Psychological Review
2010, Vol. 117, No. 1, 284–290

0033-295X/10/$12.00 DOI: 10.1037/a0017101

Locating Object Knowledge in the Brain: Comment on Bowers's (2009) Attempt to Revive the Grandmother Cell Hypothesis

David C. Plaut
Carnegie Mellon University
and The Center for the Neural Basis of Cognition

James L. McClelland
Stanford University

According to Bowers (2009), the finding that there are neurons with highly selective responses to familiar stimuli supports theories positing localist representations over approaches positing the type of distributed representations typically found in parallel distributed processing (PDP) models. However, his conclusions derive from an overly narrow view of the range of possible distributed representations and of the role that PDP models can play in exploring their properties. Although it is true that current distributed theories face challenges in accounting for both neural and behavioral data, the proposed localist account—to the extent that it is articulated at all—runs into more fundamental difficulties. Central to these difficulties is the problem of specifying the set of entities a localist unit represents.

Keywords: localist representations, distributed representations, grandmother cells, parallel distributed processing, connectionist modeling

For many years, neuroscientists and psychologists have considered how best to think about how entities such as words, faces, objects, and concepts are represented in the brain. In their chapter "Distributed Representations," Hinton, McClelland, and Rumelhart (1986) distinguished two broad alternatives.

> Given a network of simple computing elements and some entities to be represented, the most straightforward scheme is to use one computing element for each entity. This is called a *local* representation. . . . This chapter describes one type of representation that is less familiar and harder to think about than local representations. Each entity is represented by a pattern of activity distributed over many computing elements, and each computing element is involved in representing many different entities. (p. 77)

In other words, in a localist representation, the relationship of entities to units (e.g., neurons) is one-to-one, whereas in a distributed representation, it is many-to-many. On the basis of a broad range of arguments and evidence, many neuroscientists and psychologists since Barlow (1972) have rejected the notion that individual neurons would correspond to entities as complicated as one's grandmother and instead accept that the brain uses some form of distributed representation (see Gross, 2002, for a historical consideration of the grandmother cell hypothesis).

In a provocative recent article, Bowers (2009) attempted to turn these views on their head, presenting a number of arguments and findings that he believes establishes the biological plausibility of localist representations and calls into question the presumed support for distributed representations. For proper evaluation these claims, however, it is necessary to clarify what exactly counts as evidence for different types of representation, and certain aspects of Bowers's definitions and terminology are problematic in this regard. Moreover, although current distributed theories certainly face challenges in accounting for both neural and behavioral data, the proposed localist account—to the extent that it is articulated at all—runs into fundamental difficulties. Perhaps the most difficult challenge is the one we consider last: the delineation of what counts as an entity to which a localist representation would be assigned.

What Is a Localist Representation?

The interactive activation (IA) model of letter and word perception (McClelland & Rumelhart, 1981) may provide a useful context for clarifying the nature of localist and distributed representations and Bowers's claims about them. The model consists of three layers of interacting units: letter feature units at the bottom (various strokes at each of four positions), letter units in the middle (one per letter at each position; e.g., *t*, *i*, *m*, and *e*), and word units at the top (one per word; e.g., *time*). The IA model is usually thought of as a localist model because it contains single units that stand in one-to-one correspondence with words, but the current context demands more careful terminology. As the earlier quote from Hinton et al. (1986) makes clear, a representation is localist or distributed only relative to a specific set of entities. Thus, the word level of the IA model is localist relative to words, and the letter level is localist relative to (position-specific) letters. However, at the letter level, the presentation of a word results in the activation of multiple units (corresponding to its letters), and each of these units is activated by multiple words (i.e., words containing that letter in that position). Thus, according to the standard definitions, the letter level in the IA model is localist relative to letters but distributed relative to words.

David C. Plaut, Department of Psychology, Carnegie Mellon University, and The Center for Neural Basis of Cognition, Pittsburgh, Pennsylvania; James L. McClelland, Department of Psychology, Stanford University.

Correspondence concerning this article should be addressed to David C. Plaut, Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213-3890. E-mail: plaut@cmu.edu

As it turns out, however, it can be difficult to distinguish localist representations from distributed representations on the basis of activity because localist units often become active not only for the entity to which they correspond but also for entities that are similar to it. Bowers (2009) illustrated this property in the IA model, in which the input for *blur* activates its word unit strongly but also partially activates the word unit for *blue* (Figure 4, p. 226). Indeed, Hinton et al. (1986) emphasized that this off-item activation can be difficult to distinguish from the patterns that comprise distributed representations.

Moreover, in most localist theories it is assumed that there are multiple redundant copies of each dedicated unit. Thus, at a very general level, in both localist and distributed representations, multiple units become active in processing a given entity, and each unit will become at least partially active for multiple entities. This raises the question of the basis on which one might interpret neural activity as being consistent or inconsistent with one or the other of these possibilities.

One problem with evaluating Bowers's (2009) claims about representation is that his use of this term conflates two distinct aspects of a neural system. According to Bowers (2009), "the critical question is not whether a given neuron responds to more than one object, person, or word but rather whether the neuron codes for more than one thing" (p. 225). Presumably "responds to" refers to whether a neuron becomes active when certain stimuli are presented, but Bowers never defined what he means by "codes for" (nor "one thing"—more on that later). A plausible interpretation, however, is that he is referring to the knowledge that the system has about a particular entity. That is, one can distinguish whether knowledge about an entity is encoded in the connections coming into or out of a particular unit or whether it is distributed across the connections of many units. This distinction is important for understanding the operation of the system, but it is different from the question of the type of neural activity that is evoked by a given stimulus, which is the issue that frames the standard definitions of localist representations versus distributed representations (cited earlier).

For example, Bowers (2009) stated that "a key claim of [the PDP] approach is that . . . knowledge is coded as a pattern of activation across many processing units, with each unit contributing to many different representations" (p. 220). Actually, on the parallel distributed processing (PDP) approach, knowledge is encoded not in patterns of activity but in patterns of weighted connections between units. A pattern of activation (often over multiple participating brain areas) corresponds to the internal representation or interpretation of a given input, but the knowledge in the system that determines what activations will occur is to be found in the strengths of the connection weights. Distributed activity can be caused by either localist or distributed knowledge representation.

By interpreting Bowers's (2009) claims to be about locality of knowledge rather than activity, we can make sense of otherwise problematic assertions. For instance, he denied that words have distributed representations at the letter level in the IA model, stating that if this were the case, "the pattern of activation across a set of letters at layer $n-1$ should support the same (or at least similar) functions as the corresponding localist representations [of words] at layer n" (p. 223). In terms of activation, this claim is clearly false—there is no reason why the pattern of activity produced by a given stimulus at every level of the system should support the same functions.

For example, the retina is certainly necessary for recognizing a viewed object and contains all the relevant information, but it cannot support object recognition alone; rather, the information must be rerepresented by a hierarchy of visual areas before it can effectively engage object knowledge. On the other hand, viewed in terms of knowledge rather than activity, Bowers's (2009) claim makes perfect sense. Within the IA model, the lexical knowledge that the letter string *time* is a word is coded only in the connections between the corresponding word unit and its letters; remove that single unit, and *time* is no longer a word to the model.

One implication of recognizing that Bowers's (2009) claims about localist and distributed representations are actually about the degree of locality of knowledge rather than activity is that his terminology is inconsistent with most other researchers, who interpret claims about representation to refer to patterns of activity (see Hinton et al., 1986; Page, 2000). But perhaps more important, it forces a reconsideration of what type of data on neural activity would provide evidence for the locality of knowledge of words, objects, and faces.

All of the evidence that Bowers (2009) took to support localist over distributed knowledge consists of observations in which individual neurons show various types of highly selective, interpretable responses. Given that Bowers (2009) stated quite clearly that the grandmother cell hypothesis concerns the visual recognition of faces, words, and objects, many of the findings he cited (e.g., responses in simple organisms, sensory thresholds, cells in human hippocampus that respond strongly to a single individual among those tested), although intriguing in their own right, do not directly bear on the hypothesis. Regarding face selectivity, a typical reported finding is that of Young and Yamane (1992), who found one temporal-lobe neuron among 850 that responded strongly to one face and weakly to another out of 27 faces. But without a thorough exploration of the response of the cell to systematic variations of the selective face, it is difficult to know whether the cell is responding to the entire face or to some aspect of it that is distinctive within this set but shared by other faces. Moreover, the fact that only one cell out of 850 showed such a selective response raises the question of what the other cells are coding for. The natural reply on a localist account would be that they code for faces other than those that were presented during training. The problem is that the same is true of the apparently selective cell—it might also have responded to other faces if they had been presented. In fact, it is not possible to establish definitively that a neuron responds to "one thing" without testing it on all possible things; the best that can be done is to estimate a degree of sparsity in the neural response within the sampled subset of stimuli. It is interesting to note that Quian Quiroga, Kreiman, Koch, and Fried (2008) have done just this in their analysis of response properties of single neurons in human hippocampus. On the basis of the pattern of response that they saw, they estimated that each familiar pattern may activate about two out of every 1,000 neurons in the hippocampus and other areas in the medial temporal lobe (MTL). Although this seems a small number, they note that with about 1 billion neurons in the MTL, this means that around 2 million neurons participate in the pattern associated with every object. From this and further considerations, they concluded that

each MTL neuron may respond to 50–150 different objects.[1] It may be noted that the hippocampus is thought to use very sparse representations compared with other regions of the brain. Thus, it seems likely that most neurons participate in representing at least hundreds of objects.

In summary, physiological evidence does not appear to help the case in favor of the localist representation scheme that Bowers (2009) defended. Let us now consider his contention that aspects of the neurophysiological data are incompatible with the distributed alternative. Although his arguments here may seem compelling at first glance, Bowers's (2009) claims concerning the properties of distributed representations require closer scrutiny before reaching a conclusion.

Distributed Representations and the PDP Approach

Bowers (2009) distinguished three types of distributed representations based on how many units participate in the representation and on whether the multiple things to which a given unit responds are similar or unrelated: dense coding (many units, each responding to unrelated things), coarse coding (many units, each responding to similar things), and sparse coding (few units, each responding to unrelated things). Bowers treated these as distinct alternatives, arguing that on a distributed account, sparse coding applies only within the hippocampus and coarse coding applies, if at all, within the dorsal pathway and motor system. In contrast, he asserted that distributed accounts assume that the ventral pathway uses what he called dense representations.

There is room for improvement in Bowers's (2009) terminology, since it partially conflates the sparsity of the representation with the relatedness of the things to which a neuron responds. It may be helpful instead to distinguish two dimensions: *sparsity*, defined as to the fraction of neurons in a population that are activated by something,[2] and *perplexity*, defined as the degree to which the things a neuron responds to are unrelated. We say that a neuron's response is *multiplex* if it responds to a number of apparently unrelated entities. It is noteworthy that a sparse representation can be quite multiplex. For example, in the rodent hippocampus, the very same neuron can have completely nonoverlapping place fields even in two highly similar environments (Leutgeb & Leutgeb, 2007).

Most of Bowers's (2009) criticism was directed against dense, multiplex coding, which he largely equates with the type of internal representations learned by PDP networks. However, he recognized that the two issues are separable.

> Many researchers consider dense distributed representations a core theoretical claim of the PDP approach (e.g., Bowers, 2002; Elman, 1995; Hummel, 2000; Page, 2000; Smolensky, 1988; Thorpe, 1989). If it turns out that many current PDP models of memory, language, and perception do learn sparse, coarse, or local codes (contrary to the widespread assumption), or if these models are modified so that they learn these types of representations (in order to be consistent with biology), it would amount to a falsification of this theoretical assumption. At minimum, the neuroscience makes it necessary to think about the PDP approach in a fundamentally different way. (Bowers, 2009, p. 238)

Bowers's characterization of dense multiplex coding as a core theoretical claim of the PDP approach is incorrect.[3] In fact, the approach takes no specific stance on the number of units that should be active in representing a given entity or in the degree of similarity of the entities to which a given unit responds. Rather, one of the main tenets of the approach is to discover rather than stipulate representations (Plaut & McClelland, 2000). Internal representations are learned under the pressure of various demands, and the degree to which they exhibit dense or sparse activation or have units that respond to similar or unrelated things is a consequence of the basic network mechanisms, the learning procedure, and the structure of the tasks to be learned. In general, systematic tasks—in which similar inputs map to similar outputs—yield denser activation (to support generalization), whereas unsystematic tasks (e.g., word and face recognition) give rise to sparser activation (to avoid interference). Moreover, if a unit responds to one pattern, it will tend to respond to other similar patterns because its input is a linear sum of the contributions of incoming connections, although this tendency is weaker in unsystematic tasks because weights must grow larger to override the effects of similarity (for discussion, see McClelland, McNaughton, & O'Reilly, 1995; Plaut, McClelland, Seidenberg, & Patterson, 1996).

Thus, it is a mistake to treat different points in the two-dimensional space of sparsity and perplexity as distinct alternatives. Both the number of active units and the degree of similarity among the things to which they each respond are dimensions that can vary between the extremes that Bowers (2009) considered, and all of the intermediate combinations can be understood as parametric variations within the space of distributed representations. In fact, one of the theoretical strengths of the PDP approach is that it provides a computational framework in which to explore the implications of representations throughout this space to discover which types are most effective in which contexts.

To be clear, we are not claiming that the response properties of units in any particular PDP model adequately capture the relevant neurophysiological observations in the corresponding domain. Indeed, most such models (including those critiqued by Bowers, 2009) are directed at accounting for behavioral rather than neural findings. The computational principles that underlie the PDP ap-

---

[1] Bowers (2009, p. 245) claims that the analyses of Waydo, Kraskov, Quian Quiroga, Fried, and Koch (2006, cited by Quian Quiroga et al., 2008) are consistent with the possibility that the 50–150 stimuli to which a given neuron is estimated to respond might all be the same person. This is incorrect. Waydo et al.'s (2006) analysis derived a measure a defined to be the proportion of distinct stimuli to which a given neuron responds, which they estimate to be between .2%–1% for the MTL. For Bowers's claim to hold, $a = 1/U$ (where U is the universe of possible stimuli) so there could only be at most $U = 1/.002 = 500$ possible stimuli, which is far too few.

[2] Since neurons' activation is not strictly all or none, a more sophisticated definition is generally required, but this definition is a useful first approximation.

[3] It is, perhaps, telling that the majority of researchers cited by Bowers (2009) as considering dense coding to be a core claim of the PDP approach are advocates of localist representations (Bowers, Hummel, Page, Thorpe) and that the remaining researchers (Elman, Smolensky) do not claim that a lack of interpretability is a critical property of PDP systems but rather claim that the interpretability of internal representations is irrelevant to the theoretical approach.

proach are intended to capture how brain areas learn to represent and process information as patterns of activity over large groups of neurons rather than the detailed operation of the individual neurons themselves. The fact that such models have been reasonably successful at accounting for behavioral phenomena across a broad range of domains suggests that the computational principles are capturing something important about neural computation.

Even so, there are clearly many aspects of the standard PDP framework that do not emulate known aspects of neurophysiology: the lack of separate excitatory and inhibitory cell populations, the purely linear integration of inputs with no consideration of dendritic geometry, the use of a real-valued symmetric activation function, no consideration of metabolic constraints, and the propagation of error signals back through forward-going connections, to mention only a few. However, as has repeatedly been emphasized, PDP models are generally not intended to emulate all aspects of the underlying neural substrate: The models are intended to abstract away from many details. This is not to say that physiology cannot inform the PDP framework: Some of the behavioral consequences of the framework would no doubt be improved if it were brought into closer correspondence with neurophysiological findings. Given these points, we agree with Bowers (2009) that a careful consideration of findings from neuroscience motivates modifications and elaborations of the PDP approach that will ultimately enable it to provide better accounts of both neural and behavioral phenomena. Yet, it is essential to avoid the belief that the only valid model at the behavioral level is one that can capture all of the detail at the physiological level. Simplification is of the essence in successful modeling (see McClelland, 2009); what is essential is identifying the appropriate simplifications, and determining which details matter.

To us, it seems most important that the representations used in our models capture the same similarity structure that is captured by neural representations in the brain and not that the individual neurons participating in these representations have individually interpretable (i.e., low perplexity) responses. In this light, we are heartened by recent results from an extensive neurophysiological investigation of the patterns of activation produced in monkey inferotemporal cortex in response to a wide range of different pictures of natural and man-made objects (Kiani, Esteky, Mirpour, & Tanaka, 2007). These patterns appear to capture the same sort of cluster structure seen in learned distributed representations (e.g., Elman, 1990; Rogers et al., 2004). It may be instructive to analyze this data set to understand better whether the individual neurons involved in these representations have interpretable response profiles in response to the stimuli used, but this may not change the essential functional characteristic of these representations.

Problems With the Notion That a Neuron Could Represent "One Thing"

Bowers's (2009) main goal was "to show that the current findings in neuroscience are compatible with localist models in psychology" (p. 221), but he never fully specified the underlying assumptions of such models beyond the proposal that there is a dedicated unit (or set of units) for each familiar thing, among which he included words, faces, and objects. To us, a key problematic aspect of this becomes clear if we focus attention on the question of just exactly what set of (experiences with) entities in the world ought to be treated as "the same thing."

For the sake of discussion, let us begin with the concept of a grandmother cell. When one speaks of "person X's grandmother," it seems clear that one is speaking of a single entity—a certain specific person, such as Mrs. Ethyl Watts Shaffer (Jay McClelland's maternal grandmother). From an ontological point of view, this seems among the least problematic of cases. Other cases of specific objects include, for example, David Plaut's 2001 Volkswagen Jetta, one of the tulips in the vase on Jay McClelland's dining room table, and (one may imagine) the piece of toast Jeffrey Bowers had as part of his breakfast on the morning of February 28, 2009. Does it make sense to assert that people have single neurons (or groups of neurons) dedicated to each of these objects? To continue the last example, since the specific piece of toast was only ever encountered once, Bowers's recognition of it (as a piece of toast) cannot be attributed to previous experience with that particular object but must depend on previous experiences with other objects—other pieces of toast he presumably has encountered at other times and places.

Given that so many of the objects one encounters are encountered exactly once (e.g., cars passed on the road, dogs seen in the park), it seems necessary to build a theory of recognition that encompasses not only familiar entities that one encounters repeatedly, such as grandmothers and automobiles but also those entities that one generally encounters as not repeated but related instances, including tulips and pieces of toast.

A localist theory can, in fact, be extended to the latter sort of entity—indeed, the use of word units in the IA Model is an example. The localist unit for the word *time* is not a unit for a single entity like one's grandmother, but rather a unit for a class of objects generally taken to be tokens of the word *time*. A similar approach might be taken to other classes of objects, including tulips, pieces of toast, cars, dogs, and so on. In general, it would appear that for a localist theory to be of interest, the things localist units should represent should be thought of as including classes of objects in addition to specific instances.

But the use of localist representations for many of the classes of objects one encounters immediately becomes deeply problematic. To see this, consider that there are a huge number of different kinds of tulips and different kinds of pieces of toast. There are different kinds of bread that may be toasted, differences in the details of the manner of toasting, and differences in the way the bread might have been baked, sliced, or complemented with butter, jam, or (perhaps) Marmite. A localist theorist would likely claim that only some of these distinctions should be reified by assigning a localist unit to each of the alternative subvarieties, but wherever the localist stops (whether manually or on the basis of a vigilance parameter; Grossberg, 1987), a problem will remain: There will be some further distinctions that are important in some contexts but that are represented identically by the same localist unit.

A final move a localist theorist might make (and this is indeed a common approach in many areas of psycholinguistics and cognitive science) is to assert that each token is assigned its own distinct localist unit and that recognition of the next distinct token involves contributions from an ensemble of these units. For many researchers this approach has been appealing (e.g., Johnson, 1997; Pierrehumbert, 2001), and it may seem at first glance to be an escape for the localist approach. However, if this approach is adopted, it amounts to accepting that the knowledge that subserves

the recognition of an object is not stored in the connections of a single localist unit or even in a dedicated set of units corresponding to that single object but is, instead, stored in connections involving a large number of contributing units. The piece of toast Jeffrey Bowers eats every morning will depend on the contributions of all of the instances, not of the same piece of toast but of other pieces of toast Jeff has previously encountered—and this, surely, is not a localist representation.

By contrast, these problems are immediately solved by assuming instead that distinctions among different instances of a larger class depend on differences in (learned, experience-dependent) distributed representations. In such representations, similarity is captured by pattern overlap, whereas differences can be captured by the extent to which patterns do not overlap. Experience, including experience with the degree to which particular distinctions matter and the context in which they matter, will affect the degree of overlap in the distributed representations.

In short, we see little prospect for a coherent theory of localist representation. The representations the brain uses may be different from those that emerge from learning in some PDP models, but they are unlikely to be localist in nature.

Conclusions

Bowers (2009) reviewed a range of neurophysiological data indicating that although very rare, some neurons exhibit surprisingly selective responses to familiar entities, such as faces or objects. He interpreted the findings as supportive of theories in which the knowledge of such entities is localized to specific, dedicated units and as problematic for theories in which knowledge of multiple entities is distributed and overlapping. The findings themselves are certainly provocative and challenge certain types of distributed models. However, Bowers took an overly narrow view of the possible range of distributed representations and their properties and underestimated the value of PDP models in exploring these properties. Moreover, the localist theory he espoused runs into difficulty when confronting how people learn and generalize their knowledge. Distributed models need to make greater contact with neurophysiological data but should not be abandoned for localist ones.

References

Barlow, H. (1972). Single units and sensation: A neuron doctrine for perceptual psychology. *Perception, 1,* 371–394.

Bowers, J. S. (2009). On the biological plausibility of grandmother cells: Implications for neural network theories in psychology and neuroscience. *Psychological Review, 116,* 220–251.

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

Gross, C. (2002). Genealogy of the grandmother cell. *The Neuroscientist, 8,* 512–518.

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

Hinton, G. E., McClelland, J. L., & 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. Volume 1: Foundations* (pp. 77–109). Cambridge, MA: MIT Press.

Johnson, K. (1997). Speech perception without speaker normalization. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–165). San Diego, CA: Academic Press.

Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology, 97,* 4296–4309.

Leutgeb, S., & Leutgeb, J. K. (2007). Pattern separation, pattern completion, and new neuronal codes within a continuous CA3 map. *Learning & Memory, 14,* 745–757.

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science, 1,* 11–38.

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: 1. An account of basic findings. *Psychological Review, 88,* 375–407.

Page, M. P. A. (2000). Connectionist modeling in psychology: A localist manifesto. *Behavioral and Brain Sciences, 23,* 443–512.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. L. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 137–157). Amsterdam, the Netherlands: John Benjamins.

Plaut, D. C., & McClelland, J. L. (2000). Stipulating versus discovering representations. *Behavioral and Brain Sciences, 23,* 489–491.

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.

Quian Quiroga, R., Kreiman, G., Koch, C., & Fried, I. (2008). Sparse but not "Grandmother-cell" coding in the medial temporal lobe, *Trends in Cognitive Sciences, 12,* 87–91.

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., & Patterson, K. (2004). The structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review, 111,* 205–235.

Waydo, S., Kraskov, A., Quian Quiroga, R., Fried, I., & Koch, C. (2006). Sparse representation in the human medial temporal lobe. *Journal of Neuroscience, 26,* 10232–10234.

Young, M. P., & Yamane, S. (1992, May 29). Sparse population coding of faces in the inferotemporal cortex. *Science, 256,* 1327–1331.

Received March 1, 2009
Revision received June 11, 2009
Accepted June 12, 2009 ■

Postscript: Parallel Distributed Processing in Localist Models Without Thresholds

David C. Plaut
Carnegie Mellon University and the Center for the Neural Basis of Cognition

James L. McClelland
Stanford University

Bowers (2010) mischaracterizes the goals of parallel distributed processing (PDP research)—explaining performance on cognitive tasks is the primary motivation. More important, his claim that localist models, such as the interactive activation model, "recognize" their inputs when a threshold is reached runs directly counter to an essential feature of these models. This undermines his attempt to distinguish between what a neuron responds to and what it codes for and, indeed, undermines the whole localist argument he has proposed. Bowers's (2010) also continues to face difficulty in specifying what localist units correspond to, and all of the possible choices face problems. In the paragraphs below we substantiate these points.

Goals of the PDP Approach

The PDP approach, for us, is grounded in the belief that certain computational principles of neural systems are fundamental to explaining human cognitive performance. Although we agree with Bowers (2010) that part of the attraction of the approach is its potential to make contact with neural as well as behavioral data, neural verisimilitude per se has not been our primary goal; rather, the approach is directed first and foremost at accounting for performance on cognitive tasks as it occurs in real time, how performance changes over the course of normal and abnormal development and in adulthood, as well as addressing individual differences and the consequences of brain damage.

Graded Interactive Processing Versus Bowers's "Recognition Threshold"

In accounting for human behavior, one aspect of PDP models that is especially critical is their reliance on interactivity and graded constraint satisfaction to derive an interpretation of an input or to select an action that is maximally consistent with all of the system's knowledge (as encoded in connection weights between units). In this regard, models with local and distributed representations can be very similar, and a number of localist models remain highly useful and influential (e.g., Dell, 1986; McClelland & Elman, 1986; McClelland & Rumlehart, 1981; McRae, Spivey-Knowlton, & Tenenhaus, 1998). In fact, given their clear and extensive reliance on parallel distributed processing, we think it makes perfect sense to speak of localist PDP models alongside distributed ones. Some readers may imagine that Bowers's (2010) characterization of localist models is consistent with these models. However, his response to our reply to his target article brings out a crucial disagreement. Bowers (2010) claims that what makes a model localist is that there is a threshold associated with the unit corresponding to an item, such that when that threshold is reached the item is recognized or identified. However, most of the models mentioned do not employ a threshold and, moreover, imposing a threshold would undermine their ability to account for the cognitive phenomena they were designed to address.

The interactive activation model of letter perception, which Bowers (2009) repeatedly cited as his prime example, is a clear case in point. Of great interest to McClelland and Rumelhart (1981) in developing the model was the fact that context facilitates letter perception not only when the letter occurs within a familiar word (such as *cave*) but also when it occurs in an unfamiliar but wordlike pseudoword (such as *mave*). A first crucial assumption in the model was that activations of letter-level units were passed on to units at the word level in a graded and continuous way, so that partial and ambiguous activity at the letter level could propagate forward to the word level. A second crucial assumption was that partial activation at the word level, potentially spread over many word-level units, could contribute feedback activation to the letter level, thereby enhancing the activation of letter-level units for all of the letters presented—even though, taken together, these letters did not exactly match, or lead to the recognition of, any particular word. The power of the model lay precisely in allowing partial activation of units for many different entities to influence processing, without ever employing the concept of a recognition threshold. Although it is sometimes useful to apply a threshold for the purpose of measuring response times, two critical points must be noted: (a) such a threshold is external to the operation of the model and relevant only to response selection; indeed, a natural extension of the model in which partial activation over words would mutually constrain each other in sentence contexts would be precluded by the imposition of recognition thresholds; and (b) the use of a threshold for response selection is itself a simplification that can be useful when downstream processes are not of interest but is problematic when considering the more generally constructive nature of response generation (e.g., in reading aloud a novel item like *mave*, in which partial activation of many word units is thought to contribute; Glushko, 1979). The lack of recognition thresholds in localist models undermines Bowers's (2009) attempt to distinguish between units that code for and units that are activated by an input. It would be highly undesirable for the brain to make such a distinction, and we see no functional or physiological basis for it to do so.

What Does a Localist Unit Represent?

Although localist models (without recognition thresholds) have many useful properties by virtue of engaging in parallel distributed processing, there are two key reasons we prefer models that employ distributed representations: (a) it seems impossible to find a single appropriate granularity for localist representations, making every choice suspect, and (b) in PDP models trained with any one of several powerful learning methods, the granularity of the representations is determined through the course of learning and need not be stipulated by the modeler.

A particularly challenging issue facing localist models is how to capture both the shared and distinctive aspects among a set of entities—for example, the need for related but somewhat different knowledge in dealing with the different types of toast (or tulips)

one might encounter. In his reply, Bowers (2010) failed to address the core of our concern about this matter. PDP models address this issue by learning distributed representations in which regions of overlap among the patterns for different entities capture their commonalities and regions of nonoverlap capture their idiosyncrasies. We pointed out that in the absence of a well-developed learning theory, localist models face an awkward choice as to whether to allocate units to instances or to classes of entities. In his reply, Bowers (2010) reinforced this concern by continuing to equivocate on this issue. In one place, his units stand for a single familiar thing (e.g., one's particular grandmother), "The core claim of a grandmother theory is that single neurons at the top of a hierarchy represent one familiar thing, be it an object, face, or word" (Bowers, 2010, p. 303), while in another place, they stand for an equivalent class of familiar things, "a grandmother cell theory is only committed to the claim that single neurons code for an equivalent class of familiar things" (p. 303)" This of course raises a host of questions, among which are the basis for determining equivalence. Apparently, we are to believe that different grandmothers are not equivalent but that individual tulips are; but what is the rule for deciding? Bowers's (2010) attempt to address these issues is to suggest that the key criterion for allocating a unit is the level at which an entity can be individuated.

> It is only necessary to devote a single unit to a specific tulip on McClelland's dining room table if McClelland can identify it (as opposed to other tulips). Barring this, it is possible that there is a unit for tulips (in general) at the top of his visual processing hierarchy, and . . . the perceptual vividness of each tulip might be due to the specific set of coactive neurons across all the levels of the visual hierarchy. (Bowers, 2010, p. 303)

By "perceptual vividness" Bowers (2010) presumably means perceptual distinctness" of different instances, since in our response we noted that no two tulips are, in fact, the same. However, the problem is not just that different tulips are perceptually distinct but that interacting with different tulips might sometimes require somewhat different knowledge (these tulips are fresher than those and so should be placed in the living room, or these tulips bloom earlier and should be planted to make a nice display with other early bloomers). In localist models in which a single unit is responsible for identification, one may allocate a separate unit for each tulip (or subtype of tulip), but then there is no way to share knowledge based on the similarities of the different tulips; if one instead assigns them all to the same unit, there is no way to treat them differently. One common variant of localist theory addresses this issue by allocating a localist unit to each separate experience with each instance of every different type of entity and then allowing partial activation of each instance to play a role in determining the output of the system (of course, these instances then form a kind of distributed representation). To us, this idea really does not seem biologically plausible, if each localist unit is a neuron or redundant set of neurons. Fortunately, PDP models make this type of localist model unnecessary—each experience produces its own subtle pattern of adjustment to the ensemble of connections among the participating units, allowing distributed representations to capture both the general properties of classes of objects and the specific properties of individual instances (McClelland & Rumelhart, 1985).

In conclusion, we are in full agreement that the PDP approach could be further elaborated to make more direct contact with neural as well as behavioral data; recent efforts along these lines (e.g., O'Reilly & Munakata, 2000; Gotts & Plaut, 2002) suggest that the core computational principles that enable current PDP models to explain normal and impaired cognitive behavior carry forward. However, stipulating the use of localist representations would, in our view, move us away from the goal of developing a more comprehensive and integrated account of the neural basis of cognition.

References

Bowers, J. S. (2009). On the biological plausibilty of grandmother cells: Implications for neural network theories in psychology and neuroscience. *Psychological Review, 116,* 220–251.

Bowers, J. S. (2010). More on grandmother cells and the biological implausibility of PDP models of cognition: A reply to Plaut and McClelland (2010) and Quian Quiroga and Kreiman (2010). *Psychological Review, 117,* 300–308.

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

Glushko, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance, 5,* 674–691.

Gotts, S. J., & Plaut, D. C. (2002). The impact of synaptic depression following brain damage: A connectionist account of "access/refractory" and "degraded-store" semantic impairments. *Cognitive, Affective, and Behavioral Neuroscience, 2,* 187–213.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1–86.

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,* 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,* 159–188.

McRae, K., Spivey-Knowlton, M. J., & Tenenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language, 38,* 283–312.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain.* Cambridge, MA: MIT Press.