Mechanisms and Model-Based fMRI

Mark Povich

Washington University in St. Louis

Abstract. Mechanistic explanations satisfy widely held norms of explanation: the ability to control and answer counterfactual questions about the explanandum. A currently debated issue is whether any non-mechanistic explanations can satisfy these explanatory norms. Weiskopf (2011) argues that the models of object recognition and categorization, JIM, SUSTAIN, and ALCOVE, are not mechanistic, yet satisfy these norms of explanation. In this paper I will argue that these models are sketches of mechanisms. My argument will make use of model-based fMRI, a novel neuroimaging approach whose significance for current debates on psychological models and mechanistic explanation has yet to be explored.

Word count: 5000

1. Introduction

According to the mechanistic account of explanation, a phenomenon is explained by describing the entities, activities, and organization of the mechanism that produces, underlies, or maintains the phenomenon (see, e.g., Bechtel and Abrahamsen 2005). Mechanistic explanations satisfy what are widely considered normative constraints of 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* rather than prediction, description, or categorization. A currently debated issue is whether any non-mechanistic forms of explanation can satisfy these explanatory norms.[1] Weiskopf (2011) argues that the models of object recognition and categorization, JIM, SUSTAIN, and ALCOVE, are not mechanistic, yet satisfy these norms of explanation.

In this paper, in part using recent model-based fMRI research, I will argue that JIM, SUSTAIN, and ALCOVE are in fact mechanism-sketches, i.e. incomplete mechanistic explanations. Model-based approaches to neuroimaging allow cognitive neuroscientists to locate the distributed neural components of psychological models. These novel neuroimaging approaches have developed only recently and philosophers have yet to discuss their significance for current debates on psychological models and mechanistic explanation. The

1 A recent paper arguing affirmatively is Batterman and Rice (2014).

opportunity to demonstrate this significance is one advantage of responding to Weiskopf (2011) in particular.

The paper is organized as follows. In Section 2, I will motivate the mechanistic account of explanation and introduce two crucial concepts in the mechanistic account: the mechanism-sketch and the how-possibly model. In Section 3, I will introduce the models of object recognition and categorization (JIM, SUSTAIN, and ALCOVE) that Weiskopf presents as non-mechanistic, yet explanatory. In Section 4, I will present Weiskopf's arguments for thinking these models are non-mechanistic, yet explanatory, and I will begin responding to these arguments. This section demonstrates that JIM is a mechanism-sketch. Demonstrating that SUSTAIN and ALCOVE are mechanism-sketches requires covering recent studies employing model-based fMRI, a novel neuroimaging method that will be explained in section 5.

2. Mechanistic Explanation

Salmon (1984) developed the causal-mechanical account of explanation primarily in response to the covering-law or deductive-nomological model of explanation (Hempel and Oppenheim 1948). According to the deductive-nomological model, 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, explanation is showing that the explanandum phenomenon is predictable given at least one law of nature and certain specific antecedent and boundary conditions. However, tying explanation this closely to prediction generates some famous problems for the covering-law

model (see section 2.3 of Salmon [1989] for a review of these problems). 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 are many ways of situating a phenomenon in the causal structure of the world and in this paper I am solely concerned with explanations that identify the mechanism that produces, underlies, or maintains the explanandum phenomenon.[2]

Another problem with tying explanation so closely to prediction is that we miss what is distinctive about the scientific achievement of explanation. Weiskopf (2011) and I 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. These are the norms of explanation. Weiskopf and I disagree over what kinds of explanation can satisfy these norms.

Within the mechanistic framework there are two important distinctions: between complete mechanistic models and mechanism-sketches and between how-possibly and how-actually models (Craver 2007). Mechanism-sketches are incomplete descriptions of

---

2 Other ways of causally situating a phenomenon include etiologically and contextually situating it. See Bechtel (2009) for a discussion of some of these different forms of causal explanation. What Bechtel calls “looking down” I am here calling “mechanistic explanation.”

mechanisms that may contain black boxes and filler terms (Ibid., 113). Mechanistic models rest on a continuum of *more-or-less* complete (114). As more details are incorporated into the model, the more complete it becomes – though no model is ever fully complete, just complete enough for practical purposes. A more complete model is not necessarily a *better* or *more useful* model. There can certainly be *too many* details for the purposes of the modeler and the details that are included should be relevant.[3] Idealization can be readily accommodated within a mechanistic framework.

A how-possibly model describes a merely possible mechanism, whereas a how-actually model describes the mechanism actually producing, maintaining, or underlying the explanandum phenomenon. As Weiskopf (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 complete(-enough) model 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 JIM, SUSTAIN, and ALCOVE. The next section presents Weiskopf's arguments for thinking these models are non-mechanistic, yet explanatory.

---

3 See Craver (2007, section 4.8) for an 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". These geons are simple three-dimensional shapes such as cones, bricks, and cylinders. The properties of geons are intended to be non-accidental properties (NAPs), largely unaffected by rotation in depth (Biederman 2000). 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 only differ in NAPs. If you are shown a stimulus, then a series of other, 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 than the first. Sequential matching tasks with objects that differ in properties that are affected by rotation are much harder.

In JIM, this object recognition process is modeled by a seven layer neural network (Biederman, Cooper, and Fiser 1993). Layer 1 extracts image edges from an input of a line drawing that represents the orientation and depth of an object (182). Layer 2 has three components which 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, e.g., "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.

The Attention Learning Covering map (ALCOVE) is a 3-layer, feed-forward, neural network model of object categorization (Kruschke 1992). A perceived stimulus is represented as a point in a multi-dimensional 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 a 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.

The Supervised and Unsupervised Stratified Adaptive Incremental Network (SUSTAIN) is a network model of object categorization 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, including category labels. 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 associated with a category. All of SUSTAIN's clusters compete to respond, with inhibitory connections between each cluster, and the cluster closest to the stimulus in the multidimensional space is the winner. The cluster that wins activates the output unit predicting the category label. The output leads to a decision procedure that generates a category response.

4. Weiskopf's Objections

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, and the ability to manipulate and control, the explanandum phenomenon? According to Weiskopf, they satisfy explanatory norms "because these models depict one aspect of the causal structure of the system" (334). This claim is in tension with one reason Weiskopf gives for thinking these models are not mechanistic. He argues, "there may be an underlying mechanistic neural system, but this mechanistic structure is not what cognitive models capture. They capture a level of functional abstraction that this mechanistic structure realizes" (333). But the claim that these models are not mechanistic because they depict a level of functional abstraction, not causal structure, conflicts with the claim that these models are explanatory because they depict causal structure. This conflict results from the general difficulty of specifying how a model can satisfy the norms of explanation without being mechanistic.

One way of trying to reconcile 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 is depicted is not a mechanism. This is the line Weiskopf takes. Why, according to Weiskopf, are these causal structures not mechanisms? He argues that

> 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. (334)

The causal structures depicted by JIM, SUSTAIN, and ALCOVE should not be thought of as mechanisms, according to Weiskopf, because these structures are highly distributed. If mechanisms are allowed to contain distributed parts, this will make locating them difficult. The problem, then, is *practical*. Weiskopf does not give any reason to think the *philosophical* (rather than practical) problem of mechanism individuation is made more difficult by allowing distributed parts. [4] Yet numerous neuroimaging methods, especially model-based fMRI, allow cognitive neuroscientists to locate highly distributed neural mechanisms corresponding to the internal variables of computational models. Cognitive neuroscientists are interested in more than the *behavioral* accuracy of these models; they are also interested in their *mechanistic* accuracy. That cognitive neuroscientists conduct neuroimaging studies using these models shows that they are treated as mechanistic. Next I will present some of the neuroimaging studies conducted with JIM and argue that JIM is a mechanism-sketch.

---

4 Weiskopf (331) also cites the phenomenon of neural reuse as inconsistent with mechanism. This assumes that a part of one mechanism cannot be a part of another mechanism but Weiskopf has not provided any reason to think this nor to think that the possibility of reuse should give rise to any special philosophical (rather than practical) problems of mechanism individuation.

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. write, "We have concentrated on modeling primal access: The initial activation in a human brain of a basic-level representation of an image from an object exemplar, even a novel one, in the absence of any context that might reduce the set of possible objects" (Biederman, Cooper, Hummel and Fiser 1993, 176). Accordingly, Irving Biederman, one of the co-creators 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 actual object recognition mechanisms in the brain. For example, if JIM is a mechanism-sketch, there is an area or a configuration of areas in the brain where simple parts and non-accidental properties are represented. In one study (Hayworth and Biederman 2006), subjects were shown line drawings that were either local feature deleted (LFD), in which every other vertex and line was deleted from each part, removing half the contour, or part deleted (PD) in which half of the parts were removed. On each experimental run, subjects saw either LFD or PD stimuli presented as a sequential pair and had to respond whether or not the exemplars were the same or different. The second stimulus was always mirror-reversed with respect to the first. Each run was comprised of an equal number of three conditions: Identical, Complementary, and Different Exemplar. In the Identical condition, the second stimulus was the same as the first stimulus (mirror-reversed,

as all of the second stimuli were). In the Complementary condition, the second stimulus was the complement of the first, where an LFD-complement is composed of the deleted contour of the first and a PD-complement is composed of the deleted parts of the first. In the Different Exemplar condition, the second stimulus is a line-drawing of a different exemplar than the first.

An fMRI-adaptation design was used, which "relies on the assumption that neural adaptation reduces activity when two successive stimuli activate the same subpopulation but not when they stimulate different subpopulations" (Krekelberg, Boynton, van Wezel 2006, 250; see also Kourtzi and Grill-Spector 2005). 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 1824).

While these results resolve Weiskopf's worry about the difficulty of locating distributed parts, he has another argument for why JIM is not mechanistic. JIM has properties that do not and could not correspond to anything in the brain. Weiskopf (2011, 331) mentions JIM's "Fast Enabling Links" (FELs), which allow the model to bind representations and which have infinite propagation speed. According to Weiskopf, FELs are an example of

fictionalization, "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" (Ibid.), and he argues that this undermines the claim that JIM is a mechanism-sketch. Weiskopf is right that FELs are an essential fictionalization, but 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). Right now 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 like the ones discussed by Bogen (2005), such as suggesting, constraining, and sharpening questions about mechanisms. Let me explain how by comparing FELs to Bogen's example of the GHK equations.

The Goldman, Hodgkin, and Katz (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 rate (Bogen 409). About the inadequacy of these equations Bogen writes,

> While some generalizations are useful because they deliver empirically acceptable quantitative approximations, others are useful because they do not… 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. (Bogen 410)

The GHK equations provide a case of "using incorrect generalizations to articulate and develop mechanistic explanations" (Bogen 409). I argue that something similar can be said about FELs. Not only do FELs play an essential black box role, FELs suggest new questions about mechanisms, new problems to be solved. For example, Hummel and Biederman (1992) write,

> [T]he independence of FELs and standard excitatory-inhibitory connections in JIM has important computational consequences. Specifically, this independence allows 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. (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 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 studies show that SUSTAIN and ALCOVE are mechanism-sketches.

5. Model-Based fMRI

Model-based fMRI is a neuroimaging method that aims to discover the neural mechanisms that correspond to model variables. Model-based fMRI "can be used as a means of discriminating between competing computational models of cognitive and neural function. Thus, model-based fMRI 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 39). In this way, model-based fMRI provides a way of discriminating between competing, equally behaviorally confirmed cognitive models (Glascher and O'Doherty 502).

Functional magnetic resonance imaging (fMRI) is a neuroimaging method that provides an indirect measure of neuronal activity. Neuronal activity requires glucose and oxygen for fuel, which the vascular system provides. The oxygen is bound to hemoglobin molecules and the magnetic properties of deoxygenated hemoglobin are detectable by fMRI. In this way, fMRI measures a physiological indicator of oxygen consumption – deoxyhemoglobin concentration – that correlates with changes in neuronal activity (Huettel, Song, and McCarthy 159-160).

To conduct a model-based fMRI analysis, one starts with a computational model that describes the function(s) by which stimuli are transformed to result in behavioral output. Stimulus input and behavioral output are observable, but the computational model postulates internal variables linking input and output. The neural correlates of these internal variables, at

each time point in the experiment, can then be located using regression analyses (O' Doherty, Hampton, and Kim 36).

The variables that change from trial to trial are converted into a time series of the model-predicted BOLD (blood-oxygen-level dependent) response and then convolved with a canonical hemodynamic response function (Glascher and O'Doherty 505). This just means that the predicted variable values, taken over time, are mathematically combined with a stereotypical BOLD signal function. This is done to account for the usual lag in the hemodynamic response (O' Doherty, Hampton, and Kim 37). This yields a new function that, when put into a general linear model, can be regressed against fMRI data. General linear models have the following form:

$$y = B_0 + B_1 x_1 + B_2 x_2 + \ldots + B_n x_n + e$$

where y is the observed data, the x_i are regressors (the model-predicted time series), the B_i are variable weights (B_0 represents the contribution of factors held constant throughout the experiment), and e is residual noise in the data (Huettel, Song, and McCarthy 343). This allows researchers to identify brain areas where the model-predicted time series significantly correlates with the observed BOLD signal changes over time.

I should make clear that model-based fMRI has limitations and does not obviate the need for other neuroimaging methods (e.g., PET, EEG, or MEG). Like fMRI in general, model-based fMRI can only establish correlations between neural activity and behavior. In order to establish causal claims about neural activity and behavior, the same methods need to be used that were used before the introduction of model-based fMRI, such as lesioning and

transcranial magnetic stimulation (TMS) (O' Doherty, Hampton, and Kim 50). Like fMRI in general, model-based fMRI also has poor spatiotemporal resolution. This means that small computational signals such as those at the level of the single neuron will go undetected by model-based fMRI. For these reasons, a model-based approach to other neuroimaging methods is needed (Ibid.)

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 subjects were asked to classify it as "Hole A" or "Hole B," after which they received feedback. 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 might have thick legs and three out of four Hole B beetles could have thin legs. 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, during stimulus presentation SUSTAIN predicts a recognition advantage for exceptions but ALCOVE predicts no recognition advantage. This is called the recognition strength measure. This

difference in recognition strength measure 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, when subjects are given feedback, both SUSTAIN and ALCOVE predict that exceptions should lead to greater prediction error. This is called the error correction measure (Ibid., 263-4).

The results showed that the recognition strength measures and error correction measures predicted by SUSTAIN found correlations in MTL regions including bilateral hippocampus, parahippocampal cortex, and perirhinal cortex, and regions in bilateral hippocampus and perirhinal cortex, respectively. ALCOVE's predicted recognition strength measures did not find correlations in MTL, although its error correction predictions found correlations in MTL similar to SUSTAIN's (Ibid., 266-7). These results "suggest that, like SUSTAIN, the MTL contributes to category learning by forming specialized category representations appropriate for the learning context" (Davis, Love, and Preston 269). Furthermore, these correspondences to brain areas open a whole new range of opportunities for manipulation and provide answers to counterfactual questions that were not available before, thereby increasing the explanatory power of these models.

SUSTAIN and ALCOVE are mechanism-sketches. SUSTAIN is more how-actually than ALCOVE because both of SUSTAIN's prediction measures (recognition strength and error correction) were significantly correlated to areas of brain activation, whereas only one of ALCOVE's (error correction) was correlated. SUSTAIN, therefore, has more evidential support than ALCOVE. These results 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, Love, and Preston (2012) are at times quite explicit about the mechanistic nature of the models they are investigating, although they do not use the term "mechanistic." 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) and "The theory we forward relating SUSTAIN to the MTL…goes beyond the model's equations by tying model operations to brain regions" (270). Given their emphasis on mapping models to the brain, it is clear that they intend the models to be mechanistic. They are interested in more than the behavioral accuracy of these models. SUSTAIN and ALCOVE are already behaviorally well-confirmed, but model-based fMRI allowed Davis et al. to test their mechanistic accuracy**.**

6. Conclusion

Weiskopf (2011) presented three models of object recognition and categorization, JIM, ALCOVE, and SUSTAIN, that he claimed were non-mechanistic, yet explanatory. He argued that they were not mechanistic because their parts could not be neatly localized and they contained some components, such as Fast Enabling Links (FELs), which could not correspond to anything in the brain but are nevertheless essential for the proper working of the model. I argued on the contrary that these models are mechanism-sketches. In addition to

playing a black box role, FELs possess non-explanatory virtues such as suggesting new lines of investigation about feature linking and property inference.

My argument for the claim that SUSTAIN and ALCOVE are mechanism-sketches relied on model-based fMRI research. Model-based fMRI and other model-based neuroimaging approaches are beginning to allow cognitive neuroscientists to map psychological models onto the brain. Cognitive neuroscientists can then discriminate between equally behaviorally confirmed psychological 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 psychological models against neuroimaging data using model-based techniques, they will retain those models that find correspondences in the brain and reject those that do 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: 349-376.

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 E. Cooper, John E. 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-Verlag.

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 C. Love, and Alison R. 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 P. and John P. O' Doherty. 2010. “Model-based Approaches to Neuroimaging: Combining Reinforcement Learning Theory with fMRI Data.” *WIREs Cognitive Science* 1: 501-10.

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

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

Huettel, Scott A., Allen W. Song, and Gregory McCarthy. 2009. *Functional Magnetic Resonance Imaging*. Sunderland, Mass.: Sinauer Associates.

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

Kim, Jiye G. and Irving Biederman. 2012. “Greater sensitivity to nonaccidental than metric changes in the relations between simple shapes in the lateral occipital cortex.” *NeuroImage* 63: 1818-1826.

Kourtzi, Zoe and Kalanit Grill-Spector. 2005. “fMRI Adaptation: A Tool for Studying Visual Representations in the Primate Brain.” In *Fitting the Mind to the World: Adaptation and After-Effects in High-Level Vision*, ed. Colin W. G. Clifford and Gillian Rhodes, 173-88. New York: Oxford University Press.

Krekelberg, Bart, Geoffrey M. Boynton and Richard J.A. van Wezel. 2006. “Adaptation: From Single Cells to BOLD Signals.” *TRENDS in Neurosciences* 29.5: 250-56.

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

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

Love, Bradley C. and Todd M. Gureckis. 2007. “Models in Search of a Brain.”

O' Doherty, John P., 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 C. 1984. *Scientific Explanation and the Causal Structure of the World.* Princeton: Princeton University Press.

Salmon, Wesley C. 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 A. 2011. “Models and Mechanisms in Psychological Explanation.” *Synthese* 183.3: 313-38.