Models and mechanisms in psychological explanation

Daniel A. Weiskopf

Abstract: Mechanistic explanation has an impressive track record of advancing our understanding of complex, hierarchically organized physical systems, particularly biological and neural systems. But not every complex system can be understood mechanistically. Psychological capacities are often understood by providing cognitive models of the systems that underlie them. I argue that these models, while superficially similar to mechanistic models, in fact have a substantially more complex relation to the real underlying system. They are typically constructed using a range of techniques for abstracting the functional properties of the system, which may not coincide with its mechanistic organization. I describe these techniques and show that despite being non-mechanistic, these cognitive models can satisfy the normative constraints on good explanations.

1. Introduction

We are in the midst of a mania for mechanisms. In the wake of the collapse of the deductive-nomological account of explanation, philosophers of science have cast about for alternative ways of describing the structure of actual explanations in science and the normative properties that good explanations ought to have. Mechanisms and mechanistic explanation have promised to fill both of these roles, particularly in the 'fragile sciences' (Wilson, 2004): biology (Bechtel, 2006; Bechtel & Abrahamson, 2005), neuroscience (Craver, 2006, 2007), and, increasingly, psychology (Bechtel, 2008, 2009; Glennan, 2005). Besides these benefits, mechanisms have also promised to illuminate other problematic scientific notions such as capacities, causation, and causal laws (Glennan, 1996, 1997; Machamer, 2004, Woodward, 2002).

Mechanistic explanation involves isolating a set of phenomena and positing a mechanism that is capable of producing those phenomena (see Craver & Bechtel, 2006 for a capsule description). The phenomena in question are an entity or a system's exercising a certain capacity:

an insect's ability to dead reckon, my ability to tell that this is a lime, a neuron's capacity to produce an action potential, a plant's capacity for photosynthesis. What one explains mechanistically, then, is S's ability, propensity, or capacity to F. The mechanism that does the explaining is composed of some set of *entities*—the components of the mechanism—and their associated *activities* that are organized in such a way as to produce the phenomena. Mechanistic explanation involves constructing a model of such mechanisms that correctly depicts the causal interactions among their parts that enable them to produce the phenomena under various conditions.[1] Such a model should specify, among other things, the initial and termination conditions for the mechanism, how it behaves under various sorts of interventions, including abnormal inputs and internal disruptions, how it is integrated with its environment, and so on.

There is no doubt that explanation in biology and neuroscience often involves describing mechanisms. Here I'm particularly concerned with whether the mechanistic revolution should be extended to psychological explanation. A great deal of explanation in psychology involves giving models of various psychological phenomena.[2] These models can be formal (e.g., mathematical or computational) or they may be more informally presented. It can be extremely tempting to cast these models in the mold of mechanistic explanation. I'll argue that we should not succumb to this temptation, and that cognitive models are not, in general, models of mechanisms. While they have some features in common with mechanistic models, they differ significantly in the way that they relate to the underlying system whose structure they aim to

[1] I will make heavy use of the term 'model' throughout this discussion. However, I do not have any very specific conception of models in mind. What I will mean is at least the following. A model is a kind of representation of some aspect of the world. The components of models are organized entities, processes, activities, and structures that can somehow be related to such things in the real world. Models can be picked out linguistically, visuospatially, graphically, mathematically, computationally, and no doubt in many other ways. This should be a sufficiently precise conception for present purposes.

[2] See, e.g., the papers collected in Polk & Seifert (2002). For a comprehensive history of the construction of computational models of cognition, see Boden (2006).

represent. Despite this, they can be evaluated according to the standard norms that govern model construction generally, and can provide perfectly good explanations of psychological phenomena.

In the discussion to follow, I first lay out the criteria by which good models of real-world systems are to be assessed (Section 2). My starting point is Carl Craver's discussion of the norms of mechanistic explanation, which I propose should be generalized to cover other types of model-based reasoning. I then show that one type of non-mechanistic explanation, Rob Cummins' analytic functional explanations, can meet these norms despite being entirely noncomponential (Section 3). I describe several different cognitive models of psychological capacities such as object recognition and categorization (Section 4), and I show that despite being non-mechanistic, these models can also meet the normative standards for explanations (Section 5). Finally, I rebuff several attempts to either reduce these modeling strategies to some sort of mechanistic approach, or to undermine their explanatory power (Section 6).

2. Three dimensions of model assessment

In laying out the criteria for a good mechanistic explanation, Craver (2007) usefully distinguishes between two dimensions of normative evaluation that we can use in assessing these explanations. He distinguishes: (1) how-possibly, how-plausibly, and how-actually models; and (2) mechanism sketches, mechanism schemata, and complete mechanistic models. Here I will lay out what I take to be the most useful way to construe these distinctions.

Consider the first dimension, in particular the end that centers on how-possibly (HP) models. HP models are "loosely constrained conjectures" about what sort of mechanism might produce the phenomenon (Craver, 2007, p. 112). In giving these, one posits parts and operations,

but one need not have any idea if they are real parts, or whether they could do what they are posited to do. Examples here include much early work in symbolic computational simulation of cognitive capacities. Computer simulations of vision written in high-level programming languages describe a set of entities (symbolic representations) and activities (concatenation, comparison, etc.) that may produce some fragment of the relevant phenomena, but one need not know or be committed to the idea that the human visual system contains those parts and operations. Similar critiques have been made of linguists' syntactic theories: the formal sequence of operations posited in generative grammars—from transformational rules to 'Move α'—is often psychologically hard to detect.[3] How-actually (HA) models, on the other hand, "describe real components, activities, and organizational features" of the mechanism (Craver, 2007, p. 112). In between these are how-plausibly models that vary in their degree of realism.

Clearly, whether a model is nearer to the HP or HA end is not something that can be determined just by looking at its intrinsic structure. This continuum or set of distinctions turns on degrees of *evidential support*. To see this, notice that producing HP models is often part of the early stages of investigating a mechanism. This initial set of models is then evaluated to determine which one best explains the phenomena or best fits the data, if any of them do. Some rise to the level of plausibility, and eventually we may settle on our best-confirmed hypothesis as to which is the actual mechanism.

That the distinction is epistemic is suggested by the way in which models move along this dimension.[4] If we are just considering a set of models to account for a phenomenon, then we can regard them all as how-possibly, or merely conjectural. If we have some evidence that favors

[3] There may be ways to detect the presence of representations such as traces or phonologically empty categories such as PRO by comparing speakers' grammaticality judgments across pairs that differ in these hypothesized elements. But clusters of converging operations focused on these elements are difficult to come by.
[4] This is also stated explicitly in Machamer, Darden, & Craver (2000), pp. 21-2.

one or two of them, or some set of constraints that rule some of them out, they move into the how-plausibly column. This implies that a how-actually model is one that *best* accommodates all of the evidence and constraints. Terminologically, this might seem uncomfortable: whether a model captures how the mechanism *actually* is doesn't seem like a matter of evidential support, but a matter of how accurately it models the system in question. This makes it sound as if a how-actually model is just the true or accurate model of the system. But note that any one of a set of how-possibly models might turn out to accurately model the system, so the difference in how they are placed along this dimension cannot just be in terms of accuracy. So it seems that this is fundamentally an epistemic dimension. It represents something like the degree of confirmation of the claim that the model corresponds to the mechanism. Even if your model is *in fact* the one that accurately represents the mechanism in question, if you take it to be merely a guess or one possibility among many, then it's a how-possibly or how-plausibly model. More evidence that this is how the mechanism works makes it inch towards being how-actually.

The second dimension of assessment involves the continuum from mechanism sketches to mechanism schemata and complete mechanistic models. A sketch is an "incomplete model of a mechanism", or one that leaves various gaps or employs filler terms for entities and processes whose nature and functioning is unknown. These terms—'control', 'influence', 'regulate', 'process', etc.—constitute promissory notes to be cashed in by further analysis. A schema is a somewhat complete, but less than *ideally* complete, model. It may contain black boxes or other dummy items, but it incorporates more informative detail than a mere sketch. Finally an ideally complete model omits nothing, or nothing relevant to understanding the mechanism and its operations in the present context, and uses no terms that can be regarded as 'filler'.

The continuum from sketches to schemata and complete models is not epistemic. Rather it has to do with representational *accuracy*—a term which, as I use it, incorporates both grain and correctness.[5] Correctness requires that the model not include elements that are not present in the system, nor omit elements that are present. Grain has to do with the size of the chunks into which one decomposes a mechanism. This is a matter of varying degrees of precision. For example, the hippocampus can be seen as a three-part entity composed of CA1, CA3 and the dentate gyrus, or it can be seen as a more complex structure containing various cell types, layers, and their projections, etc. (Craver, 2009). But there can be coarse-grained but correct models, as this example shows. Coarse-grained models merely suppress further mechanistic details concerning their components. Presumably this would be an instance of a schema or sketch. I take it that approaching ideal accuracy involves achieving a more correct model (one that includes more of the relevant structure of the system) and also a more fine-grained model (one that achieves greater precision in its depiction of the system).

One question about this distinction is what sorts of failures of accuracy qualify a model as a 'sketch'. Every model omits something from what it represents, for instance, but not every way of omitting material seems to make for a sketch. For example, one way to omit is just not to include some component that exists in the real system. Sometimes this is innocuous, since the component may not be relevant to understanding the mechanism in the current explanatory context. Many intracellular structures are omitted in modeling neurotransmitter release, for instance. But this can also be a way of having a false or harmfully incomplete model. Alternatively one can include a 'filler' term that is known to abbreviate something about the

[5] Giere (1988) also separates accuracy into two components: similarity in certain respects and accuracy to various degrees in each of these respects. My own grain/correctness distinction does not quite correspond to his, but both illustrate the fact that we need to make more distinctions than are allowed by just the notion of undifferentiated 'accuracy' that seems to underlie the schema-sketch continuum.

system that we cannot (yet) describe any better. The question then arises what sort of relationship terms and components of models must bear to the underlying system for the model to be a good representation of the system's parts and organization. In particular, it might be that there are components of an empirically validated model that do not map onto any parts of the modeled system. I will discuss some examples of this in Section 5.

Some of these accuracy failures are harmful and others are not. It seems permissible to omit detail where it's irrelevant to our modeling purposes, so being a schema is not in and of itself a bad thing. Moreover, complete models may be intractable in practice in various ways. The larger point to notice is that the simple notion of a sketch/schema continuum runs together the notion of something's being a *false* model and its being a merely *detail-free* model. The most significant failures of models seem to arise from either including components that do not correspond to real parts of the system, or omitting real parts of the system in the model (and, correspondingly, failing to get the operations of those parts correct). These failures are lumped in with the more innocuous practices of abstracting and omitting irrelevant details in the notion of a sketch or a schema.

A third way of classifying models is with respect to whether or not they are genuinely explanatory, as mechanistic models are assumed to be. Craver (2006) draws a separate normative distinction between merely phenomenological models and genuine explanations. Phenomenological accuracy is simply capturing what the phenomena are. An example of this is the Hodgkin-Huxley equation describing the relationship between voltage and conductance for each ion channel in the cell membrane of a neuron. These may be useful for predicting and describing a system, but they do not provide explanations.[6] One possibility that Craver considers is that explanatory models are "much more useful than merely phenomenal models for the

[6] See Bokulich (forthcoming) for further discussion of explanatory versus phenomenological and fictional models.

purposes of control and manipulation" (2006, p. 358). Deeper explanations involve being able to say how things would have been otherwise, how the system would be if various perturbations occurred, how to answer a greater range of questions about the system, etc.

Here we should separate the properties of *allowing control and manipulation* from *being able to answer counterfactual questions*. Many good explanations do the latter but not the former. Our explanation for why a gaseous disc around a black hole behaves like a viscous fluid does not enable us to control or manipulate that disc in any way, nor do our explanations of how stellar fusion produces neutrinos. Many apparently phenomenological models can also describe certain sorts of counterfactual scenarios. Even the Hodgkin-Huxley model allows us to make certain counterfactual predictions about how action potentials will perform in various perturbed circumstances. But they are silent on other counterfactuals, particularly those having to do with interventions involving the system's operations. So we can still say in general that models become more explanatory the more they allow us to answer a range of counterfactual questions and the more they allow us to manipulate a system's behavior (in principle at least). This sort of normative assessment is also neutral on the question of whether the explanations in question are mechanistic or not.

What emerges, then, is a classification of models according to (1) whether they are highly confirmed or supported by the evidence and (2) whether they are representationally accurate. So stated, these dimensions of assessment are independent of whether the model is mechanistic. We can ask whether any theory, model, simulation, or other representational device conforms to the norms of accuracy and confirmation. In addition, models may be classified according to (3) whether they are genuinely explanatory, or merely phenomenological, predictive, or descriptive. Finally, there are broad requirements that models cohere with the rest of what we know. Thus we

can also assess models with respect to (4) whether they are consistent with and are plausible in light of our general background knowledge and our more local knowledge of the domain as a whole.

3. Noncomponential analysis

Mechanistic models, or many of them, can meet these normative conditions. Strictly descriptive-phenomenological models cannot. But there are effective explanatory strategies besides mechanistic explanation. Cummins (1983) argues that a kind of analytic functional explanation plays a central role in psychology. As with mechanistic explanation, the explanandum phenomenon is the fact that a system S has a capacity to F. In his account, S's capacity to F is analyzed into various further capacities $G_1,\ldots,G_n$, all of which also belong to S itself. F-ing, then, is explained as having the appropriately organized (spatially and temporally choreographed) capacities to carry out certain other operations whose exercise constitutes F-ing. This is a kind of analytic explanation, since it aims to explain one capacity by analyzing it into subcapacities. However, these are not capacities of subparts of the system. The account doesn't explain S's F-ing in terms of the G-ing of S's parts, but rather in terms of the activity of S itself. Cummins calls this *functional* analysis; it involves "analyzing a disposition into a number of less problematic dispositions such that programmed manifestation of these analyzing dispositions amounts to a manifestation of the analyzed disposition" (1983, p. 28).

In many cases, this analysis will be of one disposition of a subject or system into other dispositions of the same subject or system. In such cases, the "analysis seems to put no constraints at all on [the system's] componential analysis" (1983, p. 30). As an example, Cummins gives an analysis of the disposition to see an afterimage as shrinking if one approaches

it while it is projected onto a visible wall. This is analyzed in terms of a flowchart or program specifying the relations among various subdispositions that need to be present: the ability to determine whether an object is visually present, to determine the size of the retinal image and distance to the object, to use these to compute the apparent object size (Cummins, 1983, pp. 83-7). He offers analogous functional analyses of grammatical competence, Hull's account of conditioning, and Freudian psychodynamics.

Craver seems to reject the explanatory significance of functional analysis of this sort—call it *noncomponential analysis*, or NCA. By contrast with NCA, mechanistic explanation is "inherently componential" (2007, p. 131). From the mechanistic point of view, NCA essentially faces a dilemma. One possibility is that without appeal to components and their activities, we have no way to distinguish how-possibly from how-actually explanations, and sketches from more complete mechanistic models. In other words, NCA blocks us from making crucially important distinctions in kinds of explanations. Without some way of making these or analogous distinctions we have no way of distinguishing good explanations from non-explanations. So box-and-arrow models that do not correspond to real components are doomed to be either how-possibly or merely phenomenological models, not mechanistic models.

We can put this in the form of an argument:

1. Analytic functional explanations are noncomponential.
2. Noncomponential explanations provide only a redescription of the phenomenon or a how-possibly model.
3. Redescriptions and how-possibly models are not explanatory.
4. So analytic functional explanations are not explanatory.

The argument is valid. Premise 1 is true by definition of analytic models (at least those that are not linked with an instantiation theory). With respect to premise 3, we can agree that redescriptive models and some how-possibly models are not explanatory.[7] But the question here centers on premise 2. The issue is whether there could be genuinely explanatory but non-mechanistic and non-phenomenological models—in particular, in psychology.

Returning to our characterization of these distinctions above, we can ask whether NCA models can be assessed along our first two normative dimensions. Are we somehow blocked from *confirming* NCA models? Evidently not. We might posit one decomposition of a capacity into subcapacities only to find that, empirically, this is not the decomposition that individuals' exercise of C involves. In fact, even Cummins makes this one of his desiderata: it is a "requirement that attributions of analyzing properties should be justifiable independently of the analysis that features them" (1983, pp. 26-7). If we analyze a child's ability to divide into capacities to copy numbers, multiply, add, etc., we need separate evidence of those capacities to back this attribution. If, as I have suggested, we conceive of moving from a how-possibly model to a how-actually model as acquiring more and stronger evidence in favor of one model over the others, we can see getting this sort of confirmation for an analysis as homing in on a how-actually functional analysis.

So we can distinguish how-possibly NCA models from how-actually NCA models. Similarly, we can ask whether this NCA model *accurately* represents the subcapacities that a creature possesses, whether it does so in great detail or little detail, etc. That we can do this is evident from the fact that the attributed capacities themselves can be fine-grained or coarse-

[7] The caveat concerns contexts in which we may want to say that a how-possibly account *is* a sufficient explanation. In explaining multiple realization, for example, we explicitly consider a range of how-possibly models and treat these as explaining the fact that a capacity is displayed by physically disparate systems. See Weiskopf (forthcoming) for discussion.

grained, can be organized in different ways to produce their output, can contain different subcapacities nested within them, and so on. Consider two different ways of analyzing an image manipulation capacity: as allowing image rotation to step through 2 degrees at a time versus 5 degrees at a time; or as requiring that rotations be performed before translations in a plane, rather than the reverse; and so on. These ways of filling in the same 'black boxed' capacity correspond to the functional analytic difference between sketches, schemata, and complete models. We can, then, assess NCA models for both epistemic confirmation and for accuracy and granularity.

But Craver seems to think that NCA models can't make these distinctions, and he pins this fact on their being noncomponential (2007, p. 131):

> Box-and-arrow diagrams can depict a program that transforms relevant inputs onto relevant outputs, but if the boxes and arrows do not correspond to component entities and activities, one is providing a redescription of the phenomenon (such as the HH model of conductance change) or a how-possibly model, not a mechanistic explanation.

The way we distinguish HP from HA and sketches from schemata, etc., is by positing components and activities. Thus these facts militate in favor of mechanistic models. Call this the *Real Components Constraint* (RCC) on mechanistic models: the components described in the model should be real components in the mechanism. This is a specific instantiation of the principle that models are explanatory to the extent that they correspond with real structures. The constraint can be seen to flow from the general idea that models are better to the extent that they are accurate and complete within the explanatory demands of the context. The difference is that the focus of this principle is on components and their operations or activities rather than on accuracy in general. I discuss the role of the RCC in distinguishing good explanations from merely phenomenal accounts in Section 6.

A final objection that Craver levels against NCAs is that they do not offer *unique* explanations of cognitive capacities, and hence must only be giving us how-possibly explanations. Cummins seems to suggest this at times; for example, he says (p. 43):

> Any way of interpreting the transactions causally mediating the input-output connection as steps in a program for doing φ will, provided it is systematic and not ad hoc, make the capacity to do φ intelligible. Alternative interpretations, provided they are possible, are not competitors; hence the availability of one in no way undermines the explanatory force of another.

This appears to mean that for any system S there will be many equally good explanations for how it is able to do F, which in turn suggests that these explanations are merely how-possibly—since any how-actually explanation would necessarily have to be unique.

In fact, I don't think that we *should* presuppose that there is a unique how-actually answer to how a system carries out any of its functions. But this point aside, I think the charge rests on a misunderstanding of how the type of explanation Cummins is concerned with here works. Here he is addressing what he calls *interpretive functional analysis*. This is specifically the attempt to understand the functional organization of a system in semantically interpreted terms—not merely to describe what the system does as, e.g., opening and closing gates and relays, but as adding numbers or computing trajectories. Interpretive analysis differs from descriptive analysis precisely in its appeal to such *semantic* properties (1983, p. 34).

The point that Cummins wants to make about interpretive analyses is that for any system there may be many distinct yet still *true* explanations of what a system is doing when it exercises the capacity to F. But this fact does not suggest that the set of explanations is *merely* a how-possibly set. Consider the case of grammatical competence. There may be many predictively

equivalent yet interestingly distinct grammars underlying natural language. As Cummins notes, however, "[p]redictively adequate grammars that are not instantiated are, of course, not explanations of linguistic capacities" (p. 44). Here we may see a role for how-possibly explanations; functional analysis can result in programs that are not instantiated, and figuring out what grammar is instantiated is part of telling the how-actually story for linguistic competence. Even if a system instantiates a grammar, though, there may be other grammars that it *also* instantiates. And this may be the case even once we pin down the details of its internal structure. A decomposition of a system into components does not necessarily, in his view, uniquely fix the semantic interpretation of the components or the system that they are part of: "[i]f the structure is interpretable as a grammar, it may be interpretable as another one too" (p. 44).

Cummins' NCA-style explanations allow for structural constraints to play a role in getting a how-actually story from a how-possibly story about interpretive functional analysis. The residual multiplicity of analyses comes from the fact that these facts do not pin down a unique semantic interpretation of the system. Hence the same system may be instantiating many programs, all of which are equally good explanations of what it does. We needn't follow Cummins in thinking that it's indeterminate or pluralistic which program a system is executing; that's an idiosyncrasy of his view concerning semantic interpretation. If there are facts that can decide between these interpretive hypotheses, then NCA models can be assessed on our two normative dimensions despite not being mechanistic. Whether this is so depends ultimately on whether there can be an account of the fixation of semantic content that selects one interpretation over another, a question that is definitely beyond the scope of the discussion here.

There is a larger moral here which will serve to introduce the theme of our next sections. In trying to understand the behavior of complex systems, we can adopt different strategies. For

neurobiological systems and phenomena, it might be that compositional analysis is an obvious first step: figuring out the anatomical, morphological, and physiological properties of cells in a region, their laminar and connectional organization, their response profile to stimulation, the results of lesioning or inhibiting them, etc. But for psychological phenomena, giving an account of what is involved in their production is necessarily more indirect. It is plausible that in many cases, decomposing the target capacity into subcapacities is heuristically indispensible—if for no other reason than, often enough, we have no well-demarcated physical system to decompose, and little idea of the proper parts and operations to use in such a decomposition. The only system we can analyze is the central nervous system, and its structure is notoriously not psychologically transparent. Thus (as Cummins also argues) structural hypotheses usually *follow* interpretive functional hypotheses: we indirectly specify the structure in question by making a provisional analysis of the psychological capacity, then we look for 'fits' in the structure that can be used to interpret it. These fits obtain between the subcapacities and their flowchart relations and parts of the physical structure and their interconnections. The question, then, is how models in psychology actually operate; in particular, whether their functioning can be wedged into the mold of mechanistic explanation. In the next section I'll lay out a few such models and argue that, as with functional analysis, these are cases of perfectly good explanations that are not mechanistic.

4. The structure of cognitive models

The models I will consider are all psychological models, in the sense that they aim to explain psychological phenomena and capacities. They are models *of* parts of our psychology. They are also psychological in another sense: like interpretive functional analyses, they explain

these capacities in terms of semantic, intentional, or more generally representational states and processes. There can be models of psychological phenomena that are not psychological in this second sense. Examples are purely neurobiological models of memory encoding or attention. These explain psychological phenomena in non-psychological terms. While some models of psychological capacities employ full-blown intentional states such as beliefs, intentions, and desires (think of Freudian psychodynamic explanations and many parts of social psychology), others posit more theoretically-motivated subpersonal representational states. Constructing models of this sort is characteristic of cognitive psychology and many of its allied fields such as cognitive neuroscience and neuropsychology. Indeed, the very idea of there being a 'cognitive level' of description was introduced by appeal to explanations having just this form. I will therefore refer to models that explain psychological capacities in representational terms as *cognitive models*.[8]

To be more specific, I will focus on *representation-process-resource* models. These are models of psychological capacities that aim to explain them in terms of systems of representations, processes that operate over and transform those representations, and resources that are accessed by these processes as they carry out their operations. Specifying such a model involves specifying the set of representations (primitive and complex) that the system can employ, the relevant stock of operations, and the relevant resources available and how they interact with the operations. It also requires showing how they are organized to take the system from its inputs to its outputs in a way that implements the appropriate capacity. This involves describing at least some of the architecture of the system: how information flows through it,

[8] To be sure, in recent years there have been a number of movements in cognitive science that have proposed doing away with representational models and their attendant assumptions. These include Gibsonian versions of perceptual psychology, dynamical systems theory, behavior-based robotics, and so on. I will set these challengers aside here to focus on the properties of models that are based on the core principles of the cognitive revolution.

whether its operations are serial or parallel, whether it contains subsystems that have restricted access to the rest of the information and processes in the system, and the control structures that determine how these elements work together to mediate the input-output transitions.

Thus a cognitive model can be seen as an organized set of elements that depicts how the system takes input representations into output representations in accord with its available processes and operations, as constrained by its available resources. In what follows I briefly describe three models of object recognition and categorization to highlight the features they have in common with mechanistic models and those that set them apart.

The first model comes from studies of human object recognition. Object recognition is the capacity to judge that a (usually visually) perceived object is either the same particular one that was perceived earlier, or belongs to the same familiar class as one perceived earlier. This recognitional capacity is robust across perspectives and other viewing conditions. One doesn't have the capacity to recognize manatees, Boeing 747's, or Rodin's 'Thinker' unless one can recognize them from a variety of angles, distances, lighting, degrees of occlusion, etc. Sticking to visual object recognition, the relevant capacity takes a visual representation of the object as input and produces as output a decision as to whether the object is recognized or not, and if it is, what it is taken to be. There are many competing models of object recognition, and my goal here is to present one representative model rather than to survey all of them.[9]

The model in question, presented in Hummel & Biederman (1992), is dubbed JIM (*J*ohn and *I*rv's *M*odel). It draws on assumptions about object recognition developed in earlier work by Biederman (1987). Essentially, Biederman hypothesized that object recognition depends on a set of abstract visual primitives called *geons*. These geons are simple three-dimensional shapes that come in a range of shapes such as blocks, cylinders, and cones, and can be scaled, rotated,

[9] For a range of perspectives, see Biederman (1995), Tarr (2002), Tarr & Bülthoff (1998), and Ullman (1996).

conjoined, and otherwise modified to represent the large-scale structure of perceived objects (minus details like color, texture, etc.). Perceived objects are parsed in terms of this underlying geon structure, which is then stored in memory for comparison to new views. Since geons are three-dimensional, they provide a viewpoint independent representation of an object's spatial properties. There need, therefore, to be perceptual systems that can extract this common underlying structure despite degraded and imperfect viewing conditions, in addition to systems that will determine when a match in geon structure is 'good enough' to count as the same object (same type or same token).

In JIM this process is decomposed into a sequence of subprocesses each of which takes place in a separate layer (L1–L7). L1 is a simplified retina-like structure that represents the object from a viewpoint as a simple line drawing composed of *edges*; these can be extracted from, e.g., luminance discontinuities in the ambient light. L2 contains a set of three distinct networks each of which extracts a separate type of feature: vertices (points where multiple edges meet), axes of symmetry, and blobs (coarsely defined filled regions of space). L3 is decomposed into a set of attribute representations: axis shape (straight vs. curved), size (large to small), cross-sectional shape (straight vs. curved), orientation (vertical, diagonal, horizontal), aspect ratio (elongated to flat), etc. These attributes can be uniquely extracted from vertex, axis, and blob information. Each of them takes a unique value, and a set of active values on all attributes uniquely defines a geon; the set of all active values across all attributes at a time uniquely defines all of the geons present in a scene. L4 and L5 take their input from the L3 attributes having to do with size, orientation, and position, and they represent the relations among the geons in a scene, e.g., whether they are above, beside, or below one another. L6 is an array of individual cells each of which represents a single geon and its relations to the other geons in the scene (a geon feature

assembly), as determined by the information extracted by L3 and L5; finally, L7 represents the network's best guess as to what the object is, arrived at on the basis of the summed activity over time in the geon feature assembly layer.

The second two models come from work on concepts and categorization. Categorization and object recognition are related but distinct tasks.[10] In categorizing, one takes some information about an object—perceptual, functional, historical, contextual/ecological, theoretical/causal, etc.—and comes to a judgment about what sort of thing it is. A furry animal with pointy ears that meows is likely a cat; a strong, odorless alcoholic beverage is likely vodka; a meal bought at a fast food restaurant is likely junk food; and so on. Like object recognition, categorization can be viewed a kind of inference from evidence. But categorization can draw on a wider range of evidence than merely perceptual qualities (politicians are defined by their role, antique tables are defined by their historical properties, etc.), since concepts are representations that can group things together in ways that cross-cut their merely perceptual similarities.

As in the case of object recognition, there are far too many different models of categorization to survey here.[11] I will focus on two models that share some common assumptions and structure: the ALCOVE model (Kruschke, 1992), and the SUSTAIN model (Love, Medin, & Gureckis, 2004; Love & Gureckis, 2007). What these models have in common is that they explain how we categorize as a process of comparing new stimuli to stored *exemplars* (representations of individual category members) in memory. The similarity between the

[10] To some extent, the differences between the two may reflect not deep, underlying differences in their cognitive structure, but rather differences in the assumptions and methods of the experimental community that has investigated each one. What is called 'perceptual categorization' and object recognition may in fact be two names for the same capacity (or at least partially overlapping capacities). But the literatures on the two problems are, so far at least, substantially distinct.

[11] For a review of theories of concepts and the phenomena they aim to capture generally, see Murphy (2002). For a review of early work in exemplar theories of categorization, see Estes (1994). For a more recent review, see Kruschke (2008).

stimulus and the stored exemplars determines which ones it will be classified with. Both of these can be regarded as descendents of the Generalized Context Model (Nosofsky, 1986). The GCM assumes that all exemplars (all of the instances of a category that I have encountered and can remember) are represented by points in a large multidimensional space, where the dimensions of the space correspond to various attributes that the exemplars can have (size, color, animacy, having eyes, etc.). Each exemplar is a measurable distance in this space from every other exemplar. The psychological *similarity* between two exemplars is a function of their distance in the space.[12] Finally, categorization decisions for new stimuli are made on the basis of how similar the new stimulus is to exemplars stored in memory. If the stimulus is more similar to members of one category than another, then it will be categorized with those. Since there will usually be several possible alternatives, this is expressed as the probability of assigning a stimulus *s* to a category *C*.

The GCM gives us a set of equations relating distance, similarity, and categorization. ALCOVE (*A*ttention *L*earning *COVE*ring map) is a cognitive model that instantiates these equations. It is a feed-forward network that takes a representation of a stimulus and maps it onto a representation of a category. The input layer of the network is a set of nodes corresponding to each possible psychologically relevant dimension that a stimulus can have—that is, each property that can be encoded and stored for use in distinguishing that object from every other object. The greater the 'value' of this dimension for the stimulus, the more strongly the corresponding node is activated. The activity of these nodes is modulated by an attentional gate, which corresponds to how important that dimension is in the present categorization task, or for that type of stimulus. The values of these gates for each dimension may change across items or

[12] There are various candidate rules for computing similarity as a function of distance. Nosofsky's original rule took similarity to be a decaying exponential function of distance from an exemplar, but the details won't concern us here. The same goes for the rules determining how categorization probabilities are derived from similarities.

tasks—sometimes color is more important, sometimes shape, etc. The resulting modulated activity for the stimulus is then passed to the stored exemplar layer. In this layer, each exemplar is represented by a node, and the activity level of these nodes at a particular stage in processing is determined by their similarity to the stimulus. So highly similar exemplars will be strongly activated, moderately similar ones less so, and so on. Finally, the exemplar layer is connected to a set of nodes representing categories (*cat*, *table*, *politician*, etc.). The strength of the activated exemplars determines the activity level of the category nodes, with the most strongly activated node corresponding to the system's 'decision' about how the stimulus should be categorized.

SUSTAIN (*S*upervised and *U*nsupervised *ST*ratified *A*daptive *I*ncremental *N*etwork) is a model not just of categorization but also of category learning. Its architecture is in its initial stages similar to ALCOVE: the input layers consist of a set of separate detector representations for features, including verbally given category labels. Unlike in ALCOVE, however, these features are discrete valued rather than continuous. Examples given during training and stimuli given during testing are all represented as sets of value assignments to these features (equivalently, as vectors over features). Again, as with SUSTAIN, activation of these features is gated by attention before being passed on for processing. The next stage, however, differs significantly. Whereas in ALCOVE the system represented each exemplar individually, in SUSTAIN the system's memory contains a set of *clusters*, each of which is a single summary representation produced by averaging (or otherwise combining) individual exemplars. These clusters encode average or prototypical values of the relevant category in each feature.[13] The clusters in turn are mutually inhibiting, so activity in one tends to damp down the competition. They also pass activation to a final layer of feature nodes. The function of this layer is to infer

[13] Strictly speaking, SUSTAIN can generate exemplars as well as prototypes. Which it ends up working with depends on what distinctions are most useful in reducing errors during categorization. But this hybrid style of representation still distinguishes it from the pure exemplar-based ALCOVE model.

any properties of the stimulus that were not specified in the input. For instance, if the stimulus best fits the perceptual profile of a cat, but whether it meows is not specified, that feature would be activated or 'filled in' at the output layer. This allows the model to make inferences about properties of a category member that were not directly observed. Most importantly, the verbal label for the category is typically filled in at this stage. The label is then passed to the 'decision system', which produces it as the system's overall best guess as to what the stimulus is.

A few aspects of these models are noteworthy:

First, unlike NCAs, cognitive models clearly have a componential structure. All three models consist of (1) several distinct stages or layers, each of which (2) represents its own type of information and (3) processes it according to its own rules. Moreover, ALCOVE and SUSTAIN, at least, also (4) implement the idea of cognitive resources, since they make use of attention to modulate their performance. Representations are tokened at 'locations' within the architecture, processed, and then copied elsewhere. The representations themselves, the layers and sublayers, and the connections among them that implement the processes are all *components* of the model. And there is control over the flow of processing in the system—though in this case the control systems are fairly dumb, given that these are rigid, feed-forward systems. So cognitive models, unlike NCAs, break a system into its components and their interactions. This places them closer to mechanistic models in at least one respect.

Representations are the most obvious components of such models.[14] While in classical symbolic systems they would include propositional representations, here they include elements such as nodes representing either discrete-valued features or continuous dimensions, nodes representing parts or properties of a perceived visual scene, higher-level representations of

[14] This is not wholly uncontroversial. Some, such as Ramsey (2007), argue that connectionist networks and many other non-classical models do not in fact contain representations. Rather than enter this debate here, I am taking it at face value that the models represent what the modelers claim.

relations among these parts, nodes representing individual exemplars that the system has encountered or prototypes defined over those exemplars, and nodes representing categories themselves or their verbal labels. These models also contain processing elements that regulate the system, such as attentional gates, inhibitory connections between nodes, and ordinary weighted connections that transmit activity to the next layers. Layers or stages themselves are complex elements composed out of these basic building blocks. The properties of both sorts of elements—their functional profiles—are given by their associated sets of equations and parameters, e.g., activation functions, learning rules, and so on.

Second, the organization of these elements and operations corresponds to a map of a causal process. Earlier stages of processing in the model correspond to temporally earlier stages in real-world psychological processing, changes propagated through the elements of the model correspond to causal influences in the real system, activating or inhibiting a representation correspond to different sorts of real changes in the system, and so on. This also distinguishes these models from NCAs, since the flowchart or program of an NCA is not a causal diagram but an analytical one, displaying the logical dependence relations among functions rather than the organization of components in processing. Cognitive models are thus both componential and causal.[15]

Third, these models have all been empirically confirmed to some degree or other. JIM has been able to match human recognition performance on tasks involving scale changes, mirror image reversal, and image translations. On all of these, both humans and the model evince little performance degradation. By contrast, for images that are rotated in the visual plane, humans show systematic performance degradation, and so does the model. Similarly, both ALCOVE and

[15] This point is also endorsed by Glennan (2005). While I will disagree with his interpretation of these models as mechanistic, I agree that, as he puts it, in cognitive models "the arrows represent the causal interactions between the different parts" (p. 456).

SUSTAIN have been compared to a substantial number of datasets of human categorization performance. These include supervised and unsupervised learning, inference concerning category members, name learning, and tasks where shifting attention among features/dimensions is needed for accurate performance. Moreover, SUSTAIN has been able to succeed using a single set of parameter values for many different tasks.

Fourth, some aspects of these model clearly display black-boxing or 'filler' components. For instance, Hummel & Biederman note that "[c]omputing axes of symmetry is a difficult problem… the solution of which we are admittedly assuming" (1992, p. 487). The JIM model when it is run is simply given these values by hand rather than computing them. These components count as black boxes that presumably are intended to be filled in later.

To summarize, cognitive models are componentially organized, causally structured, semantically interpretable models of systems that are capable of producing or instantiating psychological capacities. Like mechanistic models, they can be specified at several different grains of analysis and may make use of epistemic short-cuts like black boxes or filler terms. They can also be confirmed or disconfirmed using established empirical methods. Despite these similarities, however, they are not mechanistic models. I turn now to the argument for this claim.

5. Model-based explanations without mechanisms

Reflect first on this functionalist truism: the relationship between a functional state or property of a system and the underlying state or property that realizes it is typically highly *indirect*. By 'indirect', what I mean is that one cannot in any simple or straightforward way read off the presence of the higher level state from the lower level state. The precise nature of the mapping by which functional properties (including psychological properties) are realized is often

opaque. While evidence of the lower level structure of a system can inform, constrain, and guide the construction of a theory of its higher level structure, lower level structures are not simple maps of higher level ones. Thus in psychology we have the obvious, if depressing, truth that the mind cannot simply be read off of the brain. Even if brains were less than staggeringly complex, it would still be an open question whether the organization that one discovers in the brain is the same as the one that structures the mind, and vice versa.

In attempting to understand the high level dynamics of complex systems like brains, modelers have recourse to many techniques for constructing such indirect accounts. Here I will focus on just three: *reification*, *functional abstraction*, and *fictionalization*. All of these play a role in undermining the claim that cognitive models are mechanistic.

Reification is the act of positing something with the characteristics of a more or less stable and enduring object, where in fact no such thing exists. Perhaps the canonical example of reification in cognitive science is the positing of symbolic representations in classical computational systems. Symbolic representations are purportedly akin to words on a page: discrete, able to be concatenated, moved, stored, copied, deleted. They are stable, entity-like constructs. This has given rise to a great deal of anxiety about the legitimacy of symbolic models. Nothing in the brain appears to 'stand still' in the way that symbols do, and nothing appears to have the properties of being manipulable in the way they are. This case is pushed strenuously by theorists like Clark (1992), who argues that the notion of an explicit symbol having these properties should be abandoned in favor of an account that sees explicit representation as a matter of the ease of retrieval of information and the availability of information for use in multiple tasks.[16]

[16] What this amounts to, then, is explicit representation *without* symbols. For my purposes here I don't endorse the anti-symbolic conclusion—indeed, part of what I'm arguing is that symbolic models (and representational models

In fact, from the point of view of neurophysiology, the distinction between representations and the processes that operate over them seems quite illusory. Representations and processes are inextricably entangled at the level of neural realization. In the dynamics of spike trains, excitatory and inhibitory potentials, and other events, there is no obvious separation between the two—indeed, all of the candidate vehicles of these static, entity-like symbols are themselves processes.[17] Dynamical systems theorists in particular have seen this as evidence that representational models of the mind, including all of the cognitive models considered here, should be rejected (Chemero, 2009; van Gelder, 1995). But this is an overreaction. Reification is a respectable, sometimes indispensible, tool for modeling the behavior of complex systems.

A further example of reification occurs in Just & Carpenter's (1992) model of working memory in sentence comprehension. The model (4CAPS) is a hybrid connectionist-production rule system, but one important component is a quantity, called 'activation', representing the capacity of the system to carry out various processes at a stage of comprehension. Activation is a limited-quantity property that attaches to representations and rules in the model. But while activation is a entity in the model, it does not correspond to any entity in the brain. Rather, it corresponds to a whole set of resources possessed by neural regions: "neurotransmitter function and various metabolic support systems, as well as the connectivity and structural integrity of the system" (Just, Carpenter, & Varma, 1999, p. 129). Treating this complex set of properties as a singular entity facilitates understanding the dynamics of normal comprehension, impaired comprehension due to injuries, and individual differences in comprehension. Moreover, it seems

---

more generally) can be perfectly correct and justified even if at the level of neurophysiological events there is only an entangled causal process that lacks the distinctive characteristics of localized, movable, concatenatable symbols.

[17] For a careful, thorough look at what would be required for neural systems to implement the symbolic properties of Turing machine-like computational systems, see Gallistel & King (2009).

to offer a causal explanation of the system's functioning: as these resources increase and decrease, the system becomes more or less able to process various representations.

Functional abstraction occurs when we decompose a modeled system into subsystems and other components on the basis of what they do, rather than their correspondence with organizations and groupings in the target system. To stave off an immediate objection, this isn't to say that functional groupings in systems are independent of their underlying physical, structural, and other organizational properties. But for many such obvious ways of dividing up the system there can also be *cross-cutting* functional groupings: ways of dividing the system up functionally that do not map onto the other sorts of organizational divisions in the system. Any system that instantiates functions that are not highly localized possesses this feature.

An example of this in practice is the division of processing in these three models into layer-like stages. For example, in JIM there are separate stages at which edges are extracted, vertices detected, attributes assigned, and geons represented. Earlier stages provide information to later stages and causally control them. At the same time, there appears to be a hierarchy of representation in visual processing in the brain. On the simplest view, at early stages (e.g., the retina or shortly thereafter), visual images are represented as assignments of luminance values to points in visual space. At progressively 'higher' (causally downstream and more centrally located) stages, more complex and abstract qualities are detected by successive regions that correspond to maps of visual space in terms of their proprietary features. So edges, whole line segments, movement, color, and so on, are detected by these higher level feature maps. The classic map of these areas was given by Felleman & van Essen (1991); for more recent work, see van Essen (2004). Since these have something like the structure of the layers or stages of JIM,

one might expect to be able to map the activity of layers in JIM onto these stages of neural processing.

Unfortunately, this is not likely to be possible. At the very least it would entail skipping steps, since there are likely to be several neural intermediaries between, e.g., edge detection and vertex computation. But more importantly, there may not be distinct neural maps whose component features and processes correspond to the layers of JIM. This point has even greater application to ALCOVE and SUSTAIN: since we can categorize entities according to indefinitely many features and along indefinitely many dimensions, even the idea of a structurally fixed and reasonably well-localized neural region that initiates categorization seems questionable. The same could be said of entities such as attentional gates. Positing such entities involves creating spots in the model where a complex capacity like attention can be plugged in and affect the process of categorization. But the notion of a 'place' in processing where attention modulates categorization is just the notion of attention's having a functionally defined effect on the process. There need not be any such literal, localized place.

If these models were *intended* to correspond structurally, anatomically, or in an approximate physiological way to the hierarchical organization of the brain, this would be evidence that they are at best incomplete, and at worst false. However, since these layers are functional layers, all that matters is that there is a stable pattern of organization in the brain that carries out the appropriate processes assigned to each layer, represents the appropriate information, and has the appropriate sort of internal and external causal organization. For example, there may not be three separate maps for vertex, axis, and blob information in visual cortex. This falsifies the model only if one assumes a simple correspondence between neural maps and cognitive maps.

There is some evidence that modelers are beginning to orient themselves away from localization assumptions in relating cognition to neural structures. The slogan of this movement is 'networks, not locations'. Just, Carpenter, & Varma (1999) comment that "[a]lmost every cognitive task involves the activation of a network of brain regions (say, 4-10 per hemisphere) rather than a single area" (p. 129). No single area "does" any particular cognitive function; rather, responsibility is distributed across regions. Moreover, cortical areas are multifunctional, contributing to the performance of many different tasks (p. 130). In the same spirit, Barrett (2009, p. 332) argues that

> psychological primitives are functional abstractions for brain networks that contribute to the formation of neuronal assemblies that make up each brain state. They are psychologically based, network-level descriptions. These networks are distributed across brain areas. They are not necessarily segregated (meaning that they can partially overlap). Each network exists within a context of connections to other networks, all of which run in parallel, each shaping the activity in the others.

Finally, van Orden, Pennington, & Stone (2001) amass a large amount of empirical evidence that any two putative cognitive functions can be dissociated by some pattern of lesion data, suggesting that any such localization assumptions are likely to fail.

Even in cases where there are such correspondences, they are likely to be partial. It has been proposed that several of the major functional components of SUSTAIN can be mapped onto neural regions: for instance, the hippocampus builds and recruits new clusters, the perirhinal cortex detects when a stimulus is familiar, and the prefrontal cortex directs encoding and retrieval of clusters (Love & Gureckis, 2007). First, none of these are unique functions of these areas; at best, they are among their many functions. This comports with the general idea that

neural regions are reused to carry out many cognitive functions. Second, and importantly, the exemplar storage system itself is not modeled here, a fact most plausibly explained by its being functionally distributed across wide regions of cortex. As Wimsatt (2007, pp. 191-2) remarks, "it is not merely that functionally characterized events and systems are spatially distributed or hard to locate exactly…. The problem is that a number of different functionally characterized systems, each with substantial and different powers to affect (or effect) behavior appear to be interdigitated and intermingled as the infinite regress of qualities-within-qualities of Anaxagoras' seeds." Cognitive models are poised to exploit just this sort of organization.

Finally, *fictionalization* involves 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.[18] In JIM, for instance, binding between two or more representations is achieved by synchronous firing, as is standard in many neural networks. To implement this synchrony, cells are connected by dedicated pathways called 'Fast Enabling Links' (FELs) that are distinct from the activation and inhibition-carrying pathways. Of their operation, the authors say (p. 498):

> In the current model, FELs are assumed to have functionally infinite propagation speed, allowing two cells to fire in synchrony regardless of the number of intervening FELs and active cells. Although this assumption is clearly incorrect, it is also much stronger than the computational task of image parsing requires.

The fact that FELs are independent of the usual channels by which cells communicate, and the fact that they possess physically impossible characteristics such as infinite speed suggests that not only can models contain black boxes or filler terms, they can also contain components that

[18] For an argument that the use of such fictional posits in even highly reliable models is widespread, see Winsberg (2010), Ch. 7.

cannot be filled in by any ordinary entities having the normal sorts of performance characteristics. Synchronic firing among distributed nodes needs to happen somehow, and FELs are just the devices that are introduced to fill this need.

We could think of FELs as a kind of useful fiction introduced by the modelers. Fictions are importantly unlike standard filler terms or black boxes. They are both an essential part of the operation of the model, and not clearly intended to be eliminated by any better construct in later iterations. In fact, Hummel & Biederman spend a great deal of their paper discussing the properties of FELs and their likely indispensability in future versions of the model (e.g., they discuss other ways of implementing binding through synchrony and argue for the superiority of the FEL approach; see p. 510). But like reified entities and functional abstraction, they do not correspond to parts of the modeled system. That does not mean that they are *wholly* fictional. There is reason to think that distinct neural regions, even at some distance, do synchronize their activity (Canolty, Ganguly, Kennerley, Cadieu, Koepsell, et al., 2010). How this happens remains poorly understood, but it is a certainty that there are no dedicated, high speed fiber connections linking these parts whose sole function is to synchronize firing rates. Alternatively we might say: there is something that does what FELs do, but it isn't an entity or a link or anything of that sort. FELs capture the general characteristic of neural systems that they often fire in synchrony. We can model this with FELs and lose nothing of interest. In modeling, we simply introduce a component that does the needed job—even if we recognize that there is in some sense no such *thing*.

To summarize, we should not be mislead into thinking that cognitive models are mechanistic by the fact that they are componential and causal. The reason is that even if the *intrinsic* structure of cognitive models resembles that of mechanistic models, the way in which

they *correspond* to the underlying modeled system is far less straightforward. These models often posit elements that have no mechanistic 'echo': they do not map onto parts of the realizing system in any obvious or straightforward way. To the extent that they are localizable, they are only *coarsely* so. In a good mechanistic model, elements appear in a model only when they correspond to a real part of the mechanism itself. This, recall, was the point of the Real Components Constraint (RCC) that was advanced in Section 3 to explain why Cummins' NCAs are not explanatory. But when it comes to cognitive models, not everything that counts as a component from the point of view of the *model* will look like a component in the modeled *system* itself—at least not if our notion of a component is based on a distinct, relatively localized physical entity like a cortical column, DNA strand, ribosome, or ion channel.

In light of this discussion, I would suggest that we need to make at least the following distinctions among types of models:

- *Phenomenal models*, such as the Hodgkin-Huxley equation;
- *Noncomponential analyses*, such as Cummins' analytic functional explanations;
- *Mechanistic models*, of the sort described by Craver, Bechtel, and Glennan;
- *Functional models*, of which cognitive models as described here are one example.

Phenomenal models obviously differ in their epistemic status, but the latter three types of models seem capable of meeting the general normative constraints on explanatory models perfectly well. In the spirit of explanatory pluralism, we should recognize the characteristic virtues of each modeling strategy rather than attempting to reduce them all to varieties of mechanisms.

6. Objections and replies

I now want to consider several objections to this conception of model-based explanation.

First objection: These models are in fact disguised mechanistic models—they're just bad, imperfect, or immature ones. They are models that are suffering from arrested development at the mechanism schema or sketch stage. Once all of the details are adequately filled in and their mechanistic character becomes more obvious, it can be determined to what degree they actually correspond to the underlying system.

Reply: Whether this turns out to be true in any particular case depends on the details. Some cognitive models may turn out to have components that can be localized, and to describe a sequence of events that maps onto a readily identifiable causal pathway in the neurophysiological description of the system. In these cases, the standard assumptions of mechanistic modeling will turn out to be satisfied. The cognitive model would turn out to be an abstraction from the neural system in the traditional sense: it is the result of discarding or ignoring the details of that system in favor of a coarse-grained or black-box description (much as we do with the simple three-stage description of hippocampal function).

However, there is no guarantee that this will be possible in all cases. What I am defending is the claim that these models provide legitimate explanations even when they are not sketches of mechanisms. No one should deny, first, that some capacities can be explained in terms of non-localistic or distributed systems. Many multilayer feedforward connectionist networks, as Bechtel & Richardson (1993, pp. 202-29) point out, satisfy this description. The network as a whole carries out a complex function but the subfunctions into which it might be analyzed correspond to no separate part of the network. So there is no way to localize distinct functions, but these network models are still explanatory. Indeed, Bechtel & Richardson argue that network models are themselves mechanistic insofar as their behavior is explicable in terms of the interactions of the simple components, each of which is itself an obviously mechanical

unit. They require only that "[i]f the models are well motivated, then component function will at least be consistent with physical constraints" (p. 228).

In the case envisaged here, 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. This is not like the case of mechanism schemata and sketches as described in Section 2. There we have what purports to be a model of the real parts, operations, and organization of the mechanism itself—one that may be incomplete in certain respects but which can be sharpened primarily by adding further details. Cognitive models *can* be refined in this way. But the additional details will themselves be functional abstractions of the same type, and hence will not represent an advance.

Glennan (2005) presents a view of cognitive models that is very similar to the one that I advocate here, but he argues that they are in fact mechanistic. Cognitive models—his examples are vowel normalization models—are mechanical, he claims, because they specify a list of parts along with their functional arrangement and the causal relations among them (p. 456); that is, "a set of components whose activities and interactions produce the phenomenon in question" (p. 457). Doing this is not sufficient for being a mechanistic model, in my view. The remaining condition is that the model must actually be a model of a real-world mechanism—that is, there must be the right sort of mapping from model to world.

Glennan correctly notes that asking whether a model gets a mechanism "right" is simplistic. Models need not be straightforwardly isomorphic to systems, but may be *similar* in various respects. The issue is whether the kinds of relationships that I have canvassed count in favor of their similarity or dissimilarity. Cognitive models as I construe them need not map model entities onto real-world entities, or model activities and structures onto real-world

activities and structures. Entities in models may pick out capacities, processes, distributed structures, or other large-scale functional properties of systems. Glennan shows some willingness to allow these sorts of correspondence: "[i]n the case of high level cognitive mechanisms, the parts themselves may be complex and highly distributed and may defy our strategies for localization" (2005, p. 459). But if these sorts of correspondences are allowed, and if these sorts of entities are allowed to count as 'parts', it is far from clear what content the notion of a mechanism has anymore.

It is arguable that the notion of a part of a mechanism should be closely tied to the sort of thing that the localization heuristic counsels us to seek. Mechanisms, after all, involve the coordinated *spatial and temporal* organization of parts. The heart doesn't beat unless the ventricles, valves, etc. are spatially arranged in the right sort of way, and long-term potentiation is impossible unless neurotransmitter diffusion across synaptic gaps can occur in the needed time window. Craver (2007, pp. 251-3) emphasizes this fact, noting that compartmentalizing phenomena in spatial locations and determining the spatial structure and orientation of various parts are crucial to confirming mechanistic accounts. If parts are allowed to be smeared-out processes or distributed system-level properties, the spatial organization of mechanisms becomes much more difficult to discern. In the case of ALCOVE and SUSTAIN, the 'mechanism' might include large portions of the neocortex, since this may be required for the distributed storage of exemplars. It is more than just a terminological matter whether one wants to count these as parts of mechanisms. 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.[19]

Second objection: If these are not mechanism sketches, then they are not describing the real structure of the underlying system at all. So they must be something more like merely phenomenal models: behaviorally adequate, perhaps, but non-explanatory.

Reply: This objection gives voice to a view that might be called 'mechanism imperialism'. It neglects the possibility that a system's behavior can be explained from many distinct epistemic perspectives, each of which is illuminating. Viewed from one perspective, the brain might be a hierarchical collection of neural mechanisms; viewed from another, it might instantiate a set of cognitive models that classify the system in ways that cut across mechanistic boundaries.

This point can be put more sharply. First, recall from Section 2 that explanatory models differ from phenomenal models in that they allow for control and manipulation of the system in question, and they allow us to answer various counterfactual questions about the system's behavior. Cognitive models allow us to do both of these things. Models such as ALCOVE are actually implemented as computer programs, and they can be run on various data sets, under different task demands, with various parameter values systematically permuted, and even artificially lesioned to degrade their performance. Control and manipulation can be achieved because these models depict one aspect of the causal structure of the system. They also describe the ways in which the internal configuration and output performance of the system vary with novel inputs and interventions. So this explanatory demand can be met.

[19] The individuation question is particularly important if cognitive functions take place in networks that are reused to implement many different capacities. The very same 'parts' could then simultaneously be part of many interlocking mechanisms. This too may constitute a reason to use localization as a guide to genuine mechanistic parthood.

Further, these models meet at least one form of the Real Components Constraint (RCC) described at the end of Section 2. This may seem somewhat surprising in light of the previous discussion. I have been arguing that model elements need not correspond to parts of mechanisms. How, then, can these models meet the RCC? The answer depends on different ways of interpreting the RCC's demand. Craver, for example, gives a list of criteria that a 'real part' of a mechanism must meet. Real parts, he says (2007, pp. 131-3):

1. Have a stable cluster of properties;
2. Are robust, i.e., detectable with a variety of causally and theoretically independent devices;
3. Are able to be intervened on and manipulated;
4. Are 'physiologically plausible', in the sense of existing only under regular non-pathological conditions.

The constructs posited in cognitive models satisfy these conditions. Take attentional gates as an example. These have a stable set of properties: they function to stretch or compress dimensions along which exemplars can be compared, rendering them more or less similar than they would be otherwise. The effects of attention are detectable by performance in categorization, but also in explicit similarity judgments, in errors in detecting non-attended features of stimuli, etc. Attention can be manipulated both intentionally and implicitly, e.g., by rewarding successful categorizations; it can also be disrupted by masking, presenting distractor stimuli, increasing task demands, etc. Finally, attention has a normal role to play in cognitive functioning. Since attentional gates are model entities that stand in for the functioning of this capacity, they should count as 'real parts' by these criteria.

The point here is not, of course, that these criteria show that cognitive models are mechanistic. Rather it shows that these conditions, which purport to govern 'real parts', are in fact more general. To see this, observe that these criteria could all perfectly well be accepted by Cummins. He requires of a hypothesized analysis of a capacity C that the subcapacities posited be independently attested. This is equivalent to saying that they should be robust (condition 2). Conditions 1 and 3 are also straightforwardly applicable to capacities, and a case can be made for condition 4 as well—the subcapacities in question should characterize normal, not pathological, cognitive functioning. These conditions are not merely ones under which we are warranted in hypothesizing the existence of parts of mechanisms, but general norms of explanations that aspire to transcend the merely phenomenal. Since cognitive models (and non-componential analyses) can satisfy these conditions, and since they provide the possibility of control and manipulation as well as allowing counterfactual predictions, they are not plausibly thought of as phenomenal models.

Third objection: If these models need not correspond to the underlying anatomical and physiological organization of the brain, then they are entirely unconstrained. We could in principle pick any model and find a sufficiently strange mapping according to which it would count as being realized by the brain. This approach doesn't place substantial empirical constraints on what counts as a good cognitive model.

Reply: This is a non sequitur, so I will deal with it only briefly. First, cognitive models can be confirmed or disconfirmed independent of neurobiological evidence. Many such models have been developed and tested solely by appeal to their fit with behavioral data. Where these are sufficiently numerous, they provide strong constraints on acceptable models. For example, ALCOVE and SUSTAIN differ in whether they allow solely exemplar representations to be used

or both exemplars and prototypes. Which form of representation to use is a live empirical debate, and the evidence educed for each side is largely behavioral (Malt, 1989; Minda & Smith, 2002; Smith & Minda, 2002).

Second, cognitive models are broadly required to be consistent with one another and with our background knowledge. So well-confirmed models can rule out less well-confirmed ones if they generate incompatible predictions or have conflicting assumptions. Models can be mutually constraining both within a single domain (e.g., studies of short-term memory) and across domains (e.g., memory and attention). The ease of integrating models from various cognitive task domains is in part what motivates sweeping attempts to model cognitive architecture in a unified framework, such as ACT-R and SOAR (Anderson, 1990; Newell, 1990).

And third, even models that are realized by non-localized states and processes can be empirically confirmed or disconfirmed. The nature of the evidence required to do so is, however, much more difficult to gather than in cases of simple localization. On the status of localization assumptions in cognitive neuroscience, and empirical techniques for moving beyond localization, see the papers collected in Hanson & Bunzl (2010).

7. Conclusions

Mechanistic explanation is a distinctive and powerful framework for understanding the behavior of complex systems, and it has demonstrated its usefulness in a number of domains. None of the arguments here are intended to cast doubt on these facts. However, we should bear in mind—on pain of falling prey to mechanism imperialism—that there are other tools available for modeling complex systems in ways that give us explanatory traction. Herein I have argued that, first, the norms governing mechanistic explanation are general norms that can be applied to

a variety of domains. Second, even noncomponential forms of explanation such as functional analysis can in principle satisfy these norms. And third, cognitive models of the sort that are common in many parts of psychology, cognitive neuroscience, and neuropsychology can also satisfy these normative standards without thereby being mechanistic. While psychological models do not describe mechanisms, they are not defective for failing to do so. They gain their explanatory force from picking out functionally individuated strands in the complex causal web that winds through the brain. We can gain various forms of practical and epistemic leverage by attending to these strands, but we should not expect to be able to descry them in the underlying mechanistic structure of the brain itself.

Acknowledgements

I thank the following people for helpful comments and discussion: Carl Craver, Gualtiero Piccinini, and the participants in the Society for the Metaphysics of Science symposium at the 2010 Pacific APA meeting, where an earlier version of this material was presented. Thanks also to an anonymous reviewer for this journal for very thoughtful comments on a previous draft.

References

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Lawrence Erlbaum.

Barrett, L. F. (2009). The future of psychology: Connecting mind to brain. *Perspectives on Psychological Science*, 4, 326-339.

Bechtel, W. (2006). *Discovering Cell Mechanisms: The Creation of Modern Cell Biology*. Cambridge: Cambridge University Press.

Bechtel, W. (2008). *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. New York: Routledge.

Bechtel, W. (2009). Looking down, around, and up: Mechanistic explanation in psychology. *Philosophical Psychology*, 22, 543-564.

Bechtel, W., & Abrahamson, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36, 421-441.

Bechtel, W., & Richardson, R. C. (1993). *Discovering Complexity*. Princeton: Princeton University Press.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94, 115-147.

Biederman, I. (1995). Visual object recognition. In S. F. Kosslyn & D. N. Osherson (Eds.), *An Invitation to Cognitive Science* (2nd ed.), pp. 121-165. Cambridge: MIT Press.

Boden, M. (2006). *Mind as Machine: A History of Cognitive Science* (2 vols.). Oxford: Oxford University Press.

Bokulich, A. (forthcoming). How scientific models can explain. *Synthese*.

Canolty, R. T., Ganguly, K., Kennerley, S. W., Cadieu, C. F., Koepsell, K., Wallis, J. D., & Carmena, J. M. (2010). Oscillatory phase coupling coordinates anatomically dispersed functional cell assemblies. *Proceedings of the National Academy of Sciences*, 107, 17356-17361.

Chemero, T. (2009). *Radical Embodied Cognitive Science*. Cambridge: MIT Press.

Clark, A. (1992). The presence of a symbol. *Connection Science*, 4, 193-205.

Craver, C. F. (2006). What mechanistic models explain. *Synthese*, 153, 355-376.

Craver, C. F. (2007). *Explaining the Brain*. Oxford: Oxford University Press.

Craver, C. F. (2009). Mechanisms and natural kinds. *Philosophical Psychology*, 22, 575-594.

Craver, C. F., & Bechtel, W. (2006), Mechanism. In S. Sarkar & J. Pfeifer (Eds.), *Philosophy of Science: An Encyclopedia* (pp. 469-478). New York: Routledge.

Cummins, R. (1983). *The Nature of Psychological Explanation*. Cambridge: MIT Press.

Estes, W. K. (1994). *Classification and Cognition*. Oxford: Oxford University Press.

Felleman, D. J., & van Essen, D. C. (1991). Distributed hierarchical processing in primate visual cortex. *Cerebral Cortex*, 1, 1-47.

Gallistel, C. R., & King, A. P. (2009). *Memory and the Computational Brain*. Malden, MA: Wiley-Blackwell.

Giere, R. (1988). *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.

Glennan, S. (1996). Mechanisms and the nature of causation. *Erkenntnis*, 44, 49-71.

Glennan, S. (1997). Capacities, universality, and singularity. *Philosophy of Science*, 64, 605-626.

Glennan, S. (2005). Modeling mechanisms. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36, 443-464.

Hanson, S. J., & Bunzl, M. (Eds.) (2010). *Foundational Issues in Human Brain Mapping*. Cambridge, MA: MIT Press.

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

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99, 122-149.

Just, M. A., Carpenter, P. A., & Varma, S. (1999). Computational modeling of high-level cognition and brain function. *Human Brain Mapping*, 8, 128-136.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99, 22-44.

Kruschke, J. K. (2008). Models of categorization. In R. Sun (Ed.), *The Cambridge Handbook of Computational Psychology* (pp. 267-301). Cambridge: Cambridge University Press.

Love, B. C., & Gureckis, T. M. (2007). Models in search of a brain. *Cognitive, Affective, and Behavioral Neuroscience*, 7, 90-108.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, 111, 309-332.

Machamer, P. (2004). Activities and causation: The metaphysics and epistemology of mechanisms. *International Studies in the Philosophy of Science*, 18, 27-39.

Machamer, P., Darden, L., Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science*, 67, 1-25.

Malt, B. C. (1989). An on-line investigation of prototype and exemplar strategies in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 15, 539-555.

Minda, J. P., & Smith, J. D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 275-292.

Murphy, G. (2002). *The Big Book of Concepts*. Cambridge: MIT Press.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge: Harvard University Press.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology*, 115, 39-57.

Polk, T. A., & Seifert, C. M. (Eds.) (2002). *Cognitive Modeling*. Cambridge: MIT Press.

Ramsey, W. (2007). *Rethinking Representation*. Cambridge: Cambridge University Press.

Smith, J. D., & Minda, J. P. (2002). Distinguishing prototype-based and exemplar-based processes in dot-pattern category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 800-811.

Tarr, M. J. (2002). Object Recognition. In L. Nadel (Ed.), *Encyclopedia of Cognitive Science* (pp. 490-494). London: Nature Publishing Group.

Tarr, M. J., & Bülthoff, H. H. (1998). *Object Recognition in Man, Monkey, and Machine*. Cambridge: MIT Press.

Ullman, S. (1996) *High-Level Vision: Object Recognition and Visual Cognition*. Cambridge: MIT Press.

van Essen, D. C. (2004) Organization of visual areas in macaque and human cerebral cortex. In L. Chalupa & J.S. Werner (Eds.), *The Visual Neurosciences*, pp. 507-521. Cambridge: MIT Press.

van Gelder, T. (1995). What might cognition be, if not computation? *Journal of Philosophy*, 92, 345-81.

van Orden, G. C., Pennington, B. F., & Stone, G. O. (2001). What do double dissociations prove? *Cognitive Science*, 25, 111-172.

Weiskopf, D. A. (forthcoming). The functional unity of special science kinds. *British Journal for the Philosophy of Science*.

Wilson, R. A. (2004). *Boundaries of the Mind: The Individual in the Fragile Sciences—Cognition*. Cambridge: Cambridge University Press.

Wimsatt, W. C. (2007). *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Cambridge: Harvard University Press.

Winsberg, E. (2010). *Science in the Age of Computer Simulation*. Chicago: University of Chicago Press.

Woodward, J. (2002). What is a mechanism? A counterfactual account. *Philosophy of Science*, 69, S366-S377.