Grossberg, S. (2003). Bring ART into the ACT. *Behavioral and Brain Sciences, 26,* 610–611.
Hubel, D. (1995). *Eye, brain, and vision.* New York, NY: Scientific American Library.
Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology: London, 195,* 215–243.
Keysers, C., Xiao, D. K., Foldiak, P., & Perrett, D. I. (2001). The speed of sight. *Journal of Cognitive Neuroscience, 13,* 90–101.
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.
Nadel, L., & Moscovitch, M. (1997). Memory consolidation, retrograde amnesia and the hippocampal complex. *Current Opinion Neurobiology, 7,* 217–227.
Nicholas, M. J., & Newsome, W. T. (2002). Middle temporal visual area microstimulation influences veridical judgments of motion direction. *Journal of Neuroscience, 22,* 9530–9540.
Page, M. P. A. (2000). Connectionist modeling in psychology: A localist manifesto. *Behavioral and Brain Sciences, 23,* 443–512.
Plaut, D. C., & McClelland, J. L. (2010). Locating object knowledge in the brain: Comment on Bowers's (2009) attempt to revive the grandmother cell hypothesis. *Psychological Review, 117,* 284–290.
Poggio, T., & Bizzi, E. (2004, October). Generalization in vision and motor control. *Nature, 431,* 768–774.
Quian Quiroga, R., & Kreiman, G. (2010). Measuring sparseness in the brain: Comment on Bowers (2009). *Psychological Review, 117,* 291–299.
Quian Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005, June). Invariant visual representation by single neurons in the human brain. *Nature, 435,* 1102–1107.
Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2,* 1019–1025.
Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach.* Cambridge, MA: MIT Press.
Seidenberg, M. S., & Plaut, D. C. (2006). Progress in understanding word reading: Data fitting versus theory building. In S. Andrews (Ed.), *From inkmarks to ideas: Current issues in lexical processing.* Hove, England: Psychology Press.
Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences, USA, 104,* 6424–6429.
Shoham, S., O'Connor, D. H., & Segev, R. (2006). How silent is the brain: Is there a "dark matter" problem in neuroscience? *Journal of Comparative Physiology: A. Neuroethology Sensory Neural and Behavioral Physiology, 192,* 777–784.
Sibley, D. E., Kello, C. T., Plaut, D. C., & Elman, J. L. (2009). Sequence encoders enable large-scale lexical modeling: Reply to Bowers and Davis (2009). *Cognitive Science, 33,* 1187–1191.
Smolensky, P. (1988). Putting together connectionism—Again. *Behavioral and Brain Sciences, 11,* 59–70.
Waydo, S., Kraskov, A., Quian Quiroga, R. Q., Fried, I., & Koch, C. (2006). Sparse representation in the human medial temporal lobe. *Journal of Neuroscience, 26,* 10232–10234.

Received July 6, 2009
Revision received September 21, 2009
Accepted September 22, 2009 ■

Postscript: Some Final Thoughts on Grandmother Cells, Distributed Representations, and PDP Models of Cognition

Jeffrey Bowers
University of Bristol

Below, I briefly respond to a number of terminological, theoretical, and empirical issues raised in some postscripts. The goal is not to respond to each outstanding point but rather to address some comments that in my view confuse rather than clarify matters. I respond to Plaut and McClelland (2010) and Quian Quiroga and Kreiman (2010) in turn.

According to Plaut and McClelland (2010), the parallel distributed processing (PDP) approach is defined by its commitment to interactivity and graded constraint satisfaction. Many localist models, including the interactive activation (IA) model, are characterized in this way, and accordingly, they write that "it makes perfect sense to speak of localist PDP models" (p. 289). On this definition, any evidence in support of grandmother cells constitutes a challenge not to the PDP approach per se, just to models that include distributed representations. This characterization of the PDP approach constitutes more of a terminological point than a theoretical point, but it is worth noting that it is inconsistent with many previous statements in which distributed representations are described as a core principle (e.g., Plaut & Shallice, 1993; Seidenberg, 1993). Furthermore, this definition renders the PDP approach so broad that it encompasses almost all neural networks, including network models that are typically seen as inconsistent with the PDP framework (e.g., Grossberg, 1980; Davis, 1999; Hummel & Biederman, 1992). If advocates of the PDP approach are only committed to interactivity and graded constraint satisfaction, with no commitment to the form of the representations that underpin cognition, then there is nothing unique (or novel) about the approach per se.

Even in the context of this broad definition, Plaut and McClelland (2010) argued that my version of a localist model is inconsistent with the PDP approach. That is, I am advocating models in which word, object, and face identification is achieved when a localist representation is activated beyond some threshold. This is said to undermine the key successes of localist PDP models which rely on cascaded processing. For instance, they note that the IA model can explain context effects in letter perception (e.g., a facilitation in identifying a letter embedded in a pseudoword) with the assumption that partial and ambiguous activity at the letter level propagates forward to the word level and partial and ambiguous activity at the word level feeds back to the letter level (although feedback is not strictly necessary to account for the context effects; Grainger & Jacobs, 1996). These context effects in the IA model are observed without thresholds (or identifying any words), and indeed, according to Plaut and McClelland (2010), the inclusion of thresholds would undermine a model's ability to account for the effects.

Plaut and McClelland (2010) appear to have mistaken my comments regarding thresholds with the claim that processing is discrete; that is, when partial activation of letters and words cannot be passed on to subsequent levels and can play no role in processing. In both the target article (Bowers, 2009) and my reply (Bowers, 2010), I

describe localist models in which a given input coactivates multiple units and in which the competition between coactive units plays a role in selecting the target. That is, the competition serves to restrict the number of units that pass some threshold. Thresholds and cascadedness are orthogonal issues, and accordingly, a model with thresholds can account for letter context effects in word perception. Indeed, as noted by Plaut and McClelland (2010), thresholds are often implemented in the IA model. The important point for present purposes is that thresholds in a network in no way undermine the distinction between a unit that codes for an input (e.g., a unit that codes for the word *blue*) and a unit that is only incidentally activated by virtue of form similarity (e.g., a unit coding for *blur* responding to the input *blue*). Equally important, this is all tangential to the question of whether a localist model (PDP or otherwise) is biologically plausible.

Plaut and McClelland (2010) also raised the concern that localist models have no ready way to assign units to inputs. How is the model to know whether a unit should be assigned to a particular grandmother as opposed to grandmothers in general? Or tulips in general as opposed to a particular tulip? What constitutes an equivalence class? They claimed that there are no well-developed learning theories to address these difficult problems and suggested that they may well be intractable for localist approaches in principle. But there are existing implemented localist models that show some promise in addressing these issues. For example, adaptive resonance theory (ART) models of Grossberg (1980, 1987) can learn localist representations at various levels of abstraction. A critical property of these networks is that they include a vigilance parameter that directly affects the granularity of the learned categories. The vigilance parameter is adjusted based on the feedback. If a model makes a mistake in categorizing an input (e.g., categorizing a random old lady as my grandmother or an early blooming tulip as a late bloomer), the vigilance is set higher, and as a consequence, the model learns to categorize perceptually similar inputs with separate localist units. The vigilance parameter also plays a key role in addressing the stability–plasticity dilemma, such that learning new categories (e.g., learning that this specific face belongs to my grandmother) does not erase old knowledge (that my grandmother is an old woman). As a consequence, the model does not have to decide whether to code information at either an abstract or a specific level—it can do both.

Other localist models might be developed to address these concerns as well. For example, consider the model of face identification developed by Riesenhuber and Poggio (1999). A key feature of this model is its hierarchical structure, in which information is coded at various levels of abstraction. For instance, in one layer of the network, the model includes localist units that code for specific views of familiar persons, and in a subsequent layer, units code for familiar persons independent of viewpoint. So once again, the model does not have to choose whether to code a familiar object at an abstract or specific level because it can do both. The Riesenhuber and Poggio (1999) model does not learn, but it is not implausible to imagine a learning algorithm that develops more levels of localist coding as a function of expertise. Just as we are all experts in face recognition and can distinguish one grandmother from another, a florist can distinguish different types of tulips. In both cases, this might be accomplished by the recruitment of localist representations at a subordinate level (in addition to separate units at a basic level). Of course, neither of these models provides a complete answer to these challenging questions (nor do distributed PDP models), but claims regarding the computational limitations of localist models seem premature.

With regards to the neuroscience, Quian Quiroga and Kreiman (2010) highlight that most neurons in their studies responded to more than one image. Even some of the most selective neurons with the medial temporal lobe (MTL) responded to more than one thing—for example, a neuron that fired to two basketball plays, another to two different landmarks, and yet another that responded to Luke Skywalker and Yoda (characters in Star Wars), among other examples. Nevertheless, a few neurons responded robustly to only one out of all the images tested, and the catalogue of examples is expanding. For example, Quian Quiroga, Kraskov, Koch, and Fried (2009) reported a single neuron in MTL that responded to a written word, spoken word, or image of Saddam Hussein but responded to no other stimulus in the experiment. What is to be made of the mixed set of results? Does the fact that most of these neurons responded to more than one image compromise the grandmother cell hypothesis? More generally, is the grandmother cell hypothesis falsified by the Bayesian analysis reported by Waydo, Kraskov, Quian Quiroga, Fried, and Koch (2006) that demonstrates that a given image will inevitably activate many neurons in MTL and that each of these neurons will inevitably respond to many images? I would suggest not. The critical point that needs to be reemphasized is that the units in localist models respond in a similar way; namely, each localist unit responds to more than one input, and a given input activates more than one unit. That is, lifetime sparseness and population sparseness in both localist models and the MTL are extremely high but are still not at the limit of sparseness. The analysis of Waydo et al. (2006) is an important way forward in characterizing the response profiles of neurons in the MTL, but given the range of possible estimates of these measures at present, it is not appropriate to reject localist representations (or grandmother cells) on the basis of their data just yet. What would falsify a grandmother cell theory is an estimate of lifetime and population sparseness in IT that falls outside the range of plausible values for localist models.

This relates to a more fundamental problem with Quian Quiroga and Kreiman's (2010) position. When they rejected the distinction between what a neuron "codes for" and what it "responds to," they are rejecting a fundamental distinction between localist and distributed networks. By ignoring this distinction, they only end up rejecting a straw-man version of a grandmother cell theory. Our impasse on this point might reflect a confusion of terminology between disciplines, and it might be helpful to put the issue in another way. Consider again the neurobiological model of face perception by Riesenhuber and Poggio (1999), inspired by single cell recording data. In this model, individual units are tuned to respond to specific familiar faces, and at the same time, a specific input activates more than one face unit (the target face and units tuned to other similar faces). On my definition, this constitutes a localist model in which each unit represents one specific face (and does not contribute to the representation of other faces). To see this, consider what would happen to the identification of a familiar face if all the coactive units were removed from the network (apart from the unit tuned to the target). The answer is that the model would continue to recognize the face just fine. Conversely, if this one unit was removed from the network, the model would fail to recognize the input as a familiar face. This raises the following

question: Do Quian Quiroga and Kreiman (2010) take their data as inconsistent with this modeling approach? If not, we are essentially in agreement—single cell recording data are consistent with models that work very much like the localist models in psychology.

Finally, Quian Quiroga and Kreiman (2010) reiterated their claim that all the information required for object recognition is in the retina, but in a distributed and implicit code. They reject my claim that a great deal of information required to identify words, objects, and people is located outside the retina, in higher levels of the visual processing pathway. This is said to violate a data processing inequality, according to which processing cannot add information. But there is something wrong with this characterization of the processing inequality. It is clear is that the retina does not include the information about what letter strings constitute words or what configuration of active ganglion cells constitutes an image from my grandmother. This information is stored in higher levels of the visual system, acquired through experience. Similarly, evolution may have endowed higher level visual systems with computational principles to derive shape from shading, depth cues, and so on. Although I agree that processing (transforming) information in and of itself cannot add new information (all transformations are by definition derivable from the input), bottom-up information can nevertheless access other databases of knowledge that contain new information that cannot be derived from the input alone. For example, on viewing a duck, I can predict that the duck might quack. This surely does constitute a violation of data processing inequality.

To conclude, I make one observation that should prove uncontroversial. Regardless of one's position regarding the localist–distributed debate, the target article (Bowers, 2009) highlights a promising approach for evaluating network models in the future, namely, exploring the responses of hidden units one at a time in response to a wide range of inputs. There is a striking disconnect between the methods of neurophysiology, in which neurons are studied one at a time, and the methods in cognitive science, in which hidden units in PDP models are generally studied in combination. This disconnect constitutes a missed opportunity to provide some important constraints on theorizing. An analysis of single units may provide some insights into the conditions under which different coding schemes emerge in neural network models and some insights into why the brain adopts the solutions it does. These analyses might even show that localist representations are required to solve some fundamental computational tasks in perception and memory.

References

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.

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.

Davis, C. J. (1999). The self-organising lexical acquisition and recognition (SOLAR) model of visual word recognition (Doctoral dissertation, University of New South Wales, Sydney, New South Wales, Australia, 1999). *Dissertation Abstracts International, 62,* 594.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review, 103,* 518–565.

Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review, 87,* 1–51.

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

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review, 99,* 480–517.

Plaut, D. C., & McClelland, J. L. (2010). Postscript: Parallel distributed processing in localist models without thresholds. *Psychological Review, 117,* 284–290.

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

Quian Quiroga, R., Kraskov, A., Koch, C., & Fried, I. (2009). Explicit encoding of multimodal percepts by single neurons in the human brain. *Current Biology, 19,* 1308–1313.

Quian Quiroga, R., & Kreiman, G. (2010). Measuring sparseness in the brain: Comment on Bowers (2009). *Psychological Review, 117,* 291–299.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2,* 1019–1025.

Seidenberg, M. S. (1993). Connectionist models and cognitive theory. *Psychological Science, 4,* 228–235.

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.