Comment on Page, M. (2000) Connectionist modelling in psychology: A localist manifesto
To appear in *Behavioral and Brain Sciences*, *23*(3).

Stipulating Versus Discovering Representations

David C. Plaut and James L. McClelland
Departments of Psychology and Computer Science
and the Center for the Neural Basis of Cognition
Carnegie Mellon University

Page's proposal to stipulate representations in which individual units correspond to meaningful entities is too unconstrained to support effective theorizing. An approach combining general computational principles with domain-specific assumptions, in which learning is used to discover representations that are effective in solving tasks, provides more insight into why cognitive and neural systems are organized the way they are.

Page sets up a fundamental contrast between localist versus distributed approaches to connectionist modeling. To us there appear to be several dimensions to the actual contrast he has in mind. Perhaps the most fundamental distinction is whether it is stipulated in advance that representational units be assigned to "meaningful entities" or whether, as we believe, it is better to discover useful representations in response to task constraints. We agree with Page that localist connectionist models have made important contributions to our understanding of many different cognitive phenomena. However, we think the choice of representation used in the brain reflects the operation of a set of general principles in conjunction with domain characteristics. It is a program of scientific research to discover what the principles and domain characteristics are and how they give rise to different types of representations. As a starting place in the discovery of the relevant principles, we have suggested (McClelland, 1993; Plaut, McClelland, Seidenberg, & Patterson, 1996) that the principles include the following: that the activations and connection weights that support representation and processing are graded in nature; that processing is intrinsically gradual, stochastic, and interactive; and that mechanisms underlying processing adapt to task constraints.

Constraint Versus Flexibility. Page's suggestion that we stipulate the use of representations in which the units correspond to meaningful entities would appear on the face of it to be constraining, but in practice it appears to confer too much flexibility. Indeed, throughout his target article, Page applauds the power and flexibility of localist modeling, often contrasting it with models in which representations are discovered in response to task constraints. A particularly telling example is his treatment of age-of-acquisition effects (which he considers to be "potentially difficult to model in connectionist terms", draft p. 30). Page describes a localist system, incorporating three new assumptions, that would be expected to exhibit such effects. However, it would have been even easier for Page to formulate a model that would not exhibit such effects—a localist model without the additional assumptions might suffice. A critical role of theory is to account not only for what does occur but also for what *doesn't* (see Roberts & Pashler, in press); the localist modeling framework provides no leverage in this respect. In contrast, the distributed connectionist model, which is more constrained in this regard, is potentially falsifiable by evidence of the presence or absence of age-of-acquisition effects. In fact, Page has it exactly backwards about the relationship between such effects and connectionist models that discover representations via back-propagation. Ellis and Lambon-Ralph (personal communication) have pointed out that age of acquisition effects are actually intrinsic to such models, and their characteristics provide one potential explanation for these effects.

Page is exactly right to point out that "it sometimes proves difficult to manipulate distributed representations in the same way as one can manipulate localist representations" (draft p. 50). In other words, the learning procedure discovers the representations subject to the principles governing the operation of the network and the task constraints, and the modeler is not free to manipulate them independently. Far from being problematic, however, we consider this characteristic of distributed systems to be critical to their usefulness in providing insight into cognition and behavior. By examining the adequacy of a system that applies a putative set of principles to a model that addresses performance of a particular task, we can evaluate when the principles are sufficient. When they fail, we gain the opportunity to explore how they may need to be adjusted or extended.

These considerations are relevant to Page's analysis of the complementary learning system hypothesis of McClelland, McNaughton, and O'Reilly (1995). These

We thank the CMU PDP research group for helpful comments and discussion. Correspondence regarding this article may be sent either to David Plaut (plaut@cmu.edu) or to James McClelland (jlm@cnbc.cmu.edu), Mellon Institute 115–CNBC, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh PA 15213–2683.

authors made the observation that connectionist networks trained with back-propagation or other structure-sensitive learning procedures a) discover useful representations through gradual, interleaved learning and b) exhibit catastrophic interference when trained sequentially (McCloskey & Cohen, 1989). Based on these observations, and on the fact that the gradual discovery of useful representations leads to a progressive differentiation of conceptual knowledge characteristic of human cognitive development, McClelland et al. (1995) suggested that the neocortex embodies the indicated characteristics of these learning procedures. One implication of this would be that rapid acquisition of arbitrary new information would *necessarily* be problematic for such a system, and that a solution to this problem would be provided if the brain also exploited a second, complementary approach to learning, employing sparse, conjunctive representations, that could acquire new arbitrary information quickly. The argument was that the strengths and limitations of structure-sensitive learning explained *why* there are two complementary learning systems in hippocampus and neocortex.

In contrast, Page goes to some length to illustrate how a localist approach to learning could completely avoid the problem of catastrophic interference that arises in connectionist networks trained with back-propagation. Indeed, on his approach, the hippocampal system is redundant with the neocortex as there is no need for cortical learning to be slow. Thus, within the localist framework, the existence of complementary learning systems in the hippocampus and neocortex is completely unnecessary, and hence the existence of such a division of labor in the brain is left unexplained.

Learning Representations Versus Building Them By Hand. A common approach in the early days of connectionist modeling was to wire up a network by hand, and under these circumstances there seemed to be a strong tendency among researchers to specify individual units that correspond to meaningful entities (see, e.g., Dell, 1986; McClelland & Rumelhart, 1981). However, learning is a central aspect of many cognitive phenomena, so it is essential that a modeling framework provide a natural means for acquiring and updating knowledge. Once one turns to the possibility that the knowledge embodied in a connectionist network might be learned (or even discovered by natural selection), one immediately has the chance to revisit the question of whether the individual units in a network should be expected to correspond to meaningful entities. It is not obvious that correspondence to meaningful entities per se (or the convenience of this correspondence for modelers) confers any adaptive advantage.

To his credit, Page acknowledges the central role that learning must play in cognitive modeling, and presents a modified version of the ART/competitive learning framework (Grossberg, 1976; Carpenter & Grossberg, 1987; Rumelhart & Zipser, 1985) as a proposal for learning localist representations. However, there are a number of difficulties with this proposal, all of which point to reasons for continuing to pursue other alternatives. We consider three such difficulties here.

1. On close examination, most of the positive aspects of the proposal derive from properties of the assumed distributed representations that are input to the localist learning mechanism. For example, Page points out that localist models permit graded, similarity-based activation. It is crucial to note, however, that the pattern of similarity-based activation that results depends entirely on the similarity structure of the representations providing input to the localist units. Unfortunately, nowhere in Page's article does he indicate how his localist approach could solve the problem of discovering such representations.

In contrast, a key reason for the popularity of back-propagation is that it is effective at discovering internal, distributed representations that capture the underlying structure of a domain. For example, Hinton (1986) showed that a network could discover kinship relationships within two analogous families, even in the absence of similarity structure in the input representations for individuals. Although some runs of the network produce internal representations with "meaningful" units (e.g., nationality, generation, gender, branch-of-family), the more general situation is one in which the meaningful features of the domain are captured by the *principal components* of the learned representations (McClelland, 1994; see also Anderson, Silverstein, Ritz, & Jones, 1977; Elman, 1991).

2. Page notes that localist models are capable of considerable generalization. This again arises from the similarity-based activation due to the distributed patterns of activation that are input to the localist units. We suggest that one reason localist-style models (e.g., the generalized context model, Nosofsky, 1986, or ALCOVE, Kruschke, 1992) have proven successful in modeling learning in experimental studies is because they apply to learning that occurs within the brief time frame of most psychology experiments (1 hour up to at most about 20 hours spread over a couple of weeks). Within this restricted time frame, we expect relatively little change in the relevant dimensions of the representation, so the generalization abilities of models that learn by adapting only the relative salience of existing dimensions may be sufficient.

What seems more challenging for such approaches is to address changes in the underlying representational dimensions themselves. Such shifts can occur in our task-driven approach through the progressive, incremental process by which learning assigns representations in response to exposure to examples embodying domain knowledge (McClelland, 1994). On our view, the establishment of appropriate representations is a developmental process that takes place over extended periods of time (months or years), allowing models that develop such representations to account for developmental changes such as progressive differentiation of conceptual knowledge (Keil, 1979) and developmental shifts in the basis of categorization of living things from superficial to a metabolic/reproductive basis (Johnson & Carey, 1998).

3. Within the above limitations, Page's proposed localist learning procedure sounds like it might work on paper, but it is telling that he discusses in detail how the learning process might proceed only for the case in which every presentation of an item in a psychological experiment results in a separate localist representation. This form of localism—the idea that every experience is assigned a separate unit in the representation—seems highly implausible to us. It is difficult to imagine a separate unit for every encounter with every object or written or spoken word every moment of every day. Such instance-based approaches have led to some interesting accounts of psychological data (by Logan and others, as Page reviews), but in our view it is best to treat this form of localist modeling as an interesting and useful abstraction of an underlyingly distributed, superpositional form of representation. More specifically, we agree there is a trace layed down in the brain resulting from each experience and that localist models can approximate how these traces influence processing. We believe, however, that the traces are actually the adjustments to the connections in a distributed connectionist system rather than stored instances. McClelland and Rumelhart (1985), for example, showed how a simple superpositional system can capture several patterns of data previously taken as supporting instance-based theories, and Cohen, Dunbar, and McClelland (1990) demonstrated that distributed connectionist models trained with back-propagation can capture the power law of practice just as Logan's instance models do.

It seems somewhat more plausible to us that multiple occurrences of a meaningful cognitive entity such as a letter or word might be mapped onto the same unit. However, the ability of models that exploit the kind of procedure Page proposes to actually produce such representations is unclear. In our experience, to obtain satisfactory results with such models it is necessary to tune the "vigilance" parameter very carefully, and often in ways that depend strongly on specifics of the training set. But there is a deeper problem. Whenever there is any tolerance of variation among instances of a particular item, one immediately runs into the fact that the modeler is forced to decide just what the acceptable level of mismatch should be. If, for example, a reader encounters a misspelling of the word ANTARTICA *[Note to typesetter: leave this word misspelled without inserting "sic"]*, should we necessarily imagine that the cognitive system must create a separate unit for it? Or if, in a certain Wendy's restaurant, the salad bar is not immediately opposite the ordering queue, should we create a new subcategory of the Wendy's subcategory of restaurants? Within a task-driven learning approach, in which robust patterns of covariation become representationally coherent, and in which subpatterns coexist within the larger patterns of covariation, such otherwise thorny issues become irrelevant (McClelland & Rumelhart, 1985; Rumelhart, Smolensky, McClelland, & Hinton, 1986).

Task-Driven Learning Can Discover Localist-Like Representations. As we have noted, whereas Page would stipulate localist representations for various types of problems, our approach allows an appropriate representation to be created in response to the constraints built into the learning procedure and the task at hand. At a general level, distributed representations seem most useful in systematic domains in which similar inputs map to similar outputs (e.g., English word reading), whereas localist representations (and here we mean specifically representations involving one unit per entire input pattern) are most useful in unsystematic domains in which similar inputs may map to completely unrelated outputs (e.g., word comprehension, face naming, episodic memory). It is thus interesting (although, to our knowledge, not particularly well documented) that standard connectionist learning procedures tend to produce dense, overlapping internal representations when applied to systematic tasks, whereas they tend to produce much sparser, less overlapping representations when applied to unsystematic tasks. Although Page considers the latter to be functionally equivalent to localist representations, there are at least two reasons to reject this equivalence. First, sparse distributed representations scale far better than strictly localist ones (Marr, 1970; McClelland & Goddard, 1996; Kanerva, 1988). Second, and perhaps more important, sparse distributed representations are on one end of a continuum produced by the same set of computational assumptions that yield more dense, overlapping representations when these are useful to capture the structure in a domain.

Other Comments on Page's Critique of "Distributed" Approaches. In rejecting what he calls the distributed approach, Page levels several criticisms that are either incorrect or overstated, partly because he seems to adopt an overly narrow view of the approach. For one thing, Page appears to equate the distributed approach with the application of back-propagation within feed-forward networks. He then raises questions about the biological plausibility of back-propagation but fails to acknowledge that there are a number of other, more plausible procedures for performing gradient descent learning in distributed systems which are functionally equivalent to back-propagation (see, e.g., O'Reilly, 1996). Page questions whether distributed systems can appropriately fail to generalize in unsystematic domains (e.g., mapping orthography to semantics for pseudowords; draft p. 26) when such behavior has already been demonstrated (Plaut, 1997; Plaut & Shallice, 1993). He also questions how a distributed system can decide when and how to respond without some sort of homuncular energy-monitoring system (although see Botvinick, Nystrom, Fissell, Carter, & Cohen, in press, for recent functional imaging data supporting the hypothesis that the anterior cingulate may, in fact, play such a role). In fact, no such explicit decisions are required; all that is needed is that the motor system be sufficiently damped that it initiates behavior only when driven by strongly activated, stable internal representations (see Kello, Plaut, & MacWhinney, in press, for a simple demonstration of this idea).

Based on the above, we suggest that the representations used by the brain in solving a particular task are not something we should stipulate in advance. Rather, they are selected by evolution and by learning as solutions to challenges and opportunities posed by the environment. The structure of the problem will determine whether the representation will be localist-like or more distributed in character.

References

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, *84*, 413-451.

Botvinick, M., Nystrom, L., Fissell, K., Carter, C. S., & Cohen, J. D. (in press). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature*.

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, *37*, 54-115.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review*, *97*(3), 332-361.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283-321.

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

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: Part I: Parallel development and coding of neural feature detectors. *Biological Cybernetics*, *23*, 121-134.

Hinton, G. E. (1986). Learning distributed representations of concepts. In *Proceedings of the 8th Annual Conference of the Cognitive Science Society* (p. 1-12). Hillsdale, NJ: Erlbaum.

Johnson, S. J., & Carey, S. (1998). Knowledge, enrichment and conceptual change in folkbiology: Evidence from Williams Syndrome. *Cognitive Psychology*, *38*, 156-200.

Kanerva, P. (1988). *Sparse distributed memory.* Cambridge, MA: MIT Press.

Keil, F. C. (1979). *Semantic and conceptual development: An ontological perspective.* Cambridge, MA: Harvard University Press.

Kello, C. T., Plaut, D. C., & MacWhinney, B. (in press). The task-dependence of staged versus cascaded processing: An empirical and computational study of Stroop interference on speech production. *Journal of Experimental Psychology: General*.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22-44.

Marr, D. (1970). A theory for cerebral neocortex. *Proceedings of the Royal Society of London, Series B*, *176*, 161-234.

McClelland, J. L. (1993). The GRAIN model: A framework for modeling the dynamics of information processing. In D. E. Meyer & S. Kornblum (Eds.), *Attention and Performance XIV: Synergies in experimental psychology, artifical intelligence, and cognitive neuroscience* (p. 655-688). Hillsdale, NJ: Erlbaum.

McClelland, J. L. (1994). The interaction of nature and nurture in development: A parallel distributed processing perspective. In P. Bertelson, P. Eelen, & G. d'Ydewalle (Eds.), *International perspectives on psychological science, Volume 1: Leading themes* (p. 57-88). Hillsdale, NJ: Erlbaum.

McClelland, J. L., & Goddard, N. H. (1996). Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. *Hippocampus*, *6*, 654-665.

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*, 419-457.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, *88*(5), 375-407.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, *114*(2), 159-188.

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* (Vol. 24, p. 109-165). New York: Academic Press.

Nosofsky, R. M. (1986). Attention, similiarity and the identification-categorization relationship. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *115*(1), 39-57.

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation*, *8*(5), 895-938.

Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of naming and lexical decision. *Language and Cognitive Processes*, *12*, 767-808.

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*, *10*(5), 377-500.

Roberts, S., & Pashler, H. (in press). How persuasive is a good fit? A comment on theory testing in psychology. *Psychological Review*.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In J. L. McClelland, D. E. Rumelhart, &

the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models* (p. 7-57). Cambridge, MA: MIT Press.

Rumelhart, D. E., & Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science*, *9*, 75-112.