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: 4953

1. Introduction

A mechanistic explanation explains a phenomenon 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). 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. They are not mechanistic, Weiskopf argues, because their parts cannot be neatly localized and they sometimes contain components that could never correspond to anything in the brain.

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 Barberis (2013).

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 the crucial concept of a mechanism-sketch. I will also introduce the model-mechanism-mapping (3M) constraint, which I use as a criterion for when a model is mechanistic. In Section 3, I will show that one of the models of object recognition and categorization (JIM) that Weiskopf presents as non-mechanistic, yet explanatory, is actually a mechanism-sketch. In Section 4, I will introduce model-based fMRI, which is required for the argument of the next section. In Section 5, I will use recent model-based fMRI research to show that the two other models that Weiskopf presents as non-mechanistic, yet explanatory (SUSTAIN and ALCOVE), are also mechanism-sketches.

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]

Within the mechanistic framework there is an important distinction between complete mechanistic models and mechanism-sketches (Craver 2007). Mechanism-sketches are incomplete descriptions of 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. Idealization can be readily accommodated within a mechanistic framework.

[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.”

A way of distinguishing mechanistic from non-mechanistic explanation is needed before we can assess whether or not any particular explanation is mechanistic. An account that can provide an intuitive distinction between mechanistic and non-mechanistic explanation is given by Kaplan's model-mechanism-mapping (3M) constraint (2011). According to 3M:

> A model of a target phenomenon explains that phenomenon to the extent that (a) the variables in the model correspond[3] to identifiable components, activities, and organizational features of the target mechanism that produces, maintains, or underlies the phenomenon, and (b) the (perhaps mathematical) dependencies posited among these (perhaps mathematical) variables in the model correspond to causal relations among the components of the target mechanism. (Kaplan 2011, 347)

We wanted some criteria to describe when exactly an explanation counts as mechanistic. 3M appears to give us something different – criteria which describe when a model *explains*. I think it is best to read 3M as providing criteria *both* for when a model counts as explanatory and for when it counts as mechanistic. 3M can be read as saying that a model of a target

[3] Batterman and Rice (2014) object that "correspondence" is left unanalyzed by mechanists. For the purposes of this paper, how exactly a model corresponds to reality will be clear. For example, in my model-based fMRI example correspondence between model and world is cashed out in terms of correlation between model-predicted and observed hemodynamic response functions.

phenomenon explains that phenomenon to the extent that the model is mechanistic and the model is mechanistic to the extent that (a) and (b) are satisfied. It is a more formal way of saying in model terms that an explanation is mechanistic to the extent that it accurately represents the mechanism underlying the explanandum. I will use 3M *only* to provide criteria that describe when a model counts as mechanistic. I do not want to say that only mechanistic models are explanatory (or that the only genuine explanations are mechanistic). As I mentioned above, there are other ways of situating an explanandum phenomenon in the causal structure of the world that are not mechanistic.[4]

On the 3M constraint, the continuum from mechanism-sketch to complete mechanistic model runs parallel to the continuum from less to more correspondence between the model's variables and dependencies and the world's (perhaps distributed) parts and causal relations.

3. JIM as a Mechanism-Sketch

In this section I examine a model of object recognition and categorization, JIM, that Weiskopf (2011) takes to be non-mechanistic, yet explanatory. I will argue that, given the 3M constraint, JIM is in fact a mechanism-sketch. The other two models, SUSTAIN and ALCOVE, will be examined later in Section 5 after introducing model-based fMRI. I delay

[4] I do not mean to imply that Kaplan would disagree. He presented the 3M constraint in the context of computational models in neuroscience, so he was not concerned with other forms of causal explanation.

further examination of SUSTAIN and ALCOVE until after I introduce model-based fMRI because both of these models have been investigated using model-based fMRI and I will draw on those investigations to support my claim that SUSTAIN and ALCOVE are mechanism-sketches.

According to JIM, 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.

To determine whether or not this model is a mechanism-sketch, we have to look at the extent to which it satisfies the 3M constraint. If the model is a mechanism-sketch, the systems and processes in the model required for extraction of geon structure, storage of geon structure, and comparison of geon structures must to some extent correspond to (perhaps distributed) components in the actual object recognition mechanism(s) in the brain. Is this the case?

This model was built, not merely to produce the same behavior as humans in object recognition tasks, but to model something that happens 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, Biederman and others have conducted various neuroimaging studies to investigate the neural underpinnings of the model.

If Biederman's model is correct, there is an area or 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).

It is true that 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. Right now they 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?”.

4. 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).[5] Given the 3M constraint, model-based fMRI can help us demarcate mechanistic from non-mechanistic models.

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 this neuroimaging method. 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

[5] Weiskopf (2011, 334) worries that allowing mechanisms to include distributed components will make identifying them more difficult. Model-based neuroimaging techniques allay this worry.

internal variables linking input and output. The neural correlates of these internal variables can then be located using regression analyses (O' Doherty, Hampton, and Kim 36).

I will illustrate how this works using the Rescorla-Wagner model of classical conditioning. In classical conditioning, an unconditioned stimulus, such as food, elicits an unconditioned response, such as salivation. After a neutral stimulus, such as a bell, becomes correlated[6] with the unconditioned stimulus, the neutral stimulus (now called the conditioned stimulus) elicits a response (called the conditioned response) similar to the unconditioned response.

The Rescorla-Wagner model of classical conditioning can be represented with the following equations:

$$\mathrm{V}_{n+1} = \mathrm{V}_{n} + ad$$

$$d = u - v$$

where *v* represents the value of the conditioned stimulus and *u* represents the value of the unconditioned stimulus. Conditioning occurs as the value of *v* converges toward the value of *u*. On the Rescorla-Wagner model, this occurs as the value of *v* is updated in proportion to a prediction error *d* (the difference between *u* and *v*) on each trial, with *a* being the learning

[6] This requires the caveat that the conditioned response (CR) is not already conditioned to a different conditioned stimulus (CS) that is present. Otherwise, conditioning of the CR to the first stimulus will be “blocked” by the second CS. For example, if a light is already a CS that elicits salivation, pairing a bell with the light will not make the bell a CS.

rate used to scale the updates to *v*. The value of *a* is calculated by simply finding the best fit to the behavioral data (O' Doherty, Hampton, and Kim 37).

The two variables that change from trial to trial, *v* and *d*, 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. This just means that the predicted values of *v* and *d*, 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 which, when put into a general linear model, can be regressed against fMRI data. General linear models have the following form:

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

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.)

5. SUSTAIN and ALCOVE as Mechanism-Sketches

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 model-based fMRI research. 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, the next layer consists of a set of clusters associated with a category. All the 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.

Both models were investigated in a model-based fMRI study (Davis, Love, and Preston 2012). In this study, participants completed a rule-plus-exception category learning task. 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 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.

These results show that these SUSTAIN and ALCOVE are mechanism-sketches because they have varying degrees of correspondence to brain mechanisms. SUSTAIN is less sketchy than ALCOVE because both of SUSTAIN’s prediction measures were significantly correlated to areas of brain activation, whereas only one of ALCOVE’s was correlated. SUSTAIN, therefore, gets to move slightly up the continuum from mechanism-sketch to complete mechanistic model. 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, “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. Of course, this does not show that the models are in fact mechanistic. Modelers can intend that the model variables and relations between them correspond to features of the world and be wrong. Similarly, variables that are not intended to correspond to anything in the world could turn out to correspond. The model-based fMRI results presented above indicate that SUSTAIN and ALCOVE are, in fact, sketches of mechanisms, though sketchy to different degrees.

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. 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

Barberis, Sergio Daniel. 2013. “Functional Analyses, Mechanistic Explanations, and Explanatory Tradeoffs.” *Journal of Cognitive Science* 14.3: 229-51.

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

Kaplan, David M. 2011. “Explanation and Description in Computational Neuroscience.” *Synthese* 183: 339-73.

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.

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.