Mechanisms and Model-Based Functional Magnetic Resonance Imaging

Mark Povich*†

Mechanistic explanations satisfy widely held norms of explanation: the ability to manipulate and answer counterfactual questions about the explanandum phenomenon. A currently debated issue is whether any nonmechanistic explanations can satisfy these explanatory norms. Weiskopf argues that the models of object recognition and categorization, JIM, SUSTAIN, and ALCOVE, are not mechanistic yet satisfy these norms of explanation. In this article I argue that these models are mechanism sketches. My argument applies recent research using model-based functional magnetic resonance imaging, a novel neuroimaging method whose significance for current debates on psychological models and mechanistic explanation has yet to be explored.

1. Introduction. A mechanistic explanation of a phenomenon describes the entities, activities, and organization of the mechanism that produces, underlies, or maintains that phenomenon (Bechtel and Abrahamsen 2005; Craver 2007). Mechanistic explanations satisfy what are widely considered the normative constraints on explanation: the ability to answer a range of counterfactual questions regarding the explanandum phenomenon and the ability to manipulate and control the explanandum phenomenon (Craver 2007). These norms capture what is distinctive about the scientific achievement of *explanation* as opposed to prediction, description, or categorization. A currently debated issue is whether any nonmechanistic forms of explanation can satisfy these explanatory norms.[1] Weiskopf (2011) argues that the

*To contact the author, please write to: Philosophy-Neuroscience-Psychology Program, Washington University in St. Louis, One Brookings Drive, St. Louis, MO 63130; e-mail: mapovich@gmail.com.

†I thank Carl Craver, Ron Mallon, Gualtiero Piccinini, and Dan Weiskopf for invaluable comments.

[1] Batterman and Rice (2014) is a provocative recent paper arguing affirmatively.

Philosophy of Science, 82 (December 2015) pp. 1035–1046. 0031-8248/2015/8205-0025$10.00

models of object recognition and categorization, JIM, SUSTAIN, and ALCOVE, are not mechanistic explanations and nonetheless satisfy these normative constraints.

I argue that JIM, SUSTAIN, and ALCOVE are in fact, and are intended by their creators to be, mechanism sketches, that is, incomplete mechanistic explanations. My argument applies recent research using model-based functional magnetic resonance imaging (fMRI). Model-based fMRI allows cognitive neuroscientists to locate even widely distributed neural components in psychological models. These novel neuroimaging methods have developed only recently (Glascher and O'Doherty 2010), and philosophers have yet to discuss their significance for current debates on psychological models and mechanistic explanation.

The article is organized as follows. In section 2 I motivate the mechanistic account of explanation and introduce two important distinctions in that account: complete models versus mechanism sketches, and how-possibly versus how-actually models. In section 3 I introduce the three models of object recognition and categorization that Weiskopf takes as the scientific grounds for his philosophical thesis. In section 4 I present Weiskopf's arguments for thinking that these models are nonmechanistic yet explanatory. I also begin to respond to these arguments. I show precisely why JIM should be seen as a mechanism sketch. In section 5 I show how the inventors of SUSTAIN and ALCOVE have subsequently used model-based fMRI to decide between these mechanism sketches on the basis of information about widely distributed parts.

2. Mechanistic Explanation. The mechanistic account of explanation developed out of Salmon's (1984) insight into the problems that arise when an account of explanation is tied too closely to prediction. Salmon's principal target was the deductive-nomological account. According to the deductive-nomological account (Hempel and Oppenheim 1948), an explanation is an argument with descriptions of at least one law of nature and antecedent conditions as premises and a description of the explanandum phenomenon as the conclusion. On this view, to explain is to show that the explanandum phenomenon is predictable on the basis of at least one law of nature and certain specific antecedent and boundary conditions. However, tying explanation this closely to prediction generates some famous problems (Salmon 1989). On such a view, many mere correlations come out as explanatory. For example, a falling barometer reliably predicts the weather, but the falling barometer does not explain the weather. In contrast, on the causal-mechanical view, explanation involves situating the explanandum phenomenon in the causal structure of the world. There is more than one way of situating a phenomenon in the causal structure of the world, and in this article

I am solely concerned with explanations that identify the mechanism that produces, underlies, or maintains the explanandum phenomenon.[2]

If one ties explanation so closely to prediction, one risks missing what makes explanation a distinctive scientific achievement. Weiskopf (2011) and I in fact agree on what makes explanation distinctive: explanations provide the ability to answer a range of counterfactual questions regarding the explanandum phenomenon and the ability to manipulate and control the explanandum phenomenon. Weiskopf and I disagree about what kinds of explanation or model can satisfy these norms.

Within the mechanistic framework there are two important distinctions that will be necessary in the arguments that follow: complete models versus mechanism sketches, and how-possibly versus how-actually models (Craver 2007). Mechanism sketches are incomplete descriptions of mechanisms that contain black boxes and filler terms (Craver 2007, 113). They are still partially explanatory. More details can be added to the model to fill in the gaps, though no model is ever fully complete, just complete enough for practical purposes. There can certainly be too many details for the purposes of the modeler, and the details that are included should be relevant.[3] Idealized models can be mechanistic explanations even if they are in some sense incomplete; they can exclude irrelevant detail.

A how-possibly model describes a merely possible mechanism, whereas a how-actually model describes the mechanism that (we have the most evidence to believe) actually produces, maintains, or underlies the explanandum phenomenon. As Weiskopf (2011, 315) rightly points out, this distinction is epistemic. Turning a how-possibly model into a how-actually model does not require modifying the model itself in any way; it requires testing the model. The greater the evidential support for the model, the more how-actually it is. In contrast, turning a mechanism sketch into a more complete mechanistic explanation requires modifying the model by filling in missing details.

3. JIM, SUSTAIN, and ALCOVE. In this section I introduce the models of object recognition and categorization on which Weiskopf builds his case for the existence of nonmechanistic yet explanatory models. In section 4 I present Weiskopf's arguments for thinking that these models are nonmechanistic yet explanatory.

[2] See Bechtel (2009) for a discussion of some other ways of causally situating a phenomenon. What Bechtel calls "looking down" I am here calling "mechanistic explanation."

[3] See Craver (2007, 139–60) for one account of constitutive (i.e., mechanistic) relevance.

According to JIM (John and Irv's Model), in perception objects are broken down into viewpoint-invariant primitives called "geons." Geons are simple three-dimensional shapes such as cones, bricks, and cylinders. The properties of geons are intended to be nonaccidental properties (NAPs), largely unaffected by rotation in depth (Biederman 2000). Objects are represented as spatially arranged collections of geons. The geon structure of perceived objects is extracted and stored in memory for later use in comparison and classification.

The importance of NAPs is shown by the fact that sequential matching tasks are extremely easy when stimuli differ only in NAPs. If you are first shown a stimulus and then a series of rotated stimuli, each of which differs from the first only in NAPs, it is a simple matter to judge which stimuli are the same as or different from the first. Sequential matching tasks with objects that differ in properties that are affected by rotation in depth are much harder.

In JIM, this object recognition and categorization process is modeled by a seven-layer neural network (Biederman et al. 1993). Layer 1 extracts image edges from an input of a line drawing that represents the orientation and depth of an object (Biederman et al. 1993, 182). Layer 2 has three components that represent vertices, axes, and blobs. Layer 3 represents geon attributes such as size, orientation, and aspect ratio. Layers 4 and 5 both derive invariant relations from the extracted geon attributes. Layer 6 receives inputs from layers 3 and 5 and assembles geon features, for example, "slightly elongated, vertical cone above, perpendicular to and smaller than something" (184). Layer 7 integrates successive outputs from layer 6 and produces an object judgment.

ALCOVE (Attention Learning Covering map), like JIM, is a neural network model of object categorization (Kruschke 1992). It has three layers. The perceived stimulus is represented as a point in a multidimensional psychological space with each input node representing a single, continuous psychological dimension. For example, a node may represent perceived size, in which case the greater the perceived size of the stimulus, the greater the activation of that node. Each node is modulated by an attentional gate whose strength reflects the relevance of that dimension for the categorization task. Each hidden node represents an exemplar and is activated in proportion to the psychological similarity of the input stimulus to the exemplar. Output nodes represent category responses and are activated by summing hidden nodes and multiplying by the corresponding weights.

SUSTAIN (Supervised and Unsupervised Stratified Adaptive Incremental Network) is a neural network model similar to ALCOVE (Love, Medin, and Gureckis 2004). Its input nodes also represent a multidimensional psychological space, but they can take continuous and discrete values. Like ALCOVE, inputs are modulated by an attentional gate. Unlike ALCOVE, which stores all items individually in memory in exemplar nodes, the next layer of SUSTAIN consists of a set of clusters (bundles of features) associated with a category. Each cluster activates in proportion to its proximity to the input in multidimensional psychological space; the more similar a cluster is to the input, the more it activates. There are inhibitory connections between each cluster, so that the cluster most similar to the input inhibits all others. This winning cluster activates the output unit generating the category label.

4. Weiskopf's Arguments. Weiskopf argues that the previous models are able to satisfy the norms of explanation but are not mechanistic models. How do these models provide the ability to answer counterfactual questions about the explanandum phenomenon and the ability to manipulate and control the explanandum phenomenon? According to Weiskopf, they satisfy these explanatory norms "because these models depict one aspect of the causal structure of the system" (2011, 334). This claim is prima facie in tension with Weiskopf's claim that these models are not mechanistic. He argues that "there may be an underlying mechanistic neural system, but this mechanistic structure is not what cognitive models capture" (333).

One way of reconciling the above claims is to argue that these models are explanatory because they depict causal structure, but they are not mechanistic because the causal structure that they depict is not a mechanism. This is the line Weiskopf takes. Why, according to Weiskopf, are these causal structures not mechanisms? He argues, "If parts [of mechanisms] are allowed to be smeared-out processes or distributed system-level properties, the spatial organization of mechanisms becomes much more difficult to discern. . . . Weakening the spatial organization constraint by allowing distributed, nonlocalized parts incurs costs, in the form of greater difficulty in locating the boundaries of mechanisms and stating their individuation conditions" (Weiskopf 2011, 334). The causal structures depicted by JIM, SUSTAIN, and ALCOVE should not be thought of as mechanisms, according to Weiskopf, because the structures that putatively implement them are highly distributed. If mechanisms are allowed to contain distributed, nonlocalized parts, this will make it difficult to locate them. Call this the practical problem of nonlocalization. Weiskopf does not provide any reason to think that the philosophical (rather than practical) problem of mechanism individuation is made more difficult by allowing distributed parts or that existing accounts of mechanism individuation cannot handle distributed parts.[4] Yet numerous neuroimaging methods, especially model-based fMRI, ameliorate this

[4] See n. 3 for an account of mechanism individuation. Weiskopf (2011, 331) also cites the phenomenon of neural reuse as inconsistent with mechanistic explanation, but the fact that a part of one mechanism can also be a part of a different mechanism constitutes only a practical problem for mechanism individuation.

practical problem. Model-based fMRI is well suited to mechanistically discriminate between competing, equally behaviorally confirmed psychological models.[5]

In addition to Weiskopf's practical problem, there is what I call the triviality problem of nonlocalization. Weiskopf argues that if these kinds of distributed parts are allowed, then "it is far from clear what content the notion of a mechanism has anymore" (2011, 334). First, as I have said, there has been no argument that existing accounts of mechanism individuation cannot accommodate distributed parts. If these accounts are workable while allowing distributed parts, then the notion of a mechanism remains contentful. Second, this objection misunderstands the mechanistic project, or at least a plausible way of conceiving that project. If you conceive the mechanistic project as articulating a "downward" way of causally situating an explanandum phenomenon that was neglected by Salmon and others who focused on "backward" (etiological) causal explanation (Craver 2007, 8), then a "liberalization" of the notion of mechanism that permits distributed parts is perfectly in line with that project and should not be seen as any kind of concession or retreat. Although such a "liberalization" may make mechanisms even more ubiquitous than they already were, it does not make every physical system a mechanism. For example, mere aggregates lack the organization necessary to be mechanisms (135–39).

Next, I present some of the neuroimaging studies conducted with JIM and argue that JIM is a mechanism sketch. JIM was built not merely to produce the same behavior as human beings in object recognition tasks, but to model something that might really be happening in human brains (Biederman et al. 1993, 176). Accordingly, Irving Biederman, one of the cocreators of JIM, and others have conducted various neuroimaging studies to investigate the neural underpinnings of the model.

If JIM is a mechanism sketch, the systems and processes in the model required for the extraction, storage, and comparison of geon structures must to some extent correspond to (perhaps distributed) components in the brain's actual object recognition system. For example, if JIM is a mechanism sketch, there is an area or a configuration of areas in the brain where simple parts and NAPs are represented. In one study investigating this (Hayworth and Biederman 2006), participants were shown line drawings that were either local feature deleted (LFD), in which every other vertex and line was deleted from

[5] Weiskopf (2011, 335–36) is right that evidence for psychological models can come from many places. Although psychological models can be supported and constrained behaviorally, this degree of "evidential autonomy" does not establish the explanatory autonomy Weiskopf requires. It does not affect the mechanist's point that the parts of a psychological model must correspond to brain regions that implement the relevant computations for the model to be explanatory.

each part, removing half the contour, or part deleted (PD), in which half of the parts were deleted. On each trial, participants saw either LFD or PD stimuli presented as a sequential pair and had to report whether the exemplar depicted by the second stimulus was the same as or different from that depicted by the first. The second stimulus was always mirror-reversed with respect to the first. Each experimental run was composed of an equal number of three conditions: identical, complementary, and different exemplar. In the identical condition, the second stimulus was identical to the first stimulus (though mirror-reversed). In the complementary condition, the second stimulus depicted the same exemplar as the first, but the second stimulus was a "complement" of the first stimulus. An LFD complement is composed of the deleted contour of the first stimulus, and a PD complement is composed of the deleted parts of the first stimulus. In the different exemplar condition, the second stimulus depicts a different exemplar than the first.

This study used an fMRI adaptation design that relies on the assumption that when two successive stimuli activate the same brain region, neural activity reduces (Krekelberg, Boynton, van Wezel 2006, 250). The results of the study showed adaptation between LFD complements and lack of adaptation between PD complements in lateral occipital complex, especially the posterior fusiform area, an area known to be involved in object recognition. These results imply that this area is "representing the parts of an object, rather than local features, templates, or object concepts" (Hayworth and Biederman 2006, 4029). Biederman has conducted many other fMRI experiments, including some that "suggest that LO [lateral occipital cortex] is the locus of the neural correlate for the greater detectability for nonaccidental relations" (Kim and Biederman 2012, 1824).

Although these experiments suggest that JIM should be seen as a mechanism sketch, Weiskopf has another argument for why it should not: JIM has properties that do not and could not correspond to anything in the brain. Weiskopf (2011, 331) refers to JIM's "fast enabling links" (FELs), which allow the model to bind representations and have infinite propagation speed. Weiskopf calls FELs an example of "fictionalization," or "putting components into a model that are known not to correspond to any element of the modeled system, but which serve an essential role in getting the model to operate correctly" (331). The FELs, Weiskopf argues, undermine the claim that JIM is a mechanism sketch.

Weiskopf is right that FELs are an essential fictionalization. However, playing an essential role in getting a model to operate is not the same as explaining; these parts of the model carry no explanatory information and render the model, or at least part of it, how-possibly (where the possibility involved is not physical possibility, since FELs are physically impossible). FELs play the black box role of whatever it is that accounts for binding. In addition to playing a black box role, they serve practical and epistemic purposes such as suggesting, constraining, and sharpening questions about mechanisms (Bogen 2005). Let me explain how by comparing FELs to Bogen's example of the Goldman, Hodgkin, and Katz (GHK) equations.

The GHK voltage and current equations are used to determine the reversal potential across a cell's membrane and the current across the membrane carried by an ion. These equations rely on the incorrect assumptions that each ion channel is homogeneous and that interactions among ions do not influence their flow rate (Bogen 2005, 409). Bogen highlights the effects on research of these incorrect assumptions: "Investigators used these and other GHK equation failures as problems to be solved by finding out more about how ion channels work. Fine-grained descriptions of exceptions to the GHK equations and the conditions under which they occur sharpened the problems and provided hints about how to approach them" (410). The GHK equations provide a case of "using incorrect generalizations to articulate and develop mechanistic explanations" (409). Something similar can be said about FELs. Not only do FELs play an essential black box role, but they also suggest new questions about mechanisms, new problems to be solved. For example, Hummel and Biederman write, "[FELs allow] JIM to treat the constraints on feature linking (by synchrony) separately from the constraints on property inference (by excitation and inhibition). That is, cells can phase lock without influencing one another's level of activity and vice versa. Although it remains an open question whether a neuroanatomical analog of FELs will be found to exist, we suggest that the distinction between feature linking and property inference is likely to remain an important one" (1992, 510). Like the GHK equations, FELs suggest new lines of investigation, in this case regarding the relation between feature linking, property inference, and their neural mechanisms. Specifically, FELs suggest research questions such as "Can biological neurons phase lock without influencing one another's activity?" and "Are there other ways biological neurons could implement feature linking and property inference independently?"

In the next section I will explain model-based fMRI and demonstrate how recent model-based fMRI research shows that, like JIM, SUSTAIN and ALCOVE are mechanism sketches.

5. Model-Based fMRI. fMRI is a neuroimaging method that provides an indirect measure of neuronal activity. More specifically, fMRI measures a physiological indicator of oxygen consumption that correlates with changes in neuronal activity (Huettel, Song, and McCarthy 2009, 159–60).

Model-based fMRI is a neuroimaging method that combines psychological models with fMRI data. It "provides insight into 'how' a particular cognitive function might be implemented in the brain, not only 'where' it is implemented" (O'Doherty, Hampton, and Kim 2007, 39). In this way, model-based fMRI provides a way of discriminating between competing, equally behaviorally confirmed cognitive models (Glascher and O'Doherty 2010, 502). Furthermore, "the more complex the model (and hence the more associated free parameters), the more unconstrained the behavioral fitting becomes," in which case the additional constraints imposed by neurophysiological and neuroimaging data become "even more critical" (O'Doherty et al. 2007, 37; see also White and Poldrack 2013).

To conduct a model-based fMRI analysis, one starts with a psychological model that postulates internal variables between stimulus input and behavioral output. While research participants perform a model-relevant task, researchers obtain fMRI data from which they can locate neural correlates of the internal variables (O'Doherty et al. 2007, 36). The model-predicted values of internal variables across trials are convolved (mathematically combined) with a canonical hemodynamic response function (HRF; Glascher and O'Doherty 2010, 505). This is done to account for the usual lag in the hemodynamic response (O'Doherty et al. 2007, 37). This yields a new, model-predicted HRF that can be regressed against the obtained fMRI data. This allows researchers to identify brain areas where the model-predicted HRF significantly correlates with the observed HRF across trials.[6]

I should make clear that model-based fMRI inherits the limitations of fMRI, such as poor spatiotemporal resolution, and does not obviate the need for other neuroimaging methods (e.g., positron emission tomography [PET], electroencephalography [EEG], or magnetoencephalography [MEG]), to which the model-based approach can also be applied.

Now that we have a basic understanding of how model-based fMRI works and what it can accomplish, let me return to SUSTAIN and ALCOVE and show how they are mechanism sketches by drawing on recent model-based fMRI research.

Both models were investigated in a model-based fMRI study in which participants completed a rule-plus-exception category learning task (Davis, Love, and Preston 2012). During the task, a schematic beetle was presented, and participants were asked to classify it as living in hole A or hole B. Participants then received feedback on the correctness of their classification. The beetles varied on four of the following five attributes, with the fifth held constant: eyes (green or red), tail (oval or triangular), legs (thin or thick), antennae (spindly or fuzzy), and fangs (pointy or round). Six of the eight beetles presented could be correctly categorized on the basis of a single attribute. For example, three out of four hole A beetles had thick legs, and three

[6] Batterman and Rice (2014) object that the notion of *correspondence* between model and world is never explained by mechanists. I have no general theory of correspondence, but the sense in which (parts of) a psychological model correspond(s) to (parts of) the brain should be clear in each case. Here, for example, correspondence is significant correlation between model-predicted and observed HRF.

out of four hole B beetles had thin legs. These were the rule-following beetles. The other beetles were exceptions to the rule, having legs that appeared to match the other category.

Two predictions from SUSTAIN and ALCOVE were tested. First, each model predicts specific changes in recognition strength across trials. During stimulus presentation, SUSTAIN predicts a recognition advantage for exceptions; ALCOVE predicts no recognition advantage. This difference in recognition strength predictions arises because in ALCOVE, but not in SUSTAIN, all items are stored individually in memory regardless of whether they are exceptions or rule-following items. Second, each model predicts specific changes in error correction across trials. The amount of error is given by the difference between the model's category response and the correct response. Both SUSTAIN and ALCOVE predict that exceptions will always produce more error than rule-following items, although both will produce less error as learning progresses (Davis et al. 2012, 266).

The results revealed that both the recognition strength and error correction measures predicted by SUSTAIN found significant correlations in medial temporal lobe (MTL) regions, including bilateral hippocampus, parahippocampal cortex, and perirhinal cortex. ALCOVE's predicted recognition strength measure did not find any significant correlations in MTL, although its predicted error correction measure found significant correlations in MTL regions (Davis et al. 2012, 266–67). These results "suggest that, like SUSTAIN, the MTL contributes to category learning by forming specialized category representations appropriate for the learning context" (269).

SUSTAIN is more how-actually (evidentially supported) than ALCOVE because both of SUSTAIN's prediction measures (recognition strength and error correction) were significantly correlated with observed HRF, whereas only one of ALCOVE's prediction measures (error correction) was significantly correlated. These experiments also show that cognitive neuroscientists are currently advancing the ability to map the entities and activities in psychological models to distributed neural systems, such as MTL regions spanning bilateral hippocampus, parahippocampal cortex, and perirhinal cortex.

Davis et al. (2012) are at times quite explicit that they are treating the models as mechanism sketches (see also Love and Gureckis 2007). For instance, they write, "We use a model-based functional magnetic resonance imaging (fMRI) approach to test the proposed mapping between MTL function and SUSTAIN's representational properties" (261). Given their emphasis on mapping models to the brain, it is clear that they intend these models to be mechanistic, as Biederman intends JIM to be. They are interested in more than the behavioral accuracy of these models; after all, SUSTAIN and ALCOVE are already behaviorally well confirmed. The main difference between the two is in their hidden layers, where SUSTAIN has clusters and ALCOVE stores items individually. Model-based fMRI allowed Davis et al. to gather evidence relevant to assessing which of these models was more mechanistically accurate.

6. Conclusion. Weiskopf (2011) presents three models of object recognition and categorization, JIM, ALCOVE, and SUSTAIN, that he claims are nonmechanistic yet explanatory. He argues that they are not mechanistic because their parts cannot be neatly localized and because they contain some components that cannot correspond to anything in the brain, such as FELs, but are nevertheless essential for the proper working of the model. I argue, on the contrary, that in addition to playing a black box role, FELs play useful, nonexplanatory roles such as suggesting new lines of investigation regarding feature linking and property inference.

My argument for the claim that SUSTAIN and ALCOVE are mechanism sketches relies partly on model-based fMRI research. Model-based fMRI and other model-based neuroimaging methods allow cognitive neuroscientists to explore how psychological models map onto the brain. This helps cognitive neuroscientists discriminate between equally behaviorally confirmed psychological models.

Biederman, Love, and others treat JIM, SUSTAIN, and ALCOVE as mechanism sketches, and they should because by locating mechanisms one opens a new range of opportunities for manipulating the mechanism and one obtains answers to counterfactual questions that were not available before. For example: What kinds of deficit in categorization performance would result from a lesion in bilateral hippocampus? If someone has a specific deficit in categorization performance, how might we fix it? Where might the problem lie? This increases the explanatory power of these models.

The development of these model-based approaches has broader implications, beyond the narrow dispute over JIM, SUSTAIN, and ALCOVE, for the debate over the explanatory and mechanistic status of psychological models. As cognitive neuroscientists continue to test competing models against neuroimaging data using model-based techniques, it is likely that they will, as they should, retain those models that are mechanistically accurate and discard those that are not, and in so doing reveal that explanatory progress in cognitive neuroscience consists in the development of increasingly mechanistic models.

REFERENCES

Batterman, Robert, and Collin Rice. 2014. “Minimal Model Explanations.” *Philosophy of Science* 81 (3): 349–76.

Bechtel, William. 2009. “Looking Down, Around, and Up: Mechanistic Explanation in Psychology.” *Philosophical Psychology* 22 (5): 543–64.

Bechtel, William, and Adele Abrahamsen. 2005. “Explanation: A Mechanist Alternative.” *Studies in History and Philosophy of the Biological and Biomedical Sciences* 36 (2): 421–41.

Biederman, Irving. 2000. “Recognizing Depth-Rotated Objects: A Review of Recent Research and Theory.” *Spatial Vision* 13 (2–3): 241–53.

Biederman, Irving, Eric Cooper, John Hummel, and Jozsef Fiser. 1993. “Geon Theory as an Account of Shape Recognition in Mind, Brain and Machine.” In *Proceedings of the 4th British Machine Vision Conference*, ed. John Illingworth, 175–86. London: Springer.

Bogen, Jim. 2005. “Regularities and Causality: Generalizations and Causal Explanations.” *Studies in History and Philosophy of Biology and Biomedical Sciences* 36:397–420.

Craver, Carl. 2007. *Explaining the Brain.* Oxford: Oxford University Press.

Davis, Tyler, Bradley Love, and Alison Preston. 2012. “Learning the Exception to the Rule: Model-Based fMRI Reveals Specialized Representations for Surprising Category Members.” *Cerebral Cortex* 22:260–73.

Glascher, Jan, and John O’Doherty. 2010. “Model-Based Approaches to Neuroimaging: Combining Reinforcement Learning Theory with fMRI Data.” *WIREs Cognitive Science* 1:501–10.

Hayworth, Kenneth, and Irving Biederman. 2006. “Neural Evidence for Intermediate Representations in Object Recognition.” *Vision Research* 46:4024–31.

Hempel, Carl, and Paul Oppenheim. 1948. “Studies in the Logic of Explanation.” *Philosophy of Science* 15:135–75.

Huettel, Scott, Allen Song, and Gregory McCarthy. 2009. *Functional Magnetic Resonance Imaging*. Sunderland, MA: Sinauer.

Hummel, John, and Irving Biederman. 1992. “Dynamic Binding in a Neural Network for Shape Recognition.” *Psychological Review* 99:480–517.

Kim, Jiye, and Irving Biederman. 2012. “Greater Sensitivity to Nonaccidental than Metric Changes in the Relations between Simple Shapes in the Lateral Occipital Cortex.” *NeuroImage* 63: 1818–26.

Krekelberg, Bart, Geoffrey Boynton, and Richard J. A. van Wezel. 2006. “Adaptation: From Single Cells to BOLD Signals.” *TRENDS in Neurosciences* 29 (5): 250–56.

Kruschke, John. 1992. “ALCOVE: An Exemplar-Based Connectionist Model of Category Learning.” *Psychological Review* 99:22–44.

Love, Bradley, and Todd Gureckis. 2007. “Models in Search of a Brain.” *Cognitive, Affective, and Behavioral Neuroscience* 7 (2): 90–108.

Love, Bradley C., Douglas L. Medin, and Todd M. Gureckis. 2004. “SUSTAIN: A Network Model of Category Learning.” *Psychological Review* 111:309–32.

O’Doherty, John, Alan Hampton, and Hackjin Kim. 2007. “Model-Based fMRI and Its Application to Reward Learning and Decision Making.” *Annals of the New York Academy of Sciences* 1104:35–53.

Salmon, Wesley. 1984. *Scientific Explanation and the Causal Structure of the World.* Princeton, NJ: Princeton University Press.

———. 1989. “Four Decades of Scientific Explanation.” In *Minnesota Studies in the Philosophy of Science*, vol. 13, *Scientific Explanation*, ed. Wesley Salmon and Philip Kitcher, 3–219. Minneapolis: University of Minnesota Press.

Weiskopf, Daniel. 2011. “Models and Mechanisms in Psychological Explanation.” *Synthese* 183 (3): 313–38.

White, Corey, and Russell Poldrack. 2013. “Using fMRI to Constrain Theories of Cognition.” *Perspectives on Psychological Science* 8 (1): 79–83.