The explanatory autonomy of cognitive models

Daniel A. Weiskopf

1. Many models, one world

The mind/brain, like any other complex system, can be modeled in a variety of ways.[1] Some of these involve ignoring or abstracting from most of its structure: for the purpose of understanding overall glucose metabolism in the body, we can neglect the brain's intricate internal organization and treat it simply as a suitably discretized homogeneous mass having certain energy demands (Gaohua & Kumura, 2009). Other projects demand more fine-grained modeling schemes, as when we are trying to map cortical white-matter density and connectivity (Johansen-Berg & Rushworth, 2009), or the distribution of various neurotransmitter receptor sites (Zilles & Amunts, 2009). Here, the system's detailed structural and dynamical properties matter, although not necessarily the same ones in every context. A single system may admit of many possible simplifying idealizations, and how we model a system—which of its components and properties we choose to represent, and how much detail we incorporate into that representation—is fundamentally a pragmatic choice.

When we have multiple models of a single target system, we face the problem of how to integrate these models into one coherent picture. We wish to understand these models not merely as singular glimpses, but as parts of a unified view of the system. This problem arises in a variety of domains, from atmospheric and climate modeling (Parker, 2006) to understanding the division of labor among social insects (Mitchell, 2002) to

[1] Here I am borrowing Chomsky's (2000, p. 9) hybrid term 'mind/brain' to denote the brain considered as a system instantiating both a complex neural and cognitive organization. It is meant to encompass both of these aspects, the biological and the psychological (as well as any other relevant types of causal organization).

modeling the structure of the atomic nucleus (Morrison, 2011). Concerns about integration arise whenever we are uncertain how two or more representations of the same system fit together in a way that gives us insight into the system's real organization. The problem is complicated by the fact that often models that are individually well-validated in terms of their ability to explain a range of phenomena will represent one and the same target system as having substantially different, or even seemingly contradictory, properties.[2]

There are a number of available strategies for integrating models in ways that resolve these tensions. We may be able to show one model to be an approximation to another, such as when we sharpen a model of the simple pendulum by incorporating facts about air resistance, friction, and the mass of the string. We may depict one model as an embedded component or sub-model of the other. Or we may be able to show that models apply to physically distinct aspects of the same system, so that they never actually represent the very same parts of the system in contradictory ways. This practice is standard in fluid dynamics, which treats fluids as lacking viscosity in regions where ideal fluid treatments are appropriate, and as having viscosity elsewhere, such as near walls or other boundaries.

The construction and testing of psychological models has a long history in the cognitive and behavioral sciences, and thanks to an impressive array of electrophysiological and imaging technologies we can also construct sophisticated models of the structure and dynamics of neural systems, both during the execution of tasks and in their resting or "default" state. This gives rise to what may be regarded as one modern form of the mind-body problem: how are these two types of models related?

[2] See, for example, Morrison's (2011) discussion of inconsistent models of the atomic nucleus.

Specifically, how are they to be integrated to provide a complete understanding of the mind/brain system as a whole? And what strategies are available if they cannot be neatly integrated?

The problem of integrating psychological and neuroscientific models is especially challenging, since these models are derived from different theoretical frameworks, use distinct explanatory primitives, are responsible for different experimental phenomena, and are tested and validated using different methods. Recently, some philosophers of science (e.g., Piccinini & Craver, 2011) have claimed that the framework of mechanistic explanation provides a solution to the problem of unification. They suggest that if we view psychological models as mechanistic, they can be smoothly integrated with the typical multilevel explanatory constructs of neuroscience. Here I wish to challenge this extension of the mechanistic program. While mechanistic explanation is a distinctive and important strategy in a number of scientific domains, not every attempt to capture the behavior of complex systems in terms of their causal structure should be seen as mechanistic (Woodward, 2013).

Mechanistic explanations, I suggest, are one member of the class of causal explanations, specifically the wider class of *componential causal explanations* (Clark, 1997, pp. 104-5). Many psychological explanations also fall within this class, but they differ in important respects from mechanistic explanations understood more narrowly. Despite these differences, psychological explanations are capable of fitting or capturing real aspects of the causal structure of the world, just as much as mechanistic explanations are. Thus a defense of the legitimacy of model-based psychological explanation is at the same time a defense of the reality of the cognitive structures that these models map onto.

2. Cognitive models

Psychology, like any other science, pursues multiple goals in parallel. Observational and experimentally-oriented studies often aim simply to produce and refine our characterizations of new phenomena to be explained. In other frames of mind, psychologists aim to generate theories and models that explain these phenomena. One common explanatory strategy involves producing *cognitive models*. Generally, a cognitive model takes a cognitive system as its intended target, and the structures that it contains purport to characterize this system in a way that captures its cognitive functioning. Such a model uses a proprietary set of modeling resources to explain some aspect of the system's functioning, whether normal or pathological.

These models typically describe systems in terms of the *representations*, *processes* and *operations*, and *resources* that they employ. These psychological entities constitute the basic explanatory toolkit of cognitive modeling. Representations include symbols (perceptual, conceptual, and otherwise), images and icons, units and weights, state vectors, and so on. Processes and operations are various ways of combining and transforming these representations such as comparison, concatenation, and deletion. Resources include parts of the architecture, including memory buffers, information channels, attentional filters, and process schedulers, all of which govern how and when processing can take place.[3]

[3] Some have argued that psychological processes should not be understood in representational terms (van Gelder, 1995; Chemero, 2009). Models developed within a non-representational framework will accordingly use a different toolkit of basic explanatory constructs. Cognitive models themselves are defined in terms of their explanatory targets, not whether they use representational states as their primitives.

These models can take a number of different forms, depending on the kind of format that they are presented in:[4]

1) *Verbal descriptions*: Words may be sufficient to specify some particularly simple models, or to specify models in terms of their rough qualitative features. As an example, take the levels of processing framework in memory modeling (Craik & Lockhart, 1972; Craik & Tulving, 1975; Craik & Cermak, 1979). This model makes two assumptions: (1) that novel stimuli are interpreted in terms of a fixed order of processing that operates over a hierarchy of features, starting with their superficial perceptual characteristics and leading to more conceptual or semantically elaborated characteristics; (2) that depth of processing, as defined in terms of movement through this fixed hierarchy, predicts degree of memory encoding, so that the more deeply and elaborately a stimulus is processed, the more likely it is to be recalled later. Although these two assumptions are schematic and require much more filling in, they outline a framework that is already sufficient to guide experimentation and generate determinate predictions about recall and recognition performance—they predict, for instance, that manipulating the conditions of memory encoding so that only perceptual features are processed should result in poorer recall.

2) *Mathematical formalism*: Mathematical equations and related formalisms, e.g., geometric and state-space models, have a number of applications in modeling cognition. Dynamical systems models provide one paradigmatic example. These models typically represent cognitive states as points or regions in a low-dimensional state space and

[4] For extensive discussion of further types of models and model-construction procedures, see Busemeyer & Diederich (2009), Gray (2011), Lewandowsky & Farrell (2007), and Shiffrin (2010). A taxonomy similar to the one proposed here occurs in Jacobs & Grainger (1994).

cognitive processes as trajectories through that space. The governing equations determine the trajectory that the system takes through the space under various parametric regimes.

Equations may also be used to specify the form cognitive processes take. For instance, Amos Tversky's (1977) Contrast Rule specifies that the similarity of two objects belonging to different categories (*a* and *b*) is a weighted function of their common attributes minus their distinctive attributes:

$$Sim(a,b) = \alpha f(a \cap b) - \beta f(a - b) - \gamma f(b - a).$$

The form of the equation itself carries implications about category representation, since it requires that categories *a* and *b* be associated with distinct sets of separable features whose intersection and differences can be taken. It is also possible to interpret the equation as specifying the causal process of computing similarities, in which three distinct comparison operations are carried out and then subtracted to yield an overall similarity evaluation. Support for this causal interpretation might be provided by studies that vary the common and distinctive features possessed by two categories and track the effects of these manipulations on ratings of overall similarity. One way to support a causal interpretation of an equation is to use it to design manipulations that have systematic effects such as these.

3) *Diagrams and graphics*: There are many varieties of graphical models, but the most common are so-called 'boxological' models. The main components of these models are boxes, which stand for distinct functional elements, and arrows, which stand for relationships of control or informational exchange. A cognitive architecture can be described at one level of functional analysis by a directed graph made of such elements.

Boxological models are employed in many domains. In the early so-called 'modal model', human memory was represented as having three separate stores (sensory buffers, short term memory, and long term memory), with a determinate order of processing and a set of associated control processes for orchestrating rehearsal (Atkinson & Shiffrin, 1968). In later models, the construct of working memory takes center stage; these models posit three different core components: the central executive, visuospatial sketchpad, and phonological loop (Baddeley & Hitch, 1974). In more recent iterations, they introduce more structures such as the episodic buffer and episodic long-term memory. This organization is illustrated in Figure 1 (after Baddeley, 2000). With each successive development, new functional components are added and old ones are divided into more finely specified subsystems.[5] This pattern is familiar from other domains such as the study of reading performance, which has centered on developing models that differ in how they functionally decompose the system underlying normal fluent reading (Coltheart, Curtis, Atkins, and Haller, 1993).

[5] This model in particular is discussed at greater length in section 6.

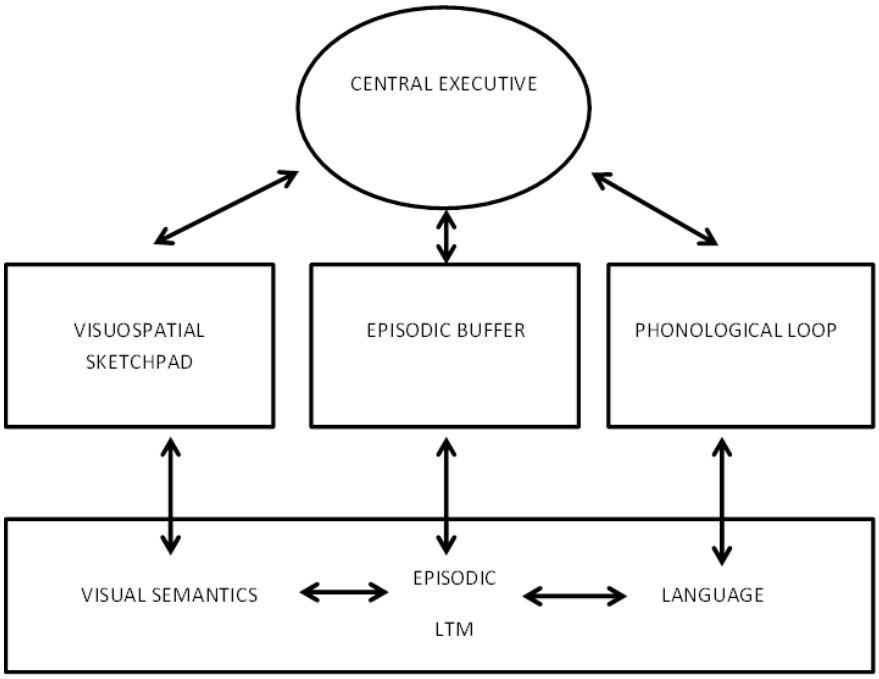

Figure 1: Baddeley's (2000) model of working memory

Diagrams often serve as the basis for hybrid models which make use of a host of representational tools (visual, verbal, mathematical) to describe cognitive systems. In all boxological models, cognitive subcomponents and their interactions are depicted as part of a directed graph, and the simplest of these models depict only this much structure. The functions of boxes, as well as the connections among them, may be specified by verbal labels or mathematical formulae. Importantly, these need not be regarded as black boxes whose inner workings are opaque: greater detail about how exactly each box carries out its function can be given by an associated description of the representations and processes that the box uses in carrying out its internal operations, and each box may be recursively decomposed into further subsystems. Boxological models offer numerous open 'slots' where further refinements may naturally be incorporated. The process of decomposition

continues until there is no longer anything useful to say, cognitively speaking, about the operations of these components.

4) *Computational models or simulations*: Computer simulations are often used to model cognitive processes because computer programs offer a concrete way to extract determinate predictions from cognitive models, and because models embodied in computer programs can be directly compared in terms of their ability to account for the same data (Winsberg, 2010). As McClelland puts it: "The essential purpose of [computational] cognitive modeling is to allow investigation of the implications of ideas, beyond the limits of human thinking. Models allow the exploration of the implications of ideas that cannot be fully explored by thought alone" (2009, p. 16). Typical examples of computer simulations of cognitive processes include some large-scale cognitive architectures such as Soar (Newell, 1990) and ACT-R (Anderson, Bothell, Byrne, Douglass, Lebiere, & Qin, 2004), as well as neural network models (Rogers & McClelland, 2004).

While mathematical models are often used as the source for computational models, the two belong to distinct types, since a set of mathematical equations can be manipulated or solved using many different computer programs implemented on many types of hardware architecture. In principle, however, any type of model can be used to construct a computer simulation, as long as its operations are capable of being described in a sufficiently precise way for us to write a program that executes them. Embodying open-ended models in programs often forces us to make a number of highly specific decisions about how the model functions, what values its parameters take, and so on, whereas other models are designed and constructed from the very start as programs.

It is important in considering computational simulations to distinguish features of the program from the features of the model that lie behind the program. So, for example, simulating object recognition with a computational routine written in LISP does not in any way commit us to thinking of the visual system itself as computing using such primitive functions as CAR, CDR, etc. These aspects of the programming language are not intended to be interpreted in terms of characteristics of the modeled system. Which of these aspects are supposed to be projected onto the cognitive system is a matter requiring careful interpretation.[6] ACT-R assumes that psychological operations consist of the application of production rules (which the program simulates), but not that they involve the execution of lines of compiled C code, and neural network models assume that cognition involves passing activation in parallel through a network of simple units, despite the fact that this activity is almost always simulated on an underlying serial computational architecture. Turning a model into a program is something of an art, and not every aspect of the resulting program should be interpreted either as part of the model that inspired it or as part of the target system itself.

This brief discussion serves to illustrate several points. First, cognitive models come in many varieties, and any discussion of their strengths and weaknesses needs to be sensitive to this diversity. Second, these models are selective simplifications: they typically aim to capture the performance of some relatively restricted aspect or subsystem of the total cognitive system, and to do so in terms of relatively few variables or factors.[7]

[6] Thus see, for example, the discussion in Cooper & Shallice (1995) of the distinction between Soar as a psychological theory and Soar as an implemented program.

[7] The main exceptions here are neural network models, which tend to be composed of hundreds or thousands of independent units, and even more weights connecting them. Recent models contain as many as 2.5 million units representing neurons, networks, or regions (Eliasmith et al., 2012). In light of this, network models are distinguished by the fact that their performance tends to be impenetrable to casual

Verbal descriptions, for instance, are obvious simplifications of cognitive processing, and mathematical models often aim for compactness of expression rather than capturing everything about a system's performance. And third, these models typically individuate their components in a way that is neutral with respect to the underlying physical structure of the system that realizes them. Although the system's physical structure and the organization of cognitive models do constrain one another, cognitive models themselves are physically noncommittal. This last point will be especially important in the forthcoming sections, which aim to distinguish cognitive models from mechanistic models.

3. The mechanist's challenge

Cognitive modeling provides a rich set of resources for representing and explaining the performance of cognitive systems. At the same time, some philosophers of science have argued that the characteristic mode of explanation in the life sciences and the sciences of complex systems more generally is *mechanistic*. Mechanistic explanations take as their targets the capacities (functions and behaviors) of particular systems, and they try to explain these capacities by breaking the system down into its component parts, enumerating their activities, operations, and interactions, and describing how they are organized (Bechtel, 2008; Bechtel & Abrahamson, 2005; Craver, 2007; Glennan, 2002). A *mechanistic model* is one that represents a system in terms of this sort of componential analysis. Such models, when they are accurate, display the system's mechanistic

---

inspection, thus giving rise to a host of analytic techniques (e.g., cluster analysis) to uncover the salient operations that explain their behavior.

organization and thus make intelligible how the dynamic activities of the components can produce the target phenomena (Kaplan & Craver, 2011).

It is indisputable that many successful explanations, particularly in neuroscience and biology, take the form of giving mechanistic models for systems. Canonical examples from neurophysiology include our best understanding of how action potentials are produced in neurons by the movement of ions across various transmembrane voltage-gated channels, and the processes by which action potentials can induce neurotransmitter release at synaptic junctions. Explaining these phenomena involves giving a detailed account of the physical organization of the components of the cell membrane and their functional profiles, the active intracellular elements that package neurotransmitters for release, the movements of various messenger molecules to key regions in the synapse, and so on. When these components are spatiotemporally integrated in just the right way and given the appropriate initiating stimulus, they will produce the phenomena associated with neural spiking and transmitter release. Mechanisms explain phenomena because they are the causes of those phenomena (or important parts of their causes). These explanations rank among the greatest modeling successes in cellular neurophysiology, and similar accounts can be given for other neural and biological phenomena at a number of spatial and temporal scales (for historical background, see Shepherd, 2010).

Supposing that many neural systems can be modeled mechanistically, the question arises: how are the cognitive models produced by psychologists related to the various multilevel mechanistic models produced by the neurosciences? An answer traditionally offered by functionalist philosophy of mind says that the domain of psychology is at least partially *autonomous* from the underlying physical details of

implementation (Fodor, 1974). Autonomy can be understood in a number of ways, but in the present context I intend it to cover both taxonomic and explanatory autonomy.

To say that psychology has *taxonomic autonomy* is to say that the range of entities, states, and processes that psychology posits as part of its basic modeling toolkit, and the kinds of structure that these models incorporate, are at most constrained only loosely by the way other sciences may model the mind/brain, and in particular by the details of physical implementation.[8] What appears as an entity or process in a cognitive model need not appear as such in any other model of the same system. Consequently, the structure that cognitive models impose on the system may differ sharply from the structure other models do. Hence cognitive models are allowed to carve up the world in a way that aligns first and foremost with the ontological and theoretical commitments of psychology.

To say that psychology has *explanatory autonomy* is to say that cognitive models are sufficient by themselves to give adequate explanations of various psychological phenomena. For example, in the domain of memory, there is a host of robust phenomena, including interactions between encoding and retrieval conditions, the specificity with which prior learning transfers to new tasks, rates of relearning, interference and serial position effects in recall, and so on. Cognitive models of memory attempt to capture some or all of these phenomena by positing underlying memory stores, types of representations and encoding schemes, and control processes. Spelling these out in

[8] There is nothing unique about psychology in this; every field involves setting out its phenomena and the set of concepts and entities it will use in explaining them. Here I follow Darden & Maull (1977) in taking *fields* to be defined by packages consisting of core problems, phenomena related to those problems, explanatory factors and goals related to solving such problems, specific techniques and methods for solving them, a proprietary vocabulary, and various concepts, laws, and theories that may be brought to bear on them.

sufficient detail describes the causal structure of the cognitive system and thereby explains how the phenomena are produced by the interactions among representations, processes, and resources. The autonomy thesis says that these phenomena can be given a wholly adequate explanation in terms of some cognitive model. That isn't to say that there might not be other possible explanations of the phenomena as well—autonomy does not imply uniqueness. It does imply that psychological modeling practices can stand on their own, however, and are not incomplete in principle.

Neither taxonomic nor explanatory autonomy requires that there is a privileged *evidential* base for the construction of cognitive models or theories. These models may be confirmed or disconfirmed by appeal to potentially any piece of evidence (introspective, behavioral, neurophysiological, clinical, etc.).[9] And neither implies that cognitive models cannot be integrated with other models to produce interlevel models. Cognitive neuropsychology, for example, is a distinctive field that explicitly aims to link psychological function with neural structure in just this way. Autonomy says only that cognitive models are capable by themselves of meeting any standards of taxonomic legitimacy and explanatory adequacy.

In a recent paper, Piccinini & Craver (2011; henceforth P&C) argue that integrating psychology with neuroscience will involve denying at least explanatory autonomy, and perhaps taxonomic autonomy as well. They argue for two related claims about the relationship between psychological and neuroscientific explanation:

[9] For an argument that cognitive theories have not, and perhaps cannot, be either supported or undermined by neuroimaging data, see Coltheart (2006); for a response, see Roskies (2009).

Common Type Claim: Psychological and neuroscientific explanations belong to a common type: both are mechanistic explanations. As P&C put it, "[f]unctional analysis cannot be autonomous from mechanistic explanation because the former is just an elliptical form of the latter" (p. 290). Consequently, these explanations take the same general form, and are subject to the same explanatory norms.

Sketch Claim: Explanations in terms of psychological mechanisms are *sketches* of more completely filled in neuroscientific mechanistic explanations. And more generally, "functional analyses are *sketches of mechanisms*, in which some structural aspects of a mechanistic explanation are omitted" (P&C, p. 284). This claim presents a picture of how models are integrated that is considerably stronger than the mere claim that psychological explanations will ultimately (somehow) be cashed out in terms of neuroscientific mechanisms, or that the psychological is *realized* by the neural.

The two claims are connected, insofar as the common type claim states that functional explanation in psychology is mechanistic, and the sketch claim says that *qua* mechanistic explanations they are incomplete. Summarizing their position, P&C say: "Psychological explanations are not distinct from neuroscientific ones; each describes aspects of the same multilevel mechanisms" (288).

I will argue that both of these claims are false. The truth of the common type claim turns on how we interpret the scope of mechanistic explanations. If they are understood in a relatively conservative way, the claim fails, whereas liberalizing the

conception of mechanistic explanation empties it of any distinctive content. The sketch claim is also false, since the relationship between psychological and neuroscientific explanations is not, in general, one in which the neuroscientific explanations involve filling in more 'missing details' or unpacking black boxes and filler terms present in psychological models. Psychology is not simply delivering approximate or idealized versions of neuroscientific explanations.

4. Against psychological mechanisms

In their influential discussion of mechanisms, Bechtel & Richardson (1993) traced the historical development of heuristics employed in the mechanistic analysis of complex systems. Chief among these are the twin heuristics of *decomposition* and *localization*. Decomposition is a form of functional analysis. It involves taking the overall function of a system and breaking it down into various simpler subfunctions whose processes and interactions jointly account for the overall system-level behavior. Localization involves mapping the component functions produced by a candidate decomposition onto relatively circumscribed component parts of the system and their activities. The joint application of strategies of decomposition and localization is central to many canonical examples of mechanistic explanation.

However, application of these strategies depends on the model in question being one that makes determinate claims about the localization of components in the first place. Not all models that display componential organization need to do this. Cognitive models, in particular, are not committed to any particular spatial organization of their

components.[10] Verbally described processing models and mathematical models are most obviously neutral on this point, but even diagrammatic models are typically compatible with many possible spatial or geometric configurations of the physical structures that realize their functional components. This is true not only when they are viewed as systems-level decompositions, but even more so when we begin to unpack the various processing stages that each subsystem implements. Even if a particular functional subsystem can be localized, it is highly unlikely that each distinct inferential stage or representational transformation that it undergoes can be.

This fact about cognitive models is often obscured by their similarity to mechanistic models, particularly when both are presented in visual or diagrammatic form. Visual representations of mechanisms often *use* space in order to *represent* space.[11] In these, "diagrams exhibit spatial relations and structural features of the entities in the mechanism" (Machamer, Darden, & Craver, 2000, p. 8; see also Bechtel & Abrahamsen, 2005, p. 428). Thus in a cross-sectional diagram of the synaptic terminal of an axon, the shape of the perimeter is roughly the shape of an idealized or 'average' terminal, the placement of intracellular structures reflects their proximity, the width of the synaptic gap is scaled to represent the relative distance between the neurons, etc. Size, scale, and location also matter in other mechanistic models, such as those describing how voltage-gated ion channels embedded in cell membranes open and close. Here the particular spatial configuration of the molecular components of the channels is essential to their

[10] See Weiskopf (2011a), pp. 332-4 for further argument on this point.
[11] This claim strikes me as clearly true of the paradigmatic mechanistic models discussed in the literature, certainly those that are drawn from cell biology and neurophysiology. Mechanisms may be described in other ways, including non-diagrammatic ones, but as we have seen this is also true of cognitive models, and it is the resemblance between these diagrammatic representations that encourages confusion between the two.

correct operation, and, importantly, this organization is reflected in their standard depiction. The same points can be made about exploded view diagrams and the zoomed-in side views used to display mereological relationships, such as how entities of different sizes may be 'nested' within each other.

In diagrammatic cognitive models such as Baddeley's working memory model (see again Figure 1), spatial relations in the representation itself need not map onto those in the target system. The length of arrows connecting boxes is irrelevant; all that matters is their directional connectivity, weight, function, etc. Similarly, the boxes themselves are represented by arbitrary shapes, whereas the particular shapes of entities in mechanistic models matters a great deal to their function. The same indifference to these characteristics is iterated at the levels of representations and processes as well; notoriously, symbolic objects and their formal properties need not *resemble* neural structures. Therefore, many of the structures posited in cognitive models lack the characteristic properties of mechanistic entities, which "often must be appropriately located, structured, and oriented" (Machamer, Darden, & Craver, 2000, p. 3).[12]

Support for the possibility of models that display this sort of spatial neutrality goes back to Herbert Simon's pioneering work on complex systems (Simon, 1996). Simon's notion of a complex system can be understood in at least two different ways. One way sees hierarchies *mereologically*, in terms of size and spatial containment relations, so that a system is decomposed into subsystems that are literally physically parts of it. Putting mereological hierarchies at the center of the notion of a complex

[12] Further, they later say: "Traditionally one identifies and individuates entities in terms of their properties and spatiotemporal location. Activities, likewise, may be identified and individuated by their spatiotemporal location" (p. 5). This repeated emphasis suggests strongly that spatial organization is central to the notion of mechanism that is prevalent in these early conceptions.

system leads naturally to the mechanist conception, since mechanistic levels themselves are partially specified in these terms, and the spatial boundaries of a mechanism are drawn around all and only the entities that account for its performance.

An alternative, however, is to define hierarchies in terms of the *interactional strength* of various components rather than their spatial relations (Haugeland, 1998). Distinguishing social from physical and biological hierarchies, Simon writes: "we propose to identify social hierarchies not by observing who lives close to whom but by observing who interacts with whom" (1996, p. 187). On this view, the boundaries of systems are determined by elements that are maximally coupled with one another, i.e., capable of frequent reliable dynamical interactions involving the flow of information and control. These are then assembled into larger elements and systems, which are bound together by further interactional relations, all the way to the top level of organization.

As Simon notes, spatial and interactional hierarchies are often related, but this pairing is at best contingent. While strongly interacting elements may be spatially contiguous, they need not be, and spatially proximate elements need not interact with each other. And since strength of interaction or influence is purely functionally defined, there is no requirement that an interactional hierarchy have any particular spatial organization, although of course it must have one of some sort in order to generate its stable functioning and the effective interactions that define it. What ties an interactional hierarchy together is the existence and strength of the causal relations among a set of elements, specifically the causal relations that support and explain the behavior of the system that is of interest to us.

This point is clear from considering the variety of systems that can manifest these hierarchies. Simon's examples of hierarchical complex systems include subsystems of the economy (those involving the production and consumption of goods), as well as various social institutions (families, tribes, states, etc.). These plainly bear no necessary spatial relationships to each other at all. Models in economics and finance offer numerous other instances, as shown by Kuorikoski (2008).[13] Consider the role of central banks in the financial system. As social institutions, central banks have effects on money markets, auctions, and regulative legislation; and in virtue of playing these roles they can do such things as selling government securities to commercial banks, setting the rate at which commercial banks can borrow, and adjusting the commercial banks' ratio of reserves to loans. All of these interactions affect the overall operation of the financial system, and as such can potentially be exploited by policymakers in designing interventions. But understanding how central banks work does not involve asking localization questions; indeed, for most cases involving component parts such as social institutions or markets, it is questionable whether localization even makes sense in principle—'markets', after all, ceased to be exclusively physical spaces long ago.

So there are many interactional systems that are highly resistant to functional localization. Cognitive models represent such systems: they are defined in terms of the functional coupling of their components, but are, considered in themselves, neutral on issues of spatial organization and the shapes of components. Cognitive models do capture a certain kind of *causal* structure. But this causal structure is modeled in ways that

[13] The example which follows is taken and slightly simplified from Kuorikoski; however, while he points out that the 'parts' in an economy are either massively distributed or bear spatial relations to each other that are essentially inscrutable, he still interprets this case as still being mechanistic. The larger point of his discussion, however, is that there are systems that *appear* mechanistic but which involve merely capturing the abstract form of the causal interactions in a system. This dovetails with the moral of the present section.

involve abstracting away many or most aspects of the physical, biological, and neural architecture that support it. These models say nothing about how this causal structure is implemented in actual underlying components and activities, and their explanatory force does not turn on such details. Mechanistic models appeal to parts, their activities, and their structure to explain a system's capacities. So if it is a requirement on being mechanistic that a model be committal about spatial or structural facts, these models will not qualify. The paradigmatic mechanistic models—the one that fix our understanding of this otherwise generic metaphysical notion—are those that themselves display the relevant spatial and temporal organization of more or less localized entities.[14]

It is in this sense that cognitive models are taxonomically autonomous: the functional divisions they impose on a system may be only loosely related to its underlying physical organization. An example discussed by P&C is the implementation of the functional distinction between beliefs and desires (p. 303). There are many possible ways to ensure that these states have distinct functional roles: one appeals to separate memory stores and processes, while another allows them to co-mingle in a single store but gives them different 'attitude tags' that assign them their typical functional roles. These are distinct realization possibilities; however (*contra* what P&C suggest) the example seems to illustrate the extremely *loose* relationship between a functional classification and its implementation, rather than any form of direct constraint.

Mechanists have responded to the possibility of nonlocalized complex systems by broadening their notion of a 'part'. Structural components, say P&C, are not necessarily spatially localizable, single-function, or "stable and unchanging": "a structural

[14] I refer here to the *paradigm* cases because, as Bechtel and Richardson (1993) point out, there is a range of cases that gradually loosen these assumptions about spatial and temporal organization.

component might be so distributed and diffuse as to defy tidy structural description, though it no doubt has one if we had the time, knowledge, and patience to formulate it" (p. 291). This strategy carries risks, however, since it appears to verge on giving up not just localization, but any requirement that parts be describable in a way that our modeling techniques can capture.[15] And this, in turn, seems to strip the mechanistic program of any substantial commitment concerning the distinctive ontology of mechanisms. Remember, as initially presented the strategy was never intended to apply to *all* complex physical systems: there were clear exit points where the heuristics of decomposition and localization broke down. If we give up localization it is no longer clear whether there is any such thing as a complex physical system that is *not* subsumable under the mechanistic program.[16]

5. Against sketches

Even if psychological models *are* mechanistic in form, it doesn't follow that they must be mechanism *sketches*. For ease of exposition in what follows, I will sometimes concessively talk as if cognitive models are mechanistic models. The real question is why

[15] This acceptance of parts that are so non-localized as to elude structural description also sits poorly with the claim that we should aim for ideally complete models that capture maximal amounts of causally relevant detail. If these parts cannot be captured by the descriptive resources we have available, these explanations seem inaccessible to us. So broadening the notion of a mechanism may have costs in terms of our ability to satisfy mechanistic explanatory norms themselves.

[16] A similar point is made in more sweeping fashion by Laura Franklin-Hall (ms.). She argues that mechanists have not said in a systematic and principled way what sorts of causal relations mechanisms contain or what counts as a genuine part of a mechanism. Without some way of fleshing out these abstract ontological categories, the notion of a mechanistic explanation remains in substantial respects a promissory note. John Campbell (2008) lodges a similar complaint, noting that the term 'mechanism' has had little specific content outside of particular historical periods and disciplines, and that treating the search for mechanisms as a general goal of scientific inquiry is misguided: "You can, of course, evacuate content from the notion of 'mechanism' and say that although there was not the kind of mechanism they expected, there was nonetheless some other kind of mechanism at work. And of course there is no point in disputing that, since the claim lacks any definite meaning" (p. 430).

they can't be *fully adequate* explanatory models, not in need of further filling-in using the various modeling tools of neuroscience.

Mechanistic models are classified according to whether they are sketches, schemata, or ideally complete models (Craver, 2006; Craver, 2007; Machamer, Darden, & Craver, 2000). The distinction is a measure of the representational accuracy of the model: a "sketch is an abstraction for which bottom out entities and activities cannot (yet) be supplied or which contains gaps in its stages. The productive continuity from one stage to the next has missing pieces, black boxes, which we do not yet know how to fill in" (Machamer, Darden, & Craver, 2000). This is not a simple continuum, however, since the notion of accuracy includes separate uncorrelated factors such as the degree to which a model abstracts away from particular details or makes use of generic 'filler' concepts, the degree to which it includes false components, the significance of these omissions or inclusions, and so on (Gervais & Weber, 2013; Weiskopf, 2011a, pp. 316-7). To move from a sketch towards an ideally complete model is to progressively remove these various omissions, generalities, and inaccuracies, on the assumption that this will result in greater predictive or explanatory power, or improved skill at intervening in the system.

In light of this definition, the claim that cognitive models *invariably and as a class* are mere sketches is suspicious on its face. It amounts to saying that no cognitive model can be ideally complete and accurate with respect how it represents a system's psychological structures and properties. We can admit that most, perhaps all, of our current best cognitive models are sketchy. This is especially true of verbally formulated models, which often give only the rough qualitative contours of the processes they represent. Mathematical models may sometimes offer precise predictions and ways of

tracking complex relationships among psychological variables, but they are often silent on systems-level facts about cognitive architecture. Diagrammatic models themselves omit many details. A box-and-arrow decomposition of a system that gives us a rough assignment of functions to subsystems may give us no clue about the detailed inner organization of these boxes: what representational formats they use, what information they encode, what control processes there are, and so on. And even where we have a detailed model of a certain subsystem, we often have no notion how to embed it into a network of other systems.

So a certain degree of sketchiness is the *de facto* norm in psychology. Part of this is due to our ignorance of the correct structure of the cognitive system itself, but part is due to ordinary idealizations common to all modeling (Weisberg, 2007). The explanatory context rarely requires us to put all of these details into our models at once. The question is whether remedying this sketchiness requires stepping out of the explanatory framework of psychology. The argument against autonomy must establish that doing so is *necessary*: if psychological explanations can meet the appropriate explanatory norms on their own, this undermines the claim that they are mere sketches.

We first need to separate two ways in which mechanistic explanations can be elaborated on or improved:

> *Intralevel elaboration*: this involves staying at the same level of the mechanistic hierarchy, but making a model more detailed and precise, adding relevant components and activities, articulating their relations and structure, and so on.

Interlevel elaboration: this involves descending a level in the mechanistic hierarchy in order to explain the behavior of the various entities and operations in the system by appeal to a further set of components and activities.

Each of these is a different way of elaborating on a simple mechanistic model, but neither undermines the autonomy of cognitive modeling.

Intralevel elaboration requires getting rid of whatever filler terms, black boxes, fictions, unspecified entities, and generic processes that the initial model incorporated and replacing them with explicit specifications of the system's elements. The end result will be a model that is de-idealized, maximally specific, and wholly veridical. For instance, a black box might be filled in by giving a description of the precise computation that it carries out, or the stages of processing involved in its operation; or an abstract arrow connecting two boxes might be elaborated on by saying something about the information it carries and in what format.

Psychologists frequently try to do this, aiming not only to distinguish functional subsystems but also to describe the information they make use of, the nature of their internal databases and operations, and the formal character of the representations they manipulate. To claim that these attempts will always stall out at the stage of a mechanism sketch is effectively to say that cognitive modeling techniques are inherently inadequate to capturing the psychological properties and states of the target system (i.e., they cannot satisfy the ideal of representational accuracy).[17] This entails, for instance, that there can

[17] Bear in mind, again, that cognitive modeling may use information from any kind of study, including lesion, imaging, and electrophysiological studies, in confirming psychological hypotheses. Whether a model is accurate or not says nothing about the kind of evidence used to support it. The claim is just that the resulting cognitive models themselves can never capture the properties of the system with full accuracy.

be no cognitive model of working memory or object recognition that is ideally complete and that captures all of the relevant phenomena. But there is no reason to believe this strong claim. An ideally complete cognitive model will still be one that is couched in the autonomous theoretical vocabulary of psychology.

Interlevel elaboration, by contrast, involves descent to a further level of mechanistic analysis of a system.[18] So we might invoke ribosomes as the site of neuropeptide synthesis in one explanation, perhaps requiring only the information about their origin within the cell so that we can account for how they are transported to the synapse. For this purpose we may ignore precisely how they carry this process out, though we can if we wish change the context and descend to a lower level by treating the ribosome itself as a new target system and attempting to explain its operations. This might be relevant if we were trying to account for the rate at which neurons can regenerate depleted neuropeptides.

Interlevel elaboration is driven by a new set of explanatory demands: a novel set of phenomena (those associated with the activities of the system's components) demand explanations of their own, and so the mechanistic hierarchy gives rise to an associated "cascade of explanations" (Bechtel & Abrahamsen, 2005, p. 426). Interlevel moves may also involve shifting from the taxonomy and explanatory toolkit of one field to another, since shifts to lower levels may involve moving to spatiotemporal scales where different principles become dominant.

[18] Following the standard view among mechanistic philosophers, we do not need to assume any ordering of nature, entities, properties, or disciplines into anything like absolute levels here. All that is needed is a relative conception. Once we fix a particular analysis of a system, a lower level is defined by the fact that it invokes a decomposition of some component or operation of the system as initially analyzed.

There are explanatory insights to be gained from this sort of descent. However, to adequately explain a system's behavior we rarely *need* to continue this recursive descent through the hierarchy. A psychological capacity may be explained by appeal to a cognitive model that captures some of the relevant internal causal structure, as Baddeley's phonological loop accounts for word-length effects, phonological similarity effects, articulatory suppression, and so on (see section 6 for more details). What does not follow is that an explanation of a psychological capacity by appeal to a cognitive model *also* requires that we have a further set of lower-level explanations for how all of the elements of the model are implemented.

Explaining one thing in terms of another does not in general require recursive explanatory pursuit. This is clear when it comes to etiological causal explanations of events: in saying why a window broke it is often sufficient to cite the proximal cause, namely its being struck by a stone. It is unnecessary to trace every causal factor that was involved all the way back to the Big Bang. Similarly, componential causal explanations of this kind typically terminate at levels far above the fundamental.[19] If this were not so, explanations in all non-fundamental sciences, including neuroscience itself, would be just as sketchy and incomplete as those in psychology. Neurobiological systems are composed of a staggering array of nested mechanisms. In explaining a particular phenomenon we ignore most of these, however, descending only low enough to uncover

[19] Salmon (1984) distinguishes between etiological and constitutive causal explanations of phenomena. Etiological explanations account for a phenomenon's existence and properties in terms of the "causal story" leading up to its occurrence. Such explanations are historical. Constitutive explanations "account for a given phenomenon by providing a causal analysis of the phenomenon itself" (p. 297); his example is explaining the pressure a gas exerts on its container in terms of the momentum exchanged by its component molecules and the walls. Salmon's example of constitutive explanation has two possible readings. On one, the phenomenon of a gas having certain pressure is *identified with* a certain pattern of momentum exchange by its molecules. Om the other, a phenomenon displayed by a system is *causally explained by* the behaviors of its components. My term "componential" is meant to have the second reading.

the immediate structures that causally explain things at the grain of detail required. To insist that these are mere sketches insofar as they fail to capture the most fundamental mechanistic dependencies of the system is to place the bar for a fully adequate model far beyond our reach.[20]

Descent down the mechanistic hierarchy, then is constrained by the fact that explanations stop at the boundaries delimited by the interests and vocabulary of particular fields of inquiry, which contain a set of 'bottom-out' activities that they take as explanatory primitives. At this point, "explanation comes to an end, and description of lower-level mechanisms would be irrelevant" (Machamer, Darden, & Craver, p. 13). This anti-fundamentalist attitude is part of what distinguishes the program of multilevel integration from a classical reductive perspective committed to pursuing explanation in terms of ultimate or fundamental structures.

Mechanists might seek a middle ground position here, saying that while we should avoid fundamentalism, we should equally avoid stopping at the level of cognitive models. P&C suggest that "the search for mechanistic details is crucial to the process of sorting correct from incorrect functional explanations" (2011, p. 306). As noted earlier, however, the evidence for a cognitive model may come from anywhere, including from neuroscience. That does not compromise its explanatory autonomy. How we confirm an explanation is one thing, whether it is autonomous is another. Further, they say: "To accept as an explanation something that need not correspond with how the system is in

[20] Many aggregative idealizations fare far better as explanations than would anything pitched at a lower level or a finer grain of detail. The premise of continuum mechanics is that these idealized representations can be explanatorily effective over a broad domain, precisely because large collections of individual atoms or molecules effectively have the causal powers of continuous substances under the appropriate conditions. This treatment not only captures their causal organization, it does so more efficiently than would a finer-grained representation of the system.

fact implemented at lower levels is to accept that the explanations simply end at that point" (2011, p. 307). But *autonomous* explanations should not be confused with *ultimate* explanations. Psychological models can be sufficient for capturing the target phenomena without being themselves inexplicable. The point is merely that explaining how these models are implemented is a separate task from explaining how to capture the original phenomena in cognitive terms.

So neither of these two ways in which models can be sharpened suggests any principled limitation on how accurate cognitive models can be. They may be enriched so as to better capture the psychological capacities that are their target phenomena, or they may be integrated with lower-level implementation details. This may provide information about how the psychological and neurobiological aspects of the mind/brain fit together, and this in turn may improve our *overall understanding* of the system. Seeing how these models fit together and achieving multilevel integration is a genuine epistemic achievement, but we should not take the fact that we can increase our understanding of the total system through this kind of integration to show anything inherently defective or incomplete in the original cognitive model itself.

6. Autonomy and realism

Stepping back, the larger question posed by the autonomy of cognitive modeling has to do with whether these models are giving us insight into real structures and processes happening in the mind/brain. P&C's argument poses a dilemma for cognitive modelers: either psychological explanations are mechanistic, or they aren't. If they aren't, then the states, entities, and processes in psychology do not map onto real components of

the mind/brain. In this case, cognitive models cannot be offering causal explanations at all, since the only way to achieve a real causal explanation is to pick out the underlying constituents of a mechanism and track how their interactions produce the phenomena. If they are, on the other hand, they can only be regarded as mechanism sketches: incomplete or partial accounts of the underlying organization of the system. Integration would then take the form of fleshing out this partial sketch of the brain's mechanisms given by psychology, with the aim of producing a fuller and more adequate model as we descend down the mechanistic hierarchy to the neural level.

Consider the first possibility. So far I have been arguing that cognitive models are often non-mechanistic in form. Despite this, they still have explanatory force. Their status as explanations derives from the fact that they are able to capture facts about the causal structure of a system. The cognitive states, processes, resources, and other components that they represent are capable of interacting to produce the psychological phenomena that lie within the domain of the model. These facts about causal structure may be verbally described, captured in sets of equations, or schematized in diagrams. The causal patterns themselves are what is important, not the mode in which they are represented.

Explanations of how a system possesses and exercises a certain capacity typically make reference to the presence of some organized structures and processes that coordinate in the right way to produce the phenomena that are characteristic of the target capacity. Sometimes these patterns conform neatly to the stereotypical examples of mechanisms in neuroscience, biology, and certain branches of engineering. On the other hand, sometimes they don't, as in the case of many of the cognitive models described here. Both kinds of models, mechanistic and non-mechanistic alike, draw their

explanatory force from the same place, namely from the fact that they pick out casual structures and patterns that produce the relevant functions and capacities. So a componential but non-mechanistic cognitive model that represents some aspects of real causal structure in the domain of psychology ought to have just as much explanatory legitimacy as a mechanistic model does.

Now consider the second case. Even if cognitive models were mechanistic, they would still be more than mere sketches. They can be fleshed out and made as close to ideally complete as any other scientific model we know how to construct. It might still seem that if cognitive models were non-mechanistic that this would somehow undermine their reality. For example, P&C argue that task analysis must ultimately be a form of mechanistic explanation because "[i]f the connection between analyzing tasks and components is severed completely, then there is no clear sense in which the analyzing sub-capacities are aspects of the actual causal structure of the system as opposed to arbitrary partitions of the system's capacities or merely possible causal structures" (p. 293). The worry is that without some appeal to realization-level facts, we cannot distinguish between competing cognitive models, and will have no grounds for saying that any of them captures the true psychological structure of the target system.[21]

As a general point, it cannot be that a model captures some causal facts only when it maps onto a mechanism. All mechanistic explanations come to an end at some point, beyond which it becomes impossible to continue to find mechanisms to account for the

[21] There are obvious echoes here of earlier debates in cognitive science, most prominently the debate about whether natural language grammars have 'psychological reality' or not. In this debate, grammars were taken to be abstract mathematical objects, and the appeal to mental structures was meant to decide which of the many possible formally equivalent grammars captures real human linguistic competence. Analogously, the issue here is whether neural facts can help to decide which cognitive model captures the real psychological facts.

behavior of a system's components. The causal capacities of these entities will have to be explained otherwise than by their mechanistic organization. For example, consider protein folding, a process which starts with a mostly linear native state of a polypeptide and terminates with a complexly structured geometric shape. There does not appear to be any *mechanism* of this process: for many proteins, given the initially generated polypeptide chain and relatively normal surrounding conditions, folding takes place automatically, under the constraints of certain basic principles of economy. The very structure of the chain itself plus this array of physical laws and constraints shapes the final outcome. This seems to be a case in which complex forms are produced not by mechanisms but by a combination of structures and the natural forces or tendencies that govern them.

But in any case, the elements of cognitive models can meet any number of tests for mapping onto real entities (Weiskopf, 2011a): they have stable properties, they are robustly detectable using a range of theoretically independent methods, they can be manipulated and intervened on, and their existence can be demonstrated under regular, non-pathological conditions. These tests are applicable to representations (such as prototypes and analog mental images), processes (such as various forms of similarity matching), and resources (such as a limited capacity working memory or an attentional filter). The manipulation condition is particularly important, since psychological experiments often aim to isolate particular cognitive processes and representations and see what effects changing them has on behavior.

The elements of cognitive models may therefore constitute *control variables* for the behavior of the cognitive system (Campbell, 2008; 2009; 2010). In Campbell's sense,

we have a control variable for a system when: (1) there is a 'good' or natural-seeming function from the variable to the set of possible outcomes; (2) changes in the variable can make a large difference to the possible outcome; (3) these differences are largely specific to the particular outcome; and (4) there is a way of systematically manipulating or changing the value of the variable. These variables are aspects of a system that, when altered in a smooth fashion, allow us to choose among its various similarly ordered outcome states. Control variables in this sense are also robustly detectable and participate in a range of causal processes.[22]

If we adopt a metaphorical view on which the elements of cognitive models are akin to the dials, knobs, and levers of a control panel, then control variables are the ones "intervention on which makes large, specific, and systematic differences to the outcome in which we are interested, and for which can be specifically changed by actual physical processes" (Campbell, 2008, p. 433). And if our cognitive architecture has a sufficiently regular causal organization such that these conditions hold for its component elements and processes, then they will constitute control variables, and we may systematically affect both particular outcomes (thoughts and behaviors) and the overall functioning of the system by manipulating them.

The existence of cognitive control variables that hover at some remove from the neural organization of the mind/brain should be no surprise, since complex systems typically instantiate many different patterns of causal structure simultaneously. Consider the many ways neurons themselves can be causally cross-classified. As living cells they have a host of processes that involve genetic regulation of their activities. They also have

[22] Elsewhere I have argued that these robust, repeatable features of models that are employed in a wide range of explanations should be thought of as functional kinds (Weiskopf, 2011b).

mechanisms for producing action potentials and other graded potentials, and they have net metabolic demands that affect how they contribute to the local BOLD signal. Further mechanisms are involved in longer-term processes like synaptic and dendritic plasticity, directing the growth and pruning of these structures with use. Many of these mechanisms are interlocking and overlapping, but they are nevertheless *different* causal patterns co-present within the same system.

To see how cognitive models can provide representations of a system's psychological control variables, return for a moment to Baddeley's recent refinement of his model of working memory (WM). In its latest version, the model contains four component systems: the phonological loop, the visuospatial sketchpad, the episodic buffer, and the central executive (Baddeley, 2000; Baddeley, 2007; Baddeley, 2012; Repovš & Baddeley, 2006).[23] Whether the components of the modeled system constitute control variables depends on whether there are ways to specifically, systematically, and significantly activate, suppress, and modulate the behavior of the components of this system. The Baddeley-Hitch WM model is a particularly good test case to measure against these criteria, since it was explicitly developed in a data-driven fashion, meaning that components were added to the model on the basis of whether they could be experimentally modified in these ways.

Consider some canonical results bearing on the properties of the phonological loop. The loop itself is composed of two subsystems: a short-term store and an articulatory rehearsal process. The former is a limited capacity buffer, while the second is

[23] These are not regarded as the ultimate or final divisions of the system: the visuospatial sketchpad itself is now thought to fractionate into two subsystems, one for retaining visual images and the other for retaining spatial coding of information (Klauer & Zhao, 2004). On the role of neuropsychological case studies in confirming the model, see Vallar & Papagno (2002).

a control system that refreshes and sustains items within the store, and also is responsible for converting visually presented material into a subvocalized phonological code. In a typical working memory span task, participants must retain an ordered sequence of items such as a list of six numbers, letters, or words. Item span can be affected by the phonological similarity between the items, so that "man, cat, map, cab, can" will be harder to recall than "pit, day, cow, sup, pen." Semantic similarities among the items, however, have no effect on how easily they can be recalled (Baddeley, 1966). The fact that only certain types of confusion can occur in working memory suggests something about the code that the system uses. Selective modification of the phonological loop component of WM is possible by manipulating the to-be-remembered stimuli along specific dimensions of similarity, consistent with the control variable paradigm.

Material manipulations provide one way to influence cognitive processing. But components of the model can also be isolated using dual-task methods. When participants are asked to hold in memory a list of heard items while performing a concurrent articulatory task such as repeating a word, their performance tends to drop precipitously (Baddeley, Thomson, & Buchanan, 1975). This is predicted by the model, since articulatory processes are involved in maintaining information within the loop's storage system. Articulatory processes are also a gateway for non-auditory information to enter the phonological store. So disabling them should prevent visual information from being recoded in an auditory format. This seems to be the case: the phonological similarity effect disappears for visually presented items when participants perform an articulation task during encoding (Baddeley, Lewis, & Vallar, 1984). Manipulating task demands,

then, provides yet another causal lever for affecting the elements of the modeled system, specifically the articulatory control processes posited as part of the phonological loop.

Finally, the phonological loop can be activated in a more or less mandatory way by certain intrusive or irrelevant sounds. Participants asked to memorize visually presented digits do poorly when they are concurrently presented with speech sounds in an unfamiliar language, relative to conditions of silence or hearing white noise (Colle & Welsh, 1976; Salame & Baddeley, 1982). This effect seems not to be specific to speech sounds, but also to include other temporally patterned sounds such as fluctuating tones (Jones & Macken, 1993). The explanation for this disruption of performance is that certain irrelevant sounds gain automatic access to the phonological store, overwriting or interfering with its existing contents. So the phonological loop may have a mandatory access channel that selects for sound patterns that share abstract qualities of variability with normal speech.

These three effects (phonological similarity, articulatory suppression, and irrelevant speech) provide evidence that the phonological loop is a real construct that can be intervened on and manipulated experimentally. It can be activated (by irrelevant speech or speechlike sounds), manipulated (by phonologically related materials), and disrupted or deactivated (by articulatory suppression). These procedures have systematic and specific effects on performance in WM tasks that, according to the model, depend on the relevant properties of this subsystem. In these respects, the phonological loop as it is modeled here satisfies Campbell's conditions for being a psychological control variable.

Of course, none of this is meant as an endorsement of Baddeley and Hitch's model, since for present purposes I am less interested in the structure of working memory

itself than I am in what the construction of working memory models can tell us about cognitive modeling practices in general.[25] What this relatively brief summary suggests is that multicomponent cognitive models can contain functionally characterized elements that may be manipulated to produce systematic effects on the phenomena in their domain. While it may be informative to ask how these elements relate to neural structures and processes, having this knowledge is not necessary for cognitive models themselves to be explanatory.

7. Conclusion

The challenge to the autonomy of cognitive modeling that I have surveyed has two parts. Against the idea that that cognitive modeling is a form of mechanistic explanation, I've argued that it is a way of capturing the causal organization of a psychological system by representing it in terms of abstract relationships among functional components. This is a kind of componential causal explanation, but one that has important differences from mechanistic modeling. Further, cognitive models are capable of giving explanations of their target phenomena that answer to all of the relevant epistemic norms and standards, and they achieve this without making essential reference to the details of those models' neural implementation. A total understanding of the mind/brain will involve both perfecting such cognitive models and coordinating them with neurobiological ones. But this is not in conflict with the autonomist claim that some explanations of our psychological capacities come to an end within psychology itself.

Acknowledgments

[25] For an alternative to Baddeley's perspective on working memory, see Postle (2006).

Thanks to Eric Winsberg and David M. Kaplan for helpful comments on an earlier draft of this paper, and to students and faculty at Georgetown University for a stimulating discussion of this material.

References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, 111, 1036-1060.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K.W. Spence (Ed.), *The Psychology of Learning and Motivation*, Vol. 2 (pp. 89-195). New York: Academic Press.

Baddeley, A. D. (1966). Short-term memory for word sequences as a function of acoustic, semantic, and formal similarity. *Quarterly Journal of Experimental Psychology*, 18, 362-365.

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, 4, 417-423.

Baddeley, A. D. (2007). *Working Memory, Thought and Action*. Oxford: Oxford University Press.

Baddeley, A. D. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, 63, 1-29.

Baddeley, A. D., & Hitch, G.J. (1974). Working memory. In G. A. Bower (Ed.), *Recent Advances in Learning and Motivation*, Vol. 8 (pp. 47-89). New York: Academic Press.

Baddeley, A. D., Lewis, V. J., & Vallar, G. (1984). Exploring the articulatory loop. *Quarterly Journal of Experimental Psychology*, 36, 233-252.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short term memory. *Journal of Verbal Learning and Verbal Behavior*, 14, 575-589.

Bechtel, W. (2008). *Mental Mechanisms*. New York: Routledge.

Bechtel, W. (2012). Reducing psychology while maintaining its autonomy. In M. Schouton & H. Looren de Jong (Eds.), *The Matter of the Mind* (pp. 172-198). Malden, MA: Blackwell.

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

Bechtel, W., & Richardson, R. C. (1993). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Princeton, NJ: Princeton University Press.

Busemeyer, J. R., & Diederich, A. (2009). *Cognitive Modeling*. Sage Publications.

Campbell, J. (2008). Interventionism, control variables, and causation in the qualitative world. *Philosophical Issues*, 18, 426-445.

Campbell, J. (2009). Control variables and mental causation. *Proceedings of the Aristotelian Society*, 110, 15-30.

Campbell, J. (2010). Independence of variables in mental causation. *Philosophical Issues*, 20, 64-79.

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

Chomsky, N. (2000). New horizons in the study of language. In *New Horizons in the Study of Language and Mind* (pp. 3-18). Cambridge: Cambridge University Press.

Clark, A. (1997). *Being There*. Cambridge: MIT Press.

Colle, H. A., & Welsh, A. (1976). Acoustic masking in primary memory. *Journal of Verbal Learning and Verbal Behavior*, 15, 17-32.

Coltheart, M. (2006). What has functional neuroimaging told us about the mind (so far)? *Cortex*, 42, 323-331.

Coltheart, M., Curtis, B., Atkins, P. & Haller, M. (1993). Models of reading aloud: Dual-route and parallel distributed processing approaches. *Psychological Review*, 100, 589-608.

Cooper, R., & Shallice, T. (1995). Soar and the case for unified theories of cognition. *Cognition*, 55, 115-149.

Craik, F. I. G., & Lockhart. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11, 671-684.

Craik, F. I. G., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, 104, 268-294.

Craik, L. S., & Craik, F. I. M. (Eds.) (1979). *Levels of Processing in Human Memory*. Hillsdale, NJ: Lawrence Erlbaum.

Craver, C. F. (2005). Beyond reduction: Mechanisms, multifield integration, and the unity of neuroscience. *Studies in the History and Philosophy of Biological and Biomedical Sciences*, 36, 373-395.

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

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

Darden, L., & Maull, N. (1977). Interfield theories. *Philosophy of Science*, 44, 43-64.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, 338, 1202-1205.

Fodor, J. A. (1974). Special sciences (or: the disunity of science as a working hypothesis). *Synthese*, 28, 97-115.

Franklin-Hall, L. (ms.). The emperor's new mechanisms. Retrieved June 7, 2012, from https://files.nyu.edu/lrf217/public/Laura_Franklin-Hall/Research_files/Franklin-Hall%20--%20The%20Emperors%20New%20Mechanisms.pdf.

Gaohua, L., & Kumura, H. (2009). A mathematical model of brain glucose homeostasis. *Theoretical Biology and Medical Modelling*, 6, 1-24.

Gervais, R., & Weber, E. (2013). Plausibility versus richness in mechanistic models. *Philosophical Psychology*, 26, 139-152.

Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science*, 69, S342-S353.

Gray, W. D. (Ed.) (2011). *Integrated Models of Cognitive Systems*. Oxford: Oxford University Press.

Haugeland, J. (1998). Mind embodied and embedded. In *Having Thought* (pp. 207-237). Cambridge: Harvard University Press.

Jacobs, A. M., & Grainger, J. (1994). Models of visual word recognition—Sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance*, 20, 1311-1334.

Johansen-Berg, H., & Rushworth, M. F. S. (2009). Using diffusion imaging to study human connectional anatomy. *Annual Reviews of Neuroscience*, 32, 75-94.

Jones, D. M., & Macken, W. J. (1993). Irrelevant tones produce an irrelevant speech effect: Implications for phonological coding in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 369-381.

Kaplan, D. M., & Craver, C. F. (2010). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science*, 78, 601-627.

Klauer, K. C., & Zhao, Z. (2004). Double dissociations in visual and spatial short-term memory. *Journal of Experimental Psychology: General*, 133, 355-381.

Kuorikoski, J. (2008). Two concepts of mechanism: Componential causal system and abstract form of interaction. *International Studies in the Philosophy of Science*, 23, 143-60.

Lewandowsky, S., & Farrell, S. (2007). *Computational Modeling in Cognition*. Sage Publications.

Lloyd, D. (2000). Terra cognita: From functional neuroimaging to the map of the mind. *Brain and Mind*, 1, 93-116.

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, 1, 11-38.

Morrison, M. (2011). One phenomenon, many models: Inconsistency and complimentarity. *Studies in History and Philosophy of Science*, 42, 342-351.

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

Parker, W. S. (2006). Understanding pluralism in climate modeling. *Foundations of Science*, 11, 349-368.

Piccinini, G., & Craver, C. F. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, 183, 283-311.

Postle, B. R. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience*, 139, 23-38.

Repovš, G., & Baddeley, A. D. (2006). The multi-component model of working memory: Explorations in experimental cognitive psychology. *Neuroscience*, 139, 5-21.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic Cognition: A Parallel Distributed Processing Approach*. Cambridge: MIT Press.

Roskies, A. (2009). Brain-mind and structure-function relationships: A methodological response to Coltheart. *Philosophy of Science*, 76, 927-939.

Salame, P., & Baddeley, A. D. (1982). Disruption of short-term memory by unattended speech: Implications for the structure of working memory. *Journal of Verbal Learning and Verbal Behavior*, 21, 150-164.

Salmon, W. (1984). Scientific explanation: Three basic conceptions. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1984 (Vol. 2), 293-305.

Shepherd, G. (2010). *Creating Modern Neuroscience: The Revolutionary 1950s*. Oxford: Oxford University Press.

Shiffrin, R. (2010). Perspectives on modeling in cognitive science. *Topics in Cognitive Science*, 2, 736-750.

Simon, H. (1996). *The Sciences of the Artificial* (3rd ed.). Cambridge: MIT Press.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, 327-352.

Vallar, G., & Papagno, C. (2002). Neuropsychological impairments of short-term memory. In A. D. Baddeley, M. D. Kopelman, & B. A. Wilson (Eds.), *The Handbook of Memory Disorders*, 2nd Ed. (pp. 249-270). West Sussex: John Wiley & Sons.

van Gelder, T. (1995). What might cognition be, if not computation? *Journal of Philosophy*, 92, 345-381.

Weisberg, M. (2007). Three kinds of idealization. *Journal of Philosophy*, 104, 639-659.

Weiskopf, D. A. (2011a). Models and mechanisms in psychological explanation. *Synthese*, 183, 313-338.

Weiskopf, D. A. (2011b). The functional unity of special science kinds. *British Journal for the Philosophy of Science*, 6, 233-258.

Woodward, J. (2013). Mechanistic explanation: Its scope and limits. *Proceedings of the Aristotelian Society*, 87, 39-65.

Zilles, K., & Amunts, K. (2009). Receptor mapping: architecture of the human cerebral cortex. *Current Opinion in Neurology*, 22, 331-339.