosely tied to the known (and largely unknown) properties of neural networks in the brain.

1. Green (1998) attempts to find a mapping between connectionist models of cognitive phenomena and traditional scientific theories. Treating nodes and connections as the postulated theoretical terms, he finds these terms to be exceedingly numerous, far removed from the observable phenomena, and to a great extent optional. Moreover, he claims, the ways in which the units interact to produce the desired behavior are generally unfathomable. This leads him to ask "what, exactly, is learned about a cognitive domain modelled by a connectionist network?" His conclusion: not much, if anything at all.

2. Green does, however, hold out hope for current and would-be connectionist modellers: In his words, "at present the only way of interpreting connectionist networks as serious candidates for theories of cognition would be as literal models of the brain activity that underpins cognition." Hence, "connectionists should start restricting themselves to units, connections, and rules that use all and only principles that are known to be true of neurons." This recommendation seems to have found favor in some of the commentaries as well (e.g., O'Brien, 1998).

3. In this commentary I take issue with both the premises and the conclusion of Green's argument: (a) Nodes and connections are not the theoretical terms of connectionist models; rather, that role is filled

by more abstract characterizations of the functional properties of the networks. (b) More is known about these functional properties than Green acknowledges, but in any case, there is no principled limit on our ability to understand these properties. (c) Exploring the capacities and functional properties of artificial neural networks can be of value in the study of cognition regardless of whether and how those capacities and properties are realized in the brain.

4. When Green looks at a connectionist network, he sees "dozens, sometimes hundreds, of simple units, bound together by hundreds, sometimes thousands, of connections... Each of the units, connections, and rules in a connectionist network is a theoretical entity..., [yet] neither the units nor the connections correspond to anything in the way that variables and rules did in traditional computational models of cognition." So, Green asks, where is the BEEF? He seems to expect a transparent isomorphism between individual ELEMENTS of the model and the cognitive phenomena that are being modeled. He argues that such is the usual case in scientific theorizing and in traditional cognitive modelling.

5. There are indeed connectionist models -- localist models -- that are relatively transparent in the mapping between the elements of the model and the actual theoretical claims about the cognitive phenomena being modeled (Feldman & Ballard, 1982; and see Grainger & Jacobs, 1998). But this is not a feature of the parallel distributed processing models that are apparently the target of Green's attack. Are the latter models truly devoid of substantive theoretical content, as Green would have us believe? Should they be abandoned in favor of more transparent localist versions (Grainger & Jacobs, 1998), or at least grounded in substantive claims about BRAIN processing (see parag. 12, below)?

6. The place of connectionism in cognitive theory has been debated extensively since its renaissance began a little over a decade ago (e.g., Broadbent, 1985; Fodor & Pylyshyn, 1988, Massaro, 1988; McCloskey, 1991; Rumelhart & McClelland, 1986; Smolensky, 1988), and many of the points that Green raises have been argued before. Seidenberg (1993) has perhaps been most explicit in fending off the central thrust of Green's argument in the distinction he draws (borrowing from Chomsky, 1965) between "descriptive" and "explanatory" theorizing. The type of theory that Green is looking for would be "descriptive" in Seidenberg's terminology: "Experiments yield data around which descriptive theories are developed... providing systematic descriptions of phenomena and generating novel predictions" (p. 230). Explanatory theories, in contrast, "appeal to a small set of concepts that are independently motivated rather than task- or phenomenon-specific' (p. 230). In Seidenberg's brand of "explanatory connectionism," theories are derived from "general connectionist principles in conjunction with domain-specific boundary conditions" (p. 231).

7. Without getting embroiled in the "more explanatory than thou" aspects of Seidenberg's argument, let us simply note that for him, connectionist theories, like other theories, are embodied in concepts and principles rather than in units and connections. Moreover, the theoretical claims are formulated at various levels of generality and abstraction. Thus, in discussing Seidenberg and McClelland's (1989) model of word reading, Seidenberg (1993, p. 232) identified "broad theoretical claims," such as those concerning the representational status of words (i.e., no explicit lexical representation) and the postulation of a single-process mechanism (as opposed to the traditional dual-process account) to handle rule-governed words, irregular words and nonwords. He also pointed to more specific claims concerning the factors that influence the generation of pronunciations from print (e.g., the importance of sublexical units, such as word bodies), and a "novel link" between frequency effects and the effects of spelling-sound consistency. Some of these claims, those directly tied to general principles, were specified in advance of the modelling, whereas others fell out of the modelling process itself.

8. The division of labor between the model and the modeller in carving out the cognitive theory deserves some amplification. Indeed, this aspect seems to have been completely overlooked in Green's analysis. Green's basic premise, that the theory is (or is not) to be found IN THE MODEL, is misguided. Theories are put forward by scientists, not by models. Simulation models are powerful tools that help researchers develop, test, present and demonstrate the plausibility of their theoretical ideas. They do not, however, "discover" the theory for the researcher, nor do they embody it. Clearly, connectionist modelling is a complex task. Because the principles of computation in connectionist (parallel distributed processing) computational architectures are not yet well understood, a large part of the discovery process comes from working with the models themselves -- trying out various architectures, input representations, learning rules and parameters, and so forth. The end product of the synergistic interaction between modeller and model, however, is not just the model (or models), but a scientific article, in which the researcher's theoretical ideas, and their justification, are articulated (viz., as an interrelated set of linguistic propositions).

9. Often in such articles, the models are treated as experimental "subjects," whose essential computational properties are inferred by the modeller. A good example is Plaut and Shallice (1993), who systematically investigated various "design issues" concerning an earlier simulation model (Hinton & Shallice, 1991) used to explain the co-occurrence of visual and semantic errors found in deep dyslexia. As summarized by Plaut (1995),

> "The design issues included the definition of the task of reading

> via meaning, the network architecture (i.e., the numbers of units, their organization into layers, and how these groups are connected), the training procedure, used for adjusting connection strengths, and the procedure for evaluating the behavior of the trained network in its normal state and after damage. The major finding was that the occurrence of the qualitative error pattern was surprisingly insensitive to these detailed aspects of the simulation. Rather, what appeared critical was a more general property that all of the implementations shared: that units learned to interact in such a way that familiar patterns of activity over semantic features -- corresponding to word meanings -- formed STABLE ATTRACTORS in the space of all possible semantic representations" (Plaut, 1995, pp. 299-300, emphasis in original).

10. Plaut and many other connectionists propose that concepts such as stable attractors, basins of attraction, clean-up processes, collateral support, superposition, gradient descent, trajectories in weight space, trajectories in state space, and so forth, offer new and valuable ways of understanding how networks -- and arguably how people -- perform cognitive tasks. Whether or not they are right (cf. Koriat & Goldsmith, 1996b), it should be clear that the building blocks of connectionist explanations of cognitive performance are not "units" or "weights," but higher order descriptions of the functional properties and dynamics (interactions) of networks, during learning, during processing, or both.

11. There are various general techniques for probing the inner workings of connectionist networks (see, for example, Berkeley et al., 1995; Elman, 1990a, 1990b; Hanson & Burr, 1990; Hinton, 1989; Hinton, McClelland and Rumelhart, 1986; Hinton & Shallice, 1991; Meddler & Dawson, 1998; Sejnowski & Rosenberg, 1988; Rumelhart & Todd, 1993), as well as many clever methods that have been tailored for specific models in particular studies. Some of the techniques are useful in attempting to map approximate, "macro-level" symbolic interpretations onto the patterns of weights and unit activations, whereas others are designed to characterize the functioning of the network in more uniquely connectionist terms (e.g., Plaut & Shallice, 1993; Plaut, McClelland, Seidenberg, & Patterson, 1996;). Either way, the goal is not to understand, say, the internal representations of the model per se, but to determine how the basic nature of those representations contributes to the psychologically relevant behavior of the model [1]. Admittedly, there is much more that can and should be done in this direction. However, the picture is far different from Green's caricatured portrayal (see parags. 15-17, 20), in which a working connectionist model is placed on the table for all to see, but no one, including the modeller, has an inkling of how it works.

12. Finally, concerning the "brain" issue: I have argued that

connectionist theories (like other cognitive theories) are propositional assertions about how certain computational tasks might be achieved, based, among other things, on the lessons learned about the functional properties of connectionist networks as they are implemented in particular cognitive domains. These theories may be evaluated in all of the usual ways -- in terms of their ability to account for existing data, and in terms of their ability to spawn novel predictions that are subsequently confirmed [2]. Hence, there is no need to ground the theories in explicit assertions about neural activity in the brain, as Green suggests (parag. 20). Moreover, it would be improper to do so. Like more traditional computational accounts, connectionist explanations are directed to the "algorithmic" (software) level of explanation in Marr's (1982) terms (see Rumelhart & McClelland, 1986; Smolensky, 1988), rather than to the implementational (hardware or "wetware") level [3]. Although connectionist computational accounts are in a unique position to be INFORMED by knowledge about the workings of the brain, they are no more committed to incorporating that knowledge than are traditional symbolic accounts (which are informed primarily by knowledge about the workings of digital computers) [4].

ENDNOTES

[1] Plaut and Shallice (1993), for example, identified four properties of their network that underlie its ability to reproduce the deep dyslexic symptom-complex: distributed orthographic and semantic representations, gradient descent learning, attractors for word meanings, and greater richness of concrete versus abstract semantics.

[2] McClelland (1988, p. 116) points out several ways in which connectionist models have already proven to be of value: (1) They have led to new interpretations of basic phenomena in the literature. (2) They have provided unified accounts of what had previously been seen as highly disparate or even contradictory phenomena. (3) They have clarified the relevance of certain kinds of evidence for adjudicating basic questions about the character of the information-processing system. (It would be a simple matter to compose a list of recent references to support each point.) In addition, he emphasizes that connectionism offers a new way of thinking about cognitive phenomena -- an alternative set of conceptual and computational tools that should not be judged in terms of being "right" or "wrong," but rather, in terms of its usefulness for capturing those aspects of cognition that are not handled well by other frameworks (cf. Koriat & Goldsmith, 1996a, 1996b).

[3] Connectionists sometimes claim to have "dissolved" the software-hardware distinction (e.g., Bates & Elman, 1993, p. 635). This is an unfortunate claim that is not supported by the sources just cited (see also, McClelland, 1988; Rumelhart, 1989). Although the "software" and "hardware" aspects of connectionist models are indeed intrinsically intertwined (unlike in

conventional digital computers), they nevertheless remain two distinct levels of description and explanation, only one of which is directly relevant to cognitive theorizing. In addition, inspection of the literature reveals that an important part of connectionist theorizing involves recasting cognitive phenomena at what Marr (1982) considered to be the highest, "computational" level, the level that specifies what computational problems need to be solved, and why (see, for example, McClelland, McNaughton, & O'Reilly, 1995; Plaut et al., 1996).

[4] Undoubtedly, part of the reason that connectionism is often judged unfavorably against a relatively strict standard of neural plausibility stems from the great emphasis that connectionists place on the positive aspects of the brain metaphor (Rumelhart, 1989). Perhaps connectionists are to blame for trying to "have their cake" and eat it too. Although not very fashionable, my view on the issue is that "brain style" computation would still be worth pursuing for its theoretical and computational advantages even if the basic computational architecture were completely artificial, credited to a prominent computer scientist, and referred to as the "Von Brain machine."

REFERENCES

Bates, E. A., & Elman, J. L. (1993). Connectionism and the study of change. In M. H. Johnson (Ed.), Brain development and cognition: A reader (pp. 623-642). Oxford: Blackwell

Berkeley, I.S.N., Dawson, M. R. W., Medler, D.A., Schopflocher, D. P., & Hornsby, L. (1995). Density plots of hidden unit activations reveal interpretable bands. Connection Science, 7, 167-186.

Broadbent, D. (1985). A question of levels: Comment on McClelland and Rumelhart. Journal of Experimental Psychology: General, 114, 198-192.

Chomsky, N. (1965). Aspects of the theory of syntax. Cambridge, MA: MIT Press.

Elman, J. L. (1990a). Finding structure in time. Cognitive Science, 14, 179-211.

Elman, J. L. (1990b). Representation and structure in connectionist models. In G. T. M. Altmann (Ed.), Cognitive models of speech processing: Psycholinguistic and computational perspectives (pp. 345-382). Cambridge, MA: MIT Press.

Feldman, J. A., & Ballard, D. H. (1982). Connectionist models and their properties. Cognitive Science, 6, 205-254.

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

Grainger, J., & Jacobs, A. M. (1998). Localist connectionism fits the bill. Psycoloquy 9(10) ftp://ftp.princeton.edu/pub/harnad/Psycoloquy/1998.volume.9/ psyc.98.9.10.connectionist-explanation.7.grainger.

Green, C. D. (1998). Are connectionist models theories of cognition? Psycoloquy 9(4) ftp://ftp.princeton.edu/pub/harnad/Psycoloquy/1998.volume.9/ psyc.98.9.04.connectionist-explanation.1.green.

Hanson, S.J., & Burr, D.J. (1990). What connectionist models learn: Learning and representation in connectionist networks. Behavioral and Brain Sciences, 13, 511-518.

Hinton, G. E. (1989). Learning distributed representations of concepts. In R. G. M. Morris (Ed.), Parallel distributed processing: Implications for psychology and neurobiology (pp. 8-45). Oxford: Oxford University Press.

Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. Psychological Review; 98, 74-95.

Hinton, G. E., McClelland, J. L., and Rumelhart, D. E. (1986) Distributed representations. In D. E. Rumelhart, J. L. McClelland & the PDP research group (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition, Vol 1. Cambridge, MA: MIT Press.

Koriat, A., & Goldsmith, M. (1996a). Memory metaphors and the real-life/laboratory controversy: Correspondence versus storehouse conceptions of memory. Behavioral and Brain Sciences, 19, 167-188.

Koriat, A. & Goldsmith, M. (1996b). The correspondence metaphor of memory: Right, wrong, or useful? Behavioral and Brain Sciences, 19, 211-228.

Marr, D. (1982). Vision. San Francisco: W. H. Freeman.

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., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. Psychological Review, 102, 419-437.

McCloskey, M. (1991). Networks and theories: The place of connectionism in cognitive science. Psychological Science, 2, 387-395.

Meddler, D. A., & Dawson, M. R. W. (1998). Connectionism and cognitive theories. Psycoloquy 9(11) ftp://ftp.princeton.edu/pub/harnad/Psycoloquy/1998.volume.9/ psyc.98.9.11.connectionist-explanation.8.medler.

O'Brien, G. J. (1998). The role of implementation in connectionist explanation. Psycoloquy 9(6) ftp://ftp.princeton.edu/pub/harnad/Psycoloquy/1998.volume.9/ psyc.98.9.06.connectionist-explanation.3.obrien.

Plaut, D. C. (1995). Double dissociation without modularity: Evidence from connectionist neuropsychology. Journal of Clinical and Experimental Neuropsychology, 17, 291-321.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. Psychological Review, 103, 56-115.

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. Cognitive Neuropsychology; 1993, 10, 377-500.

Rumelhart, D. E (1989). The architecture of mind: A connectionist approach. In M. I. Posner (Ed.) Foundations of cognitive science (pp. 133-159). Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1986). PDP models and general issues in cognitive science. In D. E. Rumelhart, J. L. McClelland & the PDP research group (Eds.), Parallel distributed processing: Explorations in the microstructure of cognition, Vol 1 (pp. 110-146). Cambridge, MA: MIT Press.

Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. In D. E. Meyer & S. Kornblum (Eds.), Attention and performance 14: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience (pp. 3-30). Cambridge, MA: MIT Press.

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

Seidenberg, M. (1993). Connectionist models and cognitive science. Psychological Science, 4, 228-235.

Sejnowski, T. J., & Rosenberg, C. R. (1988). Learning and representation in connectionist models. In M. S. Gazzaniga (Ed.), Perspectives in memory research (pp. 135-178). Cambridge, MA: MIT Press.

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