Synthese (2016) 193:1585–1614
DOI 10.1007/s11229-015-0871-5

S.I.: NEUROSCIENCE AND ITS PHILOSOPHY

Mechanisms in psychology: ripping nature at its seams

Catherine Stinson[1]

Received: 12 October 2013 / Accepted: 22 August 2015 / Published online: 3 September 2015

Abstract Recent extensions of mechanistic explanation into psychology suggest that cognitive models are only explanatory insofar as they map neatly onto, and serve as scaffolding for more detailed neural models. Filling in those neural details is what these accounts take the integration of cognitive psychology and neuroscience to mean, and they take this process to be seamless. Critics of this view have given up on cognitive models possibly explaining mechanistically in the course of arguing for cognitive models having explanatory value independent of how well they align with neural mechanisms. We can have things both ways, however. The problem with seamless integration accounts is their seamlessness, not that they take cognitive models to be mechanistic. A non-componential view of mechanisms allows for cognitive and neural models that cross cut one another, and for cognitive models that don't decompose into parts. I illustrate the inadequacy of seamless accounts of integration by contrasting how "filter" models of attention in psychology and of sodium channels in neuroscience developed; by questioning whether the mappings generated by neuroimaging subtraction studies achieve integration; and by reinterpreting the evidence for cognitive models of memory having been successfully integrated with neural models. I argue that the integrations we can realistically expect are more partial, patchy, and full of loose threads than the mosaic unity Craver describes.

Keywords Mechanism · Explanation · Integration · Cognitive psychology · Neuroscience

✉ Catherine Stinson
frau.dr.stinson@gmail.com

[1] Centre for Integrative Neuroscience, Eberhard Karls Universität Tubingen, Otfried-Müller-Str. 25, 72076 Tübingen, Germany

1 Introduction

Cognitive neuroscience is described as "The Biology of the Mind" (Gazzaniga et al. 2009). Its goal, in somewhat more technical terms, is to explain cognition in terms of neural mechanisms, and the approach it takes is usually described as 'integrative.' What exactly it might mean for cognitive and neural sciences to be integrated is not an easy question.[1] Integration is positioned as an alternative to reductionism though, so integration ought not be used as a code word for the colonization of one field by another. My aim here is to work towards an account of integration that respects this minimal criterion, but is also realistic given the evidence we have so far from cognitive neuroscience.

Two recent philosophical accounts describe how this integrative project might work, in both cases in terms of mechanisms. Piccinini and Craver (2011) argue that cognitive psychology and neuroscience can be "seamlessly" integrated by treating cognitive models as "elliptical mechanism sketches." Bechtel (2008) argues that the decomposition and localization heuristics that are so useful for discovering biological mechanisms scale up to cognitive science, and offers extended examples of how these heuristics have been used in the cognitive neuroscience of vision and memory.[2] These two accounts are similar in spirit, and both provide accurate descriptions in a limited sense. They describe a popular strategy of using the heuristic of decomposition and localization to try to integrate cognitive and neural models. Piccinini and Craver (2011) state that a cognitive model that fails to map onto neural mechanisms should be rejected or revised. Bechtel (2008) describes a representative case study in which this occurs, although does not explicitly endorse this claim. I argue that raising a fallible heuristic to the status of normative constraint on cognitive models is inappropriate. If this integrative project is to be successful, it needs to pursue a less seamless approach that acknowledges the ways that the decomposition and localization heuristic may legitimately fail. In these cases we may end up with cognitive models which do not decompose and localize into neural working parts, yet are nevertheless good models worth keeping.

1.1 Outline

In Sect. 2 I introduce the accounts of integration described in Bechtel (2008) and Piccinini and Craver (2011), and point out some of the methodological and metaphysical assumptions they make.

In Sect. 3 I question whether cognitive models can be fairly treated as elliptical mechanism sketches waiting to be filled in with neural details. I contrast two examples of model development: one from neuroscience, the other from cognitive psychology. Piccinini and Craver's account suggests that cognitive models should develop from

[1] Integration has several other meanings, but here I restrict my use of it to this project of constructing models that combine knowledge from models at different 'levels,' in this case cognitive and neural.

[2] Bechtel uses the term 'integration' in a broader sense that also includes relationships between scientists, institutions, instruments, experimental protocols, etc., in his work on integration (Bechtel 1986, 1993).

elliptical mechanism sketches into gradually more complete models that describe the details underlying the initial sketch. In my example from neuroscience, sodium channel gating, exactly the sort of progression they describe has been occurring. This isn't how cognitive models typically develop, however. In my example from cognitive psychology, the attention filter, several decades of development and improvement occurred, without any filling in of underlying details. Contemporary models of the attention filter combine knowledge from various levels, but not in the way Piccinini and Craver describe.

In Sect. 4 I question the assumption that for something to be a mechanism, it must be possible to decompose and localize it in lower-level parts. Even if one argued that cognitive models like the attention filter are incomplete as mechanistic models, it does not follow that integration must mean localizing their functional parts in brain regions, nor that they can't be mechanistic models if this fails. I note that the decomposition and localization heuristic is just that, a heuristic, and thus can't serve as a normative constraint on mechanistic models. I take up an example Bechtel offers as a success case of using the decomposition and localization heuristic to integrate a cognitive model with results from neuroscience. I argue that given the evidence Bechtel offers, this could just as easily be construed as an example of the heuristic's failure. Early cognitive models of memory, rather than aiding discovery by providing a mechanism sketch, were rather misleading guides to the discovery of neural mechanisms. More importantly, cognitive models of memory have not been integrated with neural models in the sense of having their functional parts localized. The cognitive model is incompatible in several ways with current neural models, yet it has not been rejected or significantly altered based on this mismatch. Both models are maintained, and are used to explain different things.

In Sect. 5 I suggest a less seamless, but more realistic picture of integration, in which cognitive models may count as mechanistic in their own right, even when they fail to connect tidily to a hierarchy of neural mechanisms. I describe a picture where explanations in the cognitive and neural sciences fit together not like the tiles of Craver's mosaic, but instead like a garment held together by many overlapping patches.

2 Accounts of integration

Among philosophers of neuroscience, there are widely divergent views about the prospects for integration. Craver (2007) paints an optimistic picture of "mosaic unity" achieved through multilevel mechanistic explanation. His basic story is that all the knowledge we gain about the various levels, from systems to circuits, down to cells and molecules will eventually all be pieced together into a hierarchy of mechanisms. Sullivan (2009) argues that Craver's picture of mosaic unity is unrealistic, given the difficulty in getting results even of experiments with nearly identical protocols to fit neatly together. Revonsuo claims that "there is a clear conflict of explanatory strategies and assumptions built into the ingredients of cognitive neuroscience" (Revonsuo 2001). Bechtel, in stark contrast, claims that integration is unproblematic, because "a common explanatory framework [is] employed in both the cognitive sciences and

the neurosciences—the framework of mechanistic explanation" (Bechtel 2001). In the remainder of this section, I describe recent accounts of integration in more detail. I begin with a brief look at mechanisms.

2.1 Mechanisms

Very broadly, mechanistic explanations show how things happen by referring to the proximate physical causes that bring them about, typically in terms of the coordinated actions of various parts. This is in contrast to forms of explanation that refer to historical patterns, reasons, laws, etc. For a thorough overview of accounts of what a mechanism is exactly, and how mechanistic explanations operate, see Andersen (2014a, b).

For our purposes here, just a few details of 'new mechanist' accounts are most relevant. Bechtel and Richardson (1993) define mechanisms in terms of parts and their organization. A more recent paper defines a mechanism as "a structure performing a function in virtue of its component parts, component operations, and their organization" (Bechtel and Abrahamsen 2005, p. 423). Glennan (1996) defined mechanisms in terms of interactions between parts according to causal laws, but later replaced laws with "invariant change-relating generalizations" (Glennan 2002). Glennan's interactions involve changes in properties. Machamer et al. (2000) [hereafter MDC] define mechanisms as "entities and activities organized such that they are productive of regular changes." Some key differences between these accounts are that MDC's definition allows for non-componential entities, such as wholes, as explainers; MDC insist on mechanisms acting linearly from start to finish, while Bechtel and Abrahamsen (2005) explicitly allow for cyclical mechanisms; and MDC give activities ontologically equal status to entities, rather than referring to properties of parts as Glennan does. Craver's view combines elements of Bechtel's and MDC's. He says mechanisms explain phenomena in terms of the "organization of component entities and activities" (Craver 2007, p. 8).

Bechtel and MDC both also describe the dynamics of mechanism discovery. Bechtel and Richardson (1993) introduce the decomposition and localization heuristic, which is further developed in Bechtel and Abrahamsen (2005). In this heuristic, functional decomposition is a top-down strategy where you "start with the overall functioning or behavior of the mechanism system and figure out what lower-level operations contribute to achieving it" (Bechtel and Abrahamsen 2005). The complementary strategy of structural decomposition seeks to decompose the system into working parts that "perform the operations that figure in the functional decomposition" (Bechtel and Abrahamsen 2005). The parts and functions arrived at through these two sorts of decomposition should eventually fit together for a successful mechanistic explanation. Localization is the linking together of working parts from a structural decomposition with the operations from a functional decomposition by identifying the operations as those that the parts are supposed to perform (Bechtel and Richardson 1993, p. 24).

According to MDC, mechanism discovery typically begins with a mechanism *sketch*, which is a gappy representation of a mechanism where "bottom out entities and activities" are not yet known (Machamer et al. 2000). Sketches are gradually filled in with details of the entities and activities, and corrected as necessary, until a

more complete picture is known. In this account there is no separation between the search for entities and activities. (In Bechtel's picture the two types of decomposition needn't be pursued separately, despite being distinguished in principle.) MDC's notion of mechanism *schema* is sometimes interpreted as being either a more complete sort of sketch at a later stage in mechanism discovery, or a representation of a fully understood mechanism that leaves out some details.[3] My understanding of MDC's mechanism schemata is that they can play either of two roles: abstractly representing more-or-less fully understood mechanisms, or representing types of mechanisms that one might find instantiated in various contexts.

Critical to the issue of integration is the notion of levels. Bechtel construes "nature as incorporating a hierarchy of levels" (Bechtel and Richardson 2010, p. 27). The decomposition and localization heuristic begins with the assumption that a system is hierarchically organized and decomposable. Bechtel and Richardson discuss cases where the decomposability assumption turns out to be false, and cases where the hierarchy is flat, but not where there fails to be a hierarchy. Craver's is "the leanest account of levels that will suffice for mechanistic models," according to Bechtel and Richardson (2010, p. xxix). Craver (2007) distinguishes levels of science, levels of nature, and levels of mechanisms. He describes levels of mechanisms as "levels of composition, but…the relata are behaving mechanisms at higher levels and their components at lower levels" (Craver 2007, p. 189). Which level an entity is located at depends on the mechanism under consideration. A hydrogen ion, for example, might in some contexts be at the same level as other ions, but in other contexts alongside much larger structures like amino acids or membranes. Machamer and Sullivan (2001) are skeptical of there being any notion of levels that is helpful in understanding scientific explanation. They point out that within a single mechanism it is not always clear how to assign levels to the entities involved. Hydrogen ions, for example, might be involved in several distinct activities within a mechanism, in which they act at different 'levels.'

2.2 Seamless integration

Despite these differences in how mechanisms and levels are understood, there is much in common between the views of integration described by Bechtel, and Piccinini and Craver.

Piccinini and Craver (2011) argue that cognitive psychology and neuroscience can be "seamlessly integrated with multilevel mechanistic explanations." They break down explanation in cognitive psychology into three types—task analysis, functional analysis by internal states, and boxology—then argue, for each type, that "properly constrained" these amount to sketches of neural mechanisms, as defined by MDC. Integration would then proceed by filling in the sketches that cognitive psychology provides. Mechanism sketches should gradually be filled in with plausible details and accrue evidential support, until they are adequately complete. Mature models would specify the neural sub-mechanisms that explain the higher-level components.

[3] I think both are misunderstandings. See Stinson (2013).

Cognitive psychology's models typically don't describe specific entities or activities, but rather a functional organization that might be implemented any number of ways. Nevertheless, Piccinini and Craver count these as "elliptical" mechanism sketches: bare outlines into which mechanistic models might be fitted. Their argument relies heavily on the proviso that explanations be "properly constrained." What this constraint amounts to is that psychologists "ought to acknowledge that psychological explanations describe aspects of the same multilevel neural mechanism that neuroscientists study. Thus, psychologists ought to let knowledge of neural mechanisms constrain their hypotheses" (Piccinini and Craver 2011). They make a metaphysical assumption that cognitive and neural mechanisms must be part of the same hierarchy. They also make a corollary methodological assumption that if there is a shared hierarchy, neural and cognitive models ought to place useful constraints on one another such that each should be able to guide attempts to discover the other. I call these assumptions *metaphysical seamlessness* and *methodological seamlessness*, and critique both in later sections.[4]

Bechtel (2008) uses the language of "linking" rather than integration, but his account is similar. The main strategy for linking our understanding of the mind and brain, according to Bechtel, is to localize mental components in neural ones. In this account, cognitive models featuring information-processing operations and the parts that perform them are the result of functional decompositions. Neural operations and parts are yielded by structural decompositions. The two are linked when the information-processing operations and parts are localized in neural operations and parts.

Bechtel's account makes roughly the same assumptions of seamlessness that Piccinini and Craver's does. First, it supposes that cognitive functions should be localizable in neural working parts, which assumes a shared structure between cognitive and neural parts (metaphysical seamlessness). Second, it assumes that the functional decompositions yielding cognitive models ought to be a useful heuristic for discovering neural mechanisms, and vice versa (methodological seamlessness). For Bechtel these assumptions are described as heuristics rather than norms, but he does not mention any cases where the heuristic fails.

These seamlessness assumptions are popular, and for good reason. We want to believe that nature is orderly and that we will eventually make sense of it. Piccinini and Craver, and Bechtel provide good accounts of integration in the sense that they accurately portray the typical aims and expectations of cognitive neuroscientists. The danger in seeing these aims and expectations as criteria that ought to be satisfied if we're doing science well, however, is that we may be led to reject good models on the grounds that they don't fit these idealized norms. I think we have every reason to expect models to combine in untidy ways, however. Integration doesn't have to mean a unified hierarchy or methods that are perfectly complimentary. Quite often science turns out to be more complicated than pioneers in new fields first expect.

[4] Weiskopf (2011) has also critiqued Piccinini and Craver's account of integration. I'm in agreement with many of his points, including that cognitive explanations may be non-componential, but unlike him, I do not want to claim that these explanations cannot also be mechanistic. I discuss this in Sect. 5.

3 Methodological seams

We will now look at how models in cognitive psychology and neuroscience actually develop, to see whether these accounts of integration are realistic. Research in neuroscience often does begin with a mechanism sketch and proceed to gradually fill in that sketch with further details. I first give an example of that pattern of events unfolding in the developments of models of sodium channel gating. I then give an example of a cognitive model that uses the same mechanical metaphors of channels, filters and gates. The development of the attention filter model, however, does not follow the pattern of gradually filling in a mechanism sketch. I argue that cognitive models are more than just templates into which neural details can be fit, although they might be pressed into service as such. I conclude that using cognitive models as mechanism sketches should not be expected to run seamlessly.

3.1 Neuroscience's filters

In neurophysiology the pores that allow certain ions to pass through the membranes of axons, while blocking other ions from passing through, are described in terms of channels, filters, and gates. The sodium channel, for example, which is essential for the action potential, has a selectivity filter and two gates. The selectivity filter allows Na^+ ions through but not K^+ ions. Likewise, potassium channels are selective for K^+ ions. The mechanism by which it accomplishes this involves a pore lined with negatively charged amino acids, which attract positively charged ions like Na^+ and K^+ (Hille 2001), and most likely a configuration change occurs within the pore when Na^+ interacts with glutamate, breaking its bond with lysine (Doyle et al. 1998; Lipkind and Fozzard 2008).[5] During the neuron's resting state Na^+ ions are prevented from flowing inward because of the activation gate. When the cell is stimulated, for instance during an action potential, the activation gate opens, and Na^+ flows in. The flow of ions is once again blocked by the subsequent closing of the inactivation gate, which happens a few milliseconds after depolarization (stimulation which makes the usually negatively charged neuron less negative).

The channel is made of an assembly of proteins embedded in the cell membrane, and the gates are made of loops of this protein. Ion channel proteins have sections that are variously charged, like the negatively charged pore region mentioned above, and which react to various neurochemicals. This means that when the membrane changes its charge, or when the cell encounters certain chemicals, the proteins change shape. The sodium channel's inactivation gate is a loop of protein that flaps open or closed based on changes in the neuron's electrical charge. Other types of ion channels open and close based on the presence of particular chemicals, temperature changes, or physical force, and they use various mechanisms for changing the shape and size of the pore. The $nAChR$ channel opens by rotating its helices (sections of protein forming the channel). The K^+ channel opens by bending its inner helices on a hinge point. The $MscL$ channel opens by tilting its helices (Doyle 2004).

[5] Thanks to an anonymous reviewer for pointing me towards the latest research.

From these descriptions it should be clear that these filters and channels are quite literally filters and channels. The filters are physical structures that let some objects but not others through an opening based on physical properties like size and shape. Hille (2001) describes the channel pore as "an atomic mechanical sieve." In addition, the gates are physical structures that swing, tilt or twist open and closed over an opening allowing or preventing objects to pass through. Much is known about the constitution of the proteins forming these structures, and the ways they change shape and react to various stimuli, based on methods like X-ray crystallography. Neurophysiologists are continuing to fill in more of the details of how these mechanisms work.

But even before so much was known about these structural and dynamic details, the language of channels, filters and gates was used with the expectation that structures corresponding to these names would be found there. Hille (2001) reproduces progressively more complete schematics of the Na^+ channel from 1977 onward. In the earliest version, the gate (it wasn't yet known that Na^+ channels have two gates) is represented as an amorphous blob with dotted outline, and the voltage sensor is a simple box with a probe sticking into the "Channel macromolecule." Shepherd also includes a schematic diagram of an Na^+ channel in his 1983 textbook, which looks very similar to Hille's. In Shepherd's version, the filter, gate and sensor labels are even in scare quotes, since there was at the time only speculation about the existence of entities performing approximately those functions, but no clear evidence about their physical natures. A composite of these two schematics is shown in Fig. 1.

Although many of the details of how Na^+ channels work were not yet known, the more speculative parts like the gate were included only in a provisional way. Rather than resting content with the idea that the gate should be something that changes state depending on electrical and chemical conditions, thus allowing or preventing the flow of ions, a number of specific hypotheses were put forward as to how the gate worked. Hille (2001) notes, in *Ion Channels of Excitable Membranes*, "Many models have been proposed for the nature of gates." He provides illustrations of twelve such possible mechanisms (how-possibly models) that were suggested in various published articles.

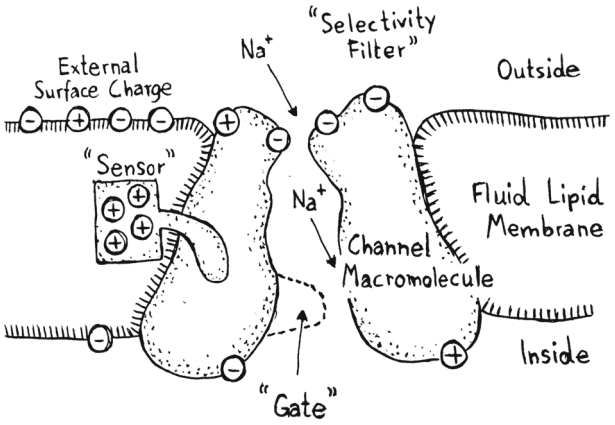

Fig. 1 Early schematic of the Na^+ channel, based on Shepherd (1983) and Hille (2001). Copyright ©2013 Boris Hennig, used with permission

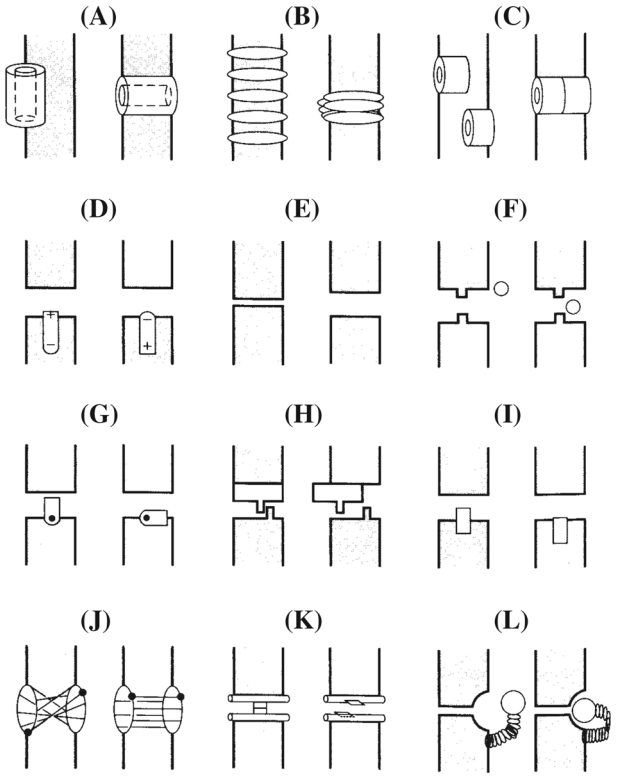

Fig. 2 Possible mechanisms for channel gating. Reprinted from Hille (2001). Copyright ©Sinauer Associates

These include gates that swing out, assemble from subunits, pinch shut, are blocked by mobile ions, rotate, slide, twist, or are plugged by a tethered ball and chain. His illustration is reproduced here as Fig. 2. Hille describes hypotheses F through L as the mechanisms that remain popular and plausible (how-plausibly models), and discusses the evidence for these mechanisms in various types of ion channels. I described a few examples of these above.

Hille's later diagram of the Na^+ channel is superficially similar, but has a number of added details, based on results that had accrued in the intervening decade. In the 1991 version, the macromolecule is called a protein, it sticks much further out into inter- and extra-cellular space, various molecules are attached to the protein, and the membrane layer and voltage sensors are much more detailed, although the gate is still represented as a schematic swinging hinge, connected to the voltage sensor.

Since 1991 it has been determined that the inactivation gate of the Na^+ channel works like a tethered ball and chain, as in Hille's hypothesis L. It is formed by the section of protein between domains III and IV (Hille 2001). Figure 3 shows a state transition diagram illustrating the Na^+ inactivation gate's opening and closing. Contemporary diagrams of the Na^+ gate show many more details, down to the twists and turns in the channel proteins.

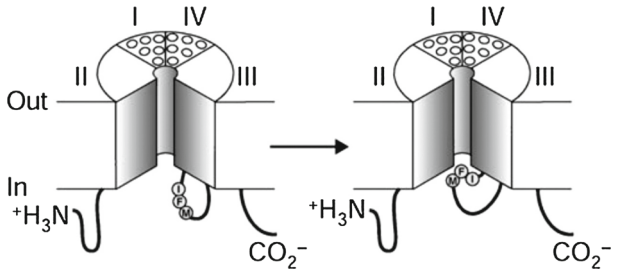

Fig. 3 Sodium channel inactivation gate. Adapted from Yu and Catterall (2003). Copyright ©2003 BioMed Central Ltd.

The development of these models from sketchy schematic drawings through to increasingly detailed models fits well with the description of mechanistic explanation MDC give, and as indicated, they progress from how-possibly to how-plausibly, then how-actually models, just as they describe. It is striking how the initial mechanism sketch was gradually filled in with more details, in precisely the way Piccinini and Craver (2011) suggest should happen with cognitive models. The diagrams of ion channels in Shepherd (1983) and Hille (2001) are perfect examples of mechanism sketches. The gate and sensors shown in dotted outline, and with scare quotes in Shepherd, began as black boxes. The discovery that the Na^+ channel has both an activation and an inactivation gate filled in more details. Hille's (2001) collection of gating mechanisms were how-possibly models: each a generic description of a possible gating mechanism, the details of which might later be further filled in and evaluated. As the structure of the proteins making up these structures is discovered, they are gradually becoming how-actually models.

In this case from neuroscience, a gradual filling-in of a mechanism sketch clearly did occur. It should be unsurprising that accounts of mechanisms work well here, since neurobiology was one of the main fields that inspired the resurgence in interest in mechanistic explanation. Indeed in the preface to Shepherd's classic (1983) textbook, *Neurobiology*, the framework of hierarchical mechanisms is described. He says, "any given region or system contains successive levels of organization, beginning with ions and molecules, and building up through cells and their circuits to behavior" (Shepherd 1983, p. viii). *Integrating* levels from cellular and molecular up to cognition is also mentioned as a goal: "Many workers in recent years have studied synaptic properties and circuits and their correlations with simple behaviors; what is still needed is an understanding of how, beginning at the single synapse, one builds up successive levels of synaptic circuits of increasing extent to mediate complex naturally occurring behaviors" (Shepherd 1983, p. ix).

3.2 Cognitive psychology's filters

I now turn to cognitive psychology to see whether comparable examples of mechanism sketches gradually being filled in with details occur there too. I begin with a classic example of a cognitive model, Broadbent's attention filter, which invokes mechanical metaphors of channels, filters and gates, just like models of the sodium channel. If

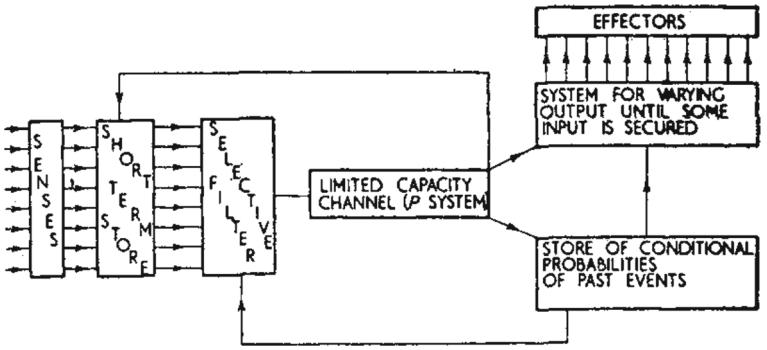

Fig. 4 Broadbent's information flow chart of the filter model of attention. Reprinted from Broadbent (1958, p. 299), Copyright (1958), with permission from Elsevier

Piccinini and Craver are right about cognitive models being elliptical mechanism sketches, we should expect later developments of Broadbent's model to fill in details about what kind of channels, filters and gates occupy the boxes in Broadbent's model, in roughly the same way as happened with the sodium channel model. Below I follow this model through its later developments, showing that its development does not proceed like that of a mechanism sketch.

Broadbent's (1958) model of attention was developed to explain results from dichotic listening experiments (where different messages are presented to the two ears simultaneously). The most basic result, known as the cocktail party effect, is that people can pay attention to a single stream of what they're hearing, and effectively block out distracting sounds. The model posits that there is a short term store where information is kept briefly, before it moves through a selective filter, which lets only some streams of information through, blocking the others. This filter was thought to be necessary, because information subsequently must move through a limited capacity channel (we can't pay attention to everything), and what makes it through the channel seems to be based on criteria fed in from higher areas; important information generally isn't lost in the shuffle.

Broadbent's model is illustrated in Fig. 4, which he labeled an "information flow chart." His approach to psychology took inspiration from radio and telephone technology, which send messages from place to place. For Broadbent, information processing refers to the journey data takes from its arrival as sensory input, its processing, to its being transmitted as output. The diagram consists of boxes representing functionally-described devices that input, manipulate or store data, and arrows representing the route of data flow.

This type of information-processing model is typically represented using a flowchart. The diagram illustrating the definition of a mechanism sketch in MDC also looks like a flowchart, so there are at least superficial reasons for Piccinini and Craver to draw comparisons between cognitive models and mechanism sketches.[6]

[6] There are of course approaches to cognitive psychology that are not connected to information processing, and psychologists who eschew the use of flowcharts. Nevertheless, this has, at least until recently, been a dominant approach.

Despite the similar appearance, these diagrams represent, and fail to represent, quite different sorts of things. There are several distinct types of flowchart, as I describe in Stinson (2013). Broadbent's is a data-flow diagram, and MDC's is a state-transition diagram. Data-flow diagrams show data sources, data sinks, and the communication pathways between data-processing modules. Diagrams of mechanism sketches (which often take the form of state-transition diagrams) show the entities constituting the mechanism and the activities they perform. Each kind of diagram highlights a different kind of detail, and neither is in principle more complete than the other.

There is nothing in Broadbent's diagram that looks particularly like a filter or a gate, although the limited capacity filter slightly resembles a tube. (This difference in height of boxes, which for Broadbent may have had some significance, disappears in reproductions of this diagram in later texts.) Broadbent steered clear of making any speculations about what the physiology of the attention filter might be, which is not surprising, since there were at the time few experimental methods available for investigating human brain functioning. Although he didn't specify physiological mechanisms, he meant his model to be compatible with their discovery. He writes, "we have tried to make our hypothetical constructs of such a kind that they could be recognized if it were possible to observe them directly: a filter or a short-term store might take different physiological forms" (Broadbent 1958). One of the advantages he cites of describing his model in "cybernetic language" is that "a description…in terms of information flow will be readily attached to physiological knowledge when the latter becomes available" (Broadbent 1958). Based on these comments, it seems initially plausible that the attention filter model could be an elliptical mechanism sketch.

But Broadbent complicates matters. He criticizes Hebb for prematurely expressing his theory in physiological terms, which leaves him open to having his theory disproved based on irrelevant physiological findings (Broadbent 1958). Broadbent then claims that the relationship between physiologists and psychologists is analogous to that between auto mechanic and driver, and quips that "for many purposes a knowledge of the mechanism is not essential to the driver" (Broadbent 1958).[7] This suggests that while knowing the physiological mechanisms is desirable, there is something worthwhile about coming up with a good psychological model independent of what the physiology later is discovered to be.

If we trace the development of Broadbent's filter theory of attention forward, we see that as further knowledge accrues, there is for many years no movement toward figuring out the physiological implementation of the boxes in cognitive models of attention. If physiological details had been available, many psychologists would have been interested, but even without such details, they still managed to do some science. Cognitive psychologists did not treat the model as a mere elliptical mechanism sketch waiting to be filled in. They considered it to be an explanatory model in its own right even without the physiological details.

Subsequent models of attention do get more complex, but the details remain of the same "cybernetic" type as in Broadbent's model. Figure 5 shows a comparison of Broadbent and Treisman's (1960) models.

[7] Of course Broadbent does not use the term 'mechanism' in the technical sense of the neo-mechanists.

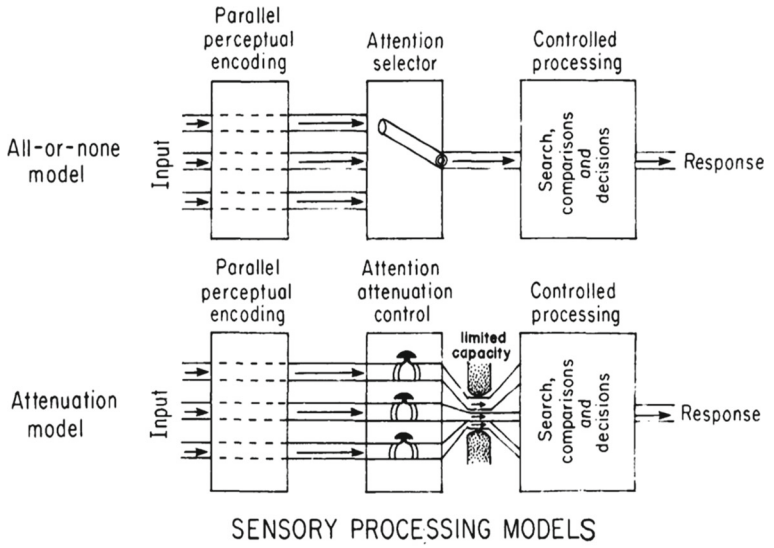

Fig. 5 Comparison of Broadbent and Treisman's filter models. Reprinted from Schneider and Shiffrin (1977). Copyright ©1977 American Psychological Association

Treisman's main alteration was to change the functioning of the filter, so that instead of just a single stream being selected at a time, many streams might be attended to at once, but with most of them attenuated. This change was in response to results suggesting that content on the unattended channels does have some effect on processing, for instance by priming certain responses, as well as the simple observation that highly salient cues, like one's name, are reliably perceived even on an unattended channel. The data flow path changed in this updated version, but it was not made any closer to being a how-possibly or how-plausibly model.

Unlike in Hille's diagram of possible sodium channel gating mechanisms, the swinging tubes and vices shown in Fig. 5 are meant purely metaphorically. Treisman (1960) discusses channels and filters at length, but the only physical entities she mentions in the paper are the ears. Broadbent's goal of eventually filling the model in with physiological details was set aside for about 30 years until the data started to become available, but psychologists still found useful work to do during this period.

The next major alteration is pictured in Fig. 6, which shows Shiffrin and Schneider's (1977) filter model. They proposed a theory that made a functional dissociation between controlled attention and automaticity, supported by experiments where they examined when one task interferes with another. Tasks that aren't adversely affected when combined with other tasks are considered automatic, while controlled processing does suffer interference when combined with other tasks. In their diagram, there are multiple levels of automatic processing within short term storage, instead of Broadbent and Treisman's short-term store followed by a single filter. These multiple filters can also have feedback effects on one another. The main change to the model was

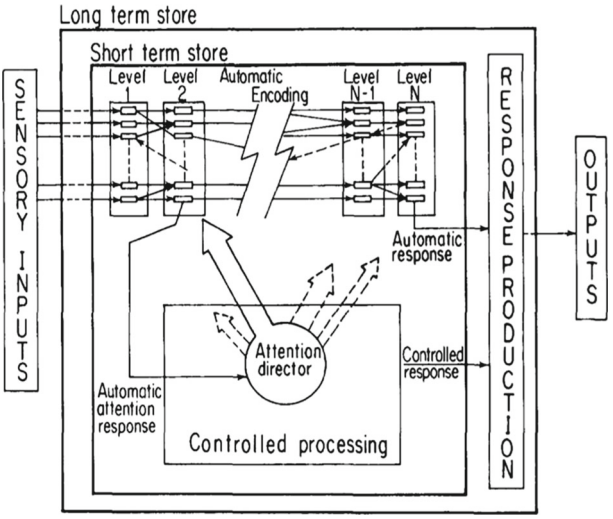

Fig. 6 Shiffrin and Schneider's filter model. Reprinted from Shiffrin and Schneider (1977). Copyright ©1977 American Psychological Association

the addition of a second pathway for controlled processing, which can exert effects at various stages during automatic filtering.

Once again, this model of attention added various complications in response to experimental data, but these complications are not details about how any of the entities might possibly or plausibly be realized in a physical mechanism. Instead the model remains, as in the older versions, essentially a data-flow diagram consisting of boxes connected by arrows showing data flow.

If we pause here before looking at more contemporary models of attention, we can already see problems with seamless accounts of integration. So far there are no signs of the cognitive model being used as a mechanism sketch. There is no equivalent to Hille's list of possible filter mechanisms in this example, no research inquiring into whether the attention filter works by pinching, twisting, sliding, or plugging, no discovery that it actually works like a ball and chain. Compared to sodium channel gating, where it is clear that a mechanism sketch was gradually filled in with details, this example so far does not look at all like a story about discovering lower-level mechanisms.

Another thing to note is that although the model is not getting any closer to revealing lower-level mechanisms, there is significant model development happening. The model responds to new experimental evidence from the original dichotic listening paradigm, as well as evidence from related phenomena like task interference. As it does so, it gains the status of a robust explanatory model within cognitive psychology. Models in other areas of psychology assume that attention is structured according this model as a starting point for their research. As a rough measure of that influence, Shiffrin and Schneider (1977) and Schneider and Shiffrin (1977) combined have been cited over 10,000 times in research covering just about every area of psychology and its applications. Papers and books citing this model deal with topics including organization theory, personality, language acquisition, visual attention, persuasion,

expert performance, perceptual symbols, reading, racial attitudes, flight control systems, consumer research, advertising effectiveness, instructional design, visual search, emotional disorders, emotional contagion, stereotypes, developmental psychology, social psychology, skill acquisition, and more. There is a sense in which the model gradually became more integrated, but this integration was lateral, connecting with other cognitive models, rather than vertical, connecting with physiological models.

Most importantly, the model isn't treated like a placeholder for a to-be-determined neural model; it is treated like an explanatory model in cognitive psychology. It is only a partial explanation, and it could at any time be rejected or revised based on new experimental findings, but all scientific explanations are partial and subject to revision in this way. A mechanism sketch, in contrast, is not yet an explanation at all. Thousands of research papers in other areas would not take it for granted that the model is more-or-less correct if it were a mere sketch.

Despite the fact that in cognitive psychology, the attention filter model is considered to explain aspects of attention, one might object that attention models remained in the elliptical sketch stage for several decades, but still ought to be filled in with neural details. Let's move ahead to the next major development in attention filter models.

3.2.1 Mapping cognition

The sense in which Piccinini and Craver are correct in their description of integration is that a popular strategy in cognitive neuroscience has been to press this sort of decades-old cognitive model into service as a starting point for investigating neural mechanisms. A common method is to generate mappings from cognitive models onto brain areas using neuroimaging technologies like fMRI, and an adapted version of Sternberg's (1969) subtraction method.[8] Descendants of Broadbent's model have been subject to this kind of investigation.

Schneider did such an fMRI study with a later iteration of his attention model called CAP2, which is described in Schneider and Chein (2003). Figure 7 shows the updated cognitive model on the top, and on the bottom, the same model overlaid on a drawing of the brain. They describe a "mapping" between their flowchart boxes or functional modules, and brain regions. The mapping is supported with neuroimaging data, meaning that these regions show more BOLD activation during tasks associated with those functional modules than during control tasks.

This may be what Piccinini and Craver (2011) have in mind as what should be done with cognitive models in order to integrate them with neuroscience: to treat them as templates for investigating the underlying neural mechanisms. In Bechtel's terms, the strategy is to try to localize cognitive components in the brain. It is clear that many projects in cognitive neuroscience use this strategy of mapping a cognitive model onto fMRI results. This localization step alone can't turn a cognitive model into an integrated, multi-level explanation, but it may be a first step towards finding connections between cognitive tasks and their implementation in the brain.

[8] fMRI methodology has recently moved on to more sophisticated methods than subtraction.

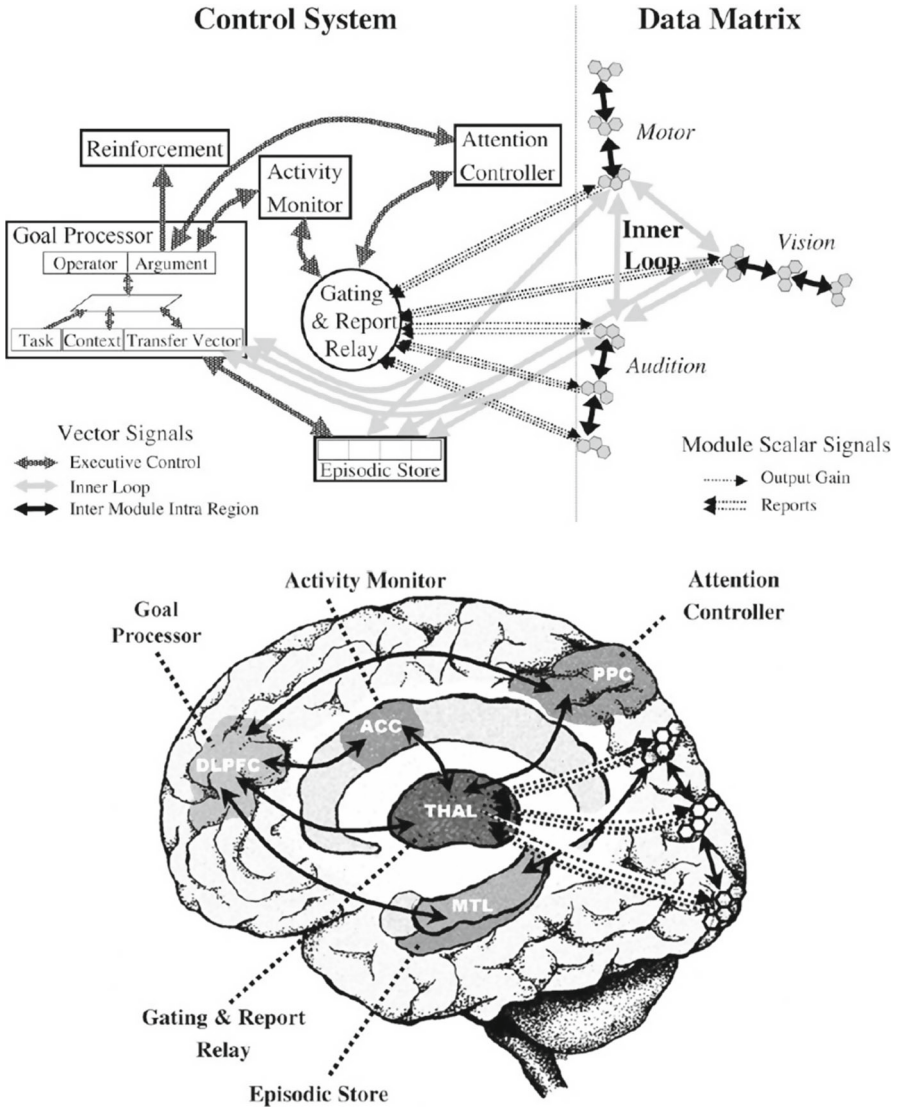

Fig. 7 Schneider and Chein's attention model. Reprinted from Schneider and Chein (2003). Copyright ©2003 Cognitive Science Society

This strategy is not without its problems. One problem is that it is difficult to tell from neuroimaging data whether a successful mapping has been achieved, because a neuroimage with hotspots will be generated under most circumstances, even in cases where localization should fail. Bechtel and Richardson (1993) outline several scenarios where localization should be expected to fail, such as in non-decompositional systems. Any two tasks differ in their brain activation patterns in some ways, which will generate hotspots in fMRI images as long as the threshold is low enough. Vul and

Kanwisher (2010) comment that these thresholds are often set too low: "Insufficient correction for multiple comparisons guarantees that some number of voxels will pass the statistical threshold, and offers no guarantee that they did so because of some true underlying effect rather than fortuitous noise" (Vul and Kanwisher 2010, p. 74). That hotspots are generated thus does not show that there are neural structures in those areas corresponding to the functional parts in the cognitive model.

Setting aside the technical difficulties, which more advanced fMRI methods may resolve, let's suppose the hotspots from such an fMRI experiment genuinely reveal the brain areas that are comparatively more active during performance of a given cognitive task. This alone doesn't get us very far, but may, if we're lucky, serve as a useful guide towards investigating the underlying neurophysiology. If we project ourselves into a hypothetical future in which the brain regions identified have subsequently been thoroughly investigated by neurophysiologists, we could be in a position to fill in the details of the mechanism sketch derived from the cognitive model.[9]

Seamless accounts of integration suggest that the neural parts so discovered should correspond to the cognitive model's functional parts. If there's a filter in the cognitive model, further investigation should reveal a filter in the brain. Another problem can arise here. Sometimes physiological mechanisms are surprising. Lateral inhibition, attractor networks, negative feedback loops, and oscillations were not the sorts of mechanisms we expected to find responsible for Mach bands, object recognition, metabolism, and sleep cycles. Kaplan (2011) and Kaplan and Craver (2011) claim that cognitive models can't be mechanistic explanations if they don't map onto the lower-level mechanisms. In cases where the lower-level mechanism turns out to be surprising, the mechanism sketch would turn out to have been wrong. In this case, should we feel compelled to alter or abandon the cognitive model that was used to generate the sketch? Piccinini and Craver suggest that "properly constrained" cognitive models should respond in this way to neural findings that challenge them. If cognitive models were just sketches of neural mechanisms, there should be no reason to hang onto them if the sketch turns out to be wrong. If we draw a distinction between a cognitive model *being* a mechanism sketch and *being used as* one, it is not so obvious that the failure of the sketch should be damning for the cognitive model too. We've already seen that cognitive models are more than just mechanism sketches; they are explanatory models with an independent set of data backing them up. As cognitive models they are still worth something. This sort of surprise or lack of correspondence at a lower level does, of course, sometimes prompt us to reject our higher level models, but, despite what Kaplan claims, there is no imperative to do so. In the next section, we'll see that psychologists sometimes hang onto their models even when they turn out not to correspond to the underlying neural mechanisms.

It is correct to say that a popular strategy in cognitive neuroscience is to *use* cognitive models as templates to guide discovery of neural mechanisms. That good cognitive models should be informative about the structure of underlying neural mechanisms

[9] At this point it is unclear whether recent research on the neural mechanisms of attention vindicate Schneider and Shiffrin's model or not. The latest publication on attention out of Schneider's lab, using white-matter imaging and fMRI, claims to "provide critical evidence for the biased competition theory of attention" (Greenberg et al. 2012, p. 2781).

is assumed as a research strategy, and I have no doubt that it is often a fruitful one. However, cognitive models are more than elliptical mechanism sketches. If they were nothing more than this, it would be hard to make sense of the decades psychologists spent doing experiments to refine the attention filter model. It would be hard to make sense of how researchers in various areas of psychology and its applications depend on the model to provide an explanation of attention, regardless of whether the neural mechanisms are known. It would also be hard to make sense of how psychologists sometimes resist making major alterations to their models based on neurophysiological findings when these conflict with independent evidence from psychology, an example of which we'll see shortly. Using cognitive models as templates in this way may often be a good research strategy, and when successful may yield well-integrated models, but this is neither a very reliable nor the only route to integration.

One might object that this single example of how a cognitive model develops over time cannot support general claims about the relationship between cognitive models and neural mechanisms. The trouble is that Piccinini and Craver did not provide any examples where their account of integration does work. The burden of proof lies with them. Without any success cases where these mappings have led to an integrated model for even one phenomenon, the promise of this method remains an article of faith. In the next section I take up an example of Bechtel's, and argue that on Piccinini and Craver's account, it is not a successful integration. First I'll say a bit more about the metaphysical assumptions underlying Piccinini, Craver, and Kaplan's claims.

4 Metaphysical seams

The metaphysical seamlessness assumption entails there being a unified hierarchy of mechanisms. This suggests that lower-level mechanisms should not cross-cut the boundaries of higher-level mechanisms. Because neural working parts are both more fundamental and seemingly more directly observable than cognitive working parts, when it turns out that they don't bear a neat mereological relationship to one another, the blame tends to be put on the cognitive model. A corollary of the metaphysical seamlessness assumption makes conforming to the divisions between neural mechanisms a criterion for a cognitive model being correct.

Kaplan, for example, insists that cognitive models are no good as explanations unless they map onto more detailed mechanisms, as described in his "model-mechanism-mapping (3M) constraint" (Kaplan 2011; Kaplan and Craver 2011). He starts from the reasonable requirement that for a model to be a mechanistic explanation it must identify causally-relevant factors.[10] From this Kaplan concludes that to have explanatory power, a model must explain in terms of component parts. Weiskopf (2011) points out that cognitive models sometimes also explain in terms of wholes, but argues that cognitive models do have explanatory power by denying that they are *mechanistic* explanations. It's not necessary to deny that cognitive models explain

[10] This leaves out some things we might want to call explanations, but when restricted to causal-mechanistic explanation seems uncontroversial.

mechanistically in order to avoid this problem, however, because Kaplan's conclusion does not follow. Mechanistic causes needn't be parts.

One difference between the MDC account of mechanism and others is that where others talk about parts, MDC talk about entities. The import of this terminological disagreement is that there might be causally-relevant entities in cognitive models that are not constituent parts. Entities or activities might be multiply realized, realized by diffuse, distributed networks, their realizers might overlap with one another (brain regions typically perform many functions and engage in many activities, which needn't be functionally related), or they might depend on higher-level properties of brain mechanisms. There is little reason to believe that cognitive and neural entities and activities must be similarly organized. In complex systems, what looks stable and robust at one scale may not be so at another scale.

No compelling reason has been offered as to why hierarchies of mechanisms must be unified. It is hard to see how Craver's idea of levels being local to specific mechanisms is compatible with a unified hierarchy of mechanisms. Machamer and Sullivan (2001) and Sullivan (2009), taken together, suggest that levels are local not just to specific mechanisms, but even more specifically to mechanisms investigated with a given paradigm. Nature's joints when seen at the cognitive level might not happen to fall at the same locations as neural-level joints.

The difficult question is what to do when a well accepted cognitive model does not map neatly onto brain mechanisms. A standard norm of good scientific models is that they be consistent with models in cognate fields. This means that if there are several cognitive models that, based on all other criteria, are approximately equally good, but one maps much more neatly onto the brain than the others, the one with the neat mapping should be preferred. This is a perfectly legitimate way of using neuroscience to constrain cognitive models. Alternatively, if a non-essential aspect of a cognitive model is contradicted by neuroscientific discoveries, and the cognitive model can be altered to better fit the neuroscience, again this would be a perfectly legitimate way of using neuroscience to constrain cognitive models. In clear-cut cases like these, psychologists ought to revise their models or model allegiances to better match the neuroscience. One caveat is that there may be some specific niches in which a generally worse model does a better job. Many sciences keep multiple imperfect models in their toolboxes for these uses.

A slightly more complicated sort of case would be where there is a cognitive model that is clearly preferred based on the norms of explanation internal to psychology, but its entities cross-cut what look like the natural divisions in the underlying neural structures. This is particularly likely to occur in cases where knowledge about the relevant neural structures is not initially available to psychologists when building their models, as in Broadbent's case. If later developing knowledge about the neural underpinnings of the phenomenon suggests a different decomposition into working parts than the well-accepted cognitive model, it is not immediately clear what the resolution should be. If we think of cognitive models as mechanism sketches, it seems like the solution should be to alter the sketch, thus alter the cognitive model. This is not what psychologists in fact do, however, as I'll describe below. It would be uncharitable not to at least consider the possibility that cognitive psychologists might have a good reason for holding on to their models under these circumstances.

4.1 A putative case of integration

Bechtel (2008) offers a case study of how work on memory mechanisms developed from early phenomenal models to integrated multi-level mechanisms. It is meant to illustrate the decomposition and localization heuristic at work in a case where a top-down functional decomposition preceded bottom-up work on the structural details. As in Broadbent's case, the physiological data was not available until much later, so psychological models were developed first.

The main lesson Bechtel draws from this case, in which progress was a little rocky, is that working at either too high or too low a level can hinder the search for mechanisms. That seems right. But what this case study also demonstrates, is that a perfectly acceptable cognitive model can prove to be a very poor sketch of the underlying neural mechanisms. In this case, the cognitive model arguably hindered progress towards discovering neural mechanisms. Furthermore, although the cognitive model turns out to fit quite badly with current neural models, psychologists have not given up their models, contrary to the normative claims made by Piccinini and Craver (2011), Kaplan (2011) and Kaplan and Craver (2011).[11]

Psychologists' most basic divisions in terms of memory processes are between encoding, storage, and retrieval. Storage is typically divided into three types in terms of how long they last: long-term, short-term, and a very brief type usually called working memory. Types of memory can also be divided into separate stores for different sensory modalities, such as iconic or olfactory memory. Long term memory is often divided into declarative and procedural memory, both of which also subdivide into further types. A combination of sources of data contribute to this basic taxonomy, including introspection, behavioral experiments, and studies of amnesic patients.

Bechtel relates how these basic divisions guided attempts to localize memory processes in the human brain, using PET and fMRI, animal physiology, and lesion data from amnesic patients. One key source of data that drove localization attempts was the case of HM. HM's combination of retrograde and anterograde amnesia after hippocampus resection led researchers to look in the hippocampus for the neural mechanisms responsible for memory encoding, storage, and retrieval. The expectation was that neural mechanisms performing each of the functions in the cognitive model might be found there. The search for mechanisms performing these functions was largely unsuccessful.

Another line of research took a different starting point: single-cell recordings from rats while they performed spatial memory tasks. Encoding, storage and retrieval cells were not found in the hippocampus. Instead, place cells and grid cells that encode the location and direction of the animal in its environment were discovered. This suggested that the main function of the hippocampus may instead be spatial processing and navigation, not memory. The search for memory mechanisms in the hippocampus seemed off target given these findings. One suggestion was that hippocampal lesions may affect memory indirectly by damaging axon pathways from the basal ganglia to the basal forebrain which pass through the white matter of the medial temporal lobe.

[11] Bechtel evidently does not support those claims, since he considers this episode a successful case of integration.

Researchers studying humans continued to look for memory mechanisms in the hippocampus, however. A recent account in this tradition suggests that the recurrent connections and rapid synaptic plasticity in region CA3 combine to create representations of sequences of information (Eichenbaum 2004). A popular view now is that the hippocampus is responsible for both spatial navigation and declarative memory, although there is less agreement on how it manages to do both. Eichenbaum and Cohen (2014) argue that the two can be reconciled by understanding the role of hippocampus in terms of both spatial and nonspatial relational processing. This reconciliation requires a significant restructuring of cognitive models of memory, however.

Cognitive models of memory did not provide a helpful sketch of the neural mechanisms discovered. Trying to locate memory functions like encoding, storage and retrieval in the hippocampus hindered progress towards discovering the neural mechanisms of memory by leading researchers to look for the wrong sorts of things, and it did nothing to aid in the discovery of the role of hippocampus in spatial navigation. The cognitive model of memory led researchers to these discoveries only in the sense that failure to find what they were looking for forced them to try other tacks. Decomposition and localization is a good heuristic generally, but in this case, it fared no better than trial and error. Proceeding with the physiological work without the guidance of the cognitive model of memory arguably would have been more productive. In this case, the methodology was not seamless.

Further problems arise beyond the methodological assumptions not panning out. As Bechtel describes, there is increasing evidence "that challenges the distinctions between episodic and semantic memory, between short- and long-term memory, and between memory and other cognitive functions" (Bechtel 2008). The first piece of research reviewed "brings into question the assumption that encoding and retrieval are really distinct operations" (Bechtel 2008), despite Tulving's hypothesis that encoding and retrieval are performed, respectively, in left and right prefrontal cortex. This is followed by a summary of evidence suggesting that some semantic retrieval is done in the area supposed to be devoted to episodic retrieval, and vice versa. Next comes evidence that long- and short-term memory are not separate systems, but rather depend on shared operations in the same brain areas. Finally comes evidence that memory and language processing are not entirely distinct, and evidence suggesting that memory storage is inseparable from neural processing in general. Essentially all of the distinctions made in the cognitive model of memory fail to map onto divisions between neural working parts. If one insists that the two decompositions ought to correspond (i.e., metaphysical seamlessness), it must look like the cognitive model of memory is deeply flawed.

The cognitive model of memory fails to map neatly onto the neuroscience. Nevertheless, the categorizations into short- and long-term memory, encoding and storage, etc., have not by any means been abandoned by psychologists. They are still found in psychology textbooks (often alongside the seemingly contradictory results from neuroscience), they still form the foundation for experimental paradigms, they are still assumed in models of other cognitive functions that interact with memory, and they are still considered useful for clinical purposes. In short, it is still a good cognitive model, based on all the other criteria psychologists use for judging models, and despite being well aware of its awkward fit with the neuroscience, psychologists have not abandoned it.

One example of how the cognitive model of memory is used in a related field is Harvey et al. (2014) which makes recommendations for improving psychosocial treatment outcomes in depression, bipolar disorder and schizophrenia. Because memory for clinical diagnosis and advice is poor among medical patients generally, they suggest that treatments should be supplemented with exercises to increase memory of therapy points. Their recommendations are based on classic memory models from cognitive psychology involving encoding in short-term memory, transfer to long-term memory, rehearsal in the phonological loop, etc. The strategies for increasing memory retention that they recommend have been extensively tested by cognitive psychologists and shown to be effective. One of these strategies is categorization, or "binding information into meaningful chunks" (Harvey et al. 2014, p. 165), which should increase memory capacity, based on Baddeley's model of the episodic buffer, which can only hold four chunks of information at a time. Regardless of whether the memory model that underlies these strategies fits well with neuroscience's discoveries about the lower level mechanisms, these memory techniques have proven their value. Its continued usefulness in a wide range of contexts justifies retaining the cognitive model of memory.

Moreover, neuroscientists haven't abandoned the cognitive model of memory either. Neuroscientists may not any longer try to find brain parts that correspond to anything like a short-term memory store, but they do look for processes that are necessary and/or sufficient for short-term memory traces of specific kinds. For example, Zars et al. (2000) is just one of many papers investigating short-term memory in *Drosophila*. They investigate an enzyme that mediates synaptic plasticity for olfactory learning. This is the sort of integration occurring in cognitive neuroscience, but it is not integration as Piccinini and Craver imagine it.

Bechtel is perfectly right that cognitive models of memory were used as the starting point for neural investigations of memory, and that decomposition and localization was used as a heuristic. Cognitive models of memory were not a very helpful guide though, and it seems fair to say that the heuristic failed in this case (as even the best heuristics sometimes do).

More troubling for seamless accounts of integration is the fact that the disconnect between cognitive and neural models, which the scientists seem to be well aware of, does not seem to bother them. Both neuroscientists and psychologists proceed with their work as though the lack of agreement between the hierarchies of mechanisms each investigates is not a problem at all. The two sets of concepts peacefully coexist despite neither acting as a sketch of the other, and research into the complex relationship between the two distinct hierarchies of mechanisms continues.

This is not a case of integration in Piccinini and Craver's sense, since the models of memory psychologists talk about do not work as a sketch of the neural mechanisms involved. This case study is the only example of integration starting from a cognitive model that we've been presented. It is neither a successful case of decomposition and localization nor of treating cognitive models as elliptical mechanism sketches. Something more is needed to make integration work. What is needed is an account that allows for the very much to be expected outcome where cognitive and neural mechanisms don't match up neatly.

4.2 What to do with cognitive models

According to Piccinini and Craver (2011), the sort of schism between psychology and neuroscience that we see in work on memory shouldn't happen. They even address this example explicitly. Discussing the task analysis of memory into encoding, storage and retrieval, they claim that, "If the study of brain mechanisms forces us to lump, split, eliminate or otherwise rethink any of these sub-capacities, the functional analysis of memory will have to change" (Piccinini and Craver 2011). The functional analysis of memory has not changed though, and Piccinini and Craver's argument as to why it should, amounts to the appeal that explanations be "properly constrained." They offer no arguments for why this constraint is proper. The situation is reminiscent (but even more challenging to deniers of psychology's autonomy, I think) of what Aizawa and Gillett (2011) describe in vision science, where discoveries about the neural underpinnings of color vision do not result in vision scientists changing their higher level categories about types of color vision, despite the wishes of some philosophers of mind.

The sort of mismatch we've seen between cognitive models that by psychology's standards are perfectly good, and what neuroscientists find out about the neural working parts is to be expected though. Bechtel and Richardson (1993) argue that the decomposition and localization heuristic should only work reliably for nearly decomposable systems, which the brain certainly is not. Some visual phenomena have been successfully localized in visual cortex or retinal structures, but in the rest of cognitive psychology, partial localizations are probably the best we can hope for.

When a cognitive model does not match neatly with neural working parts, it certainly warrants a close look at whether the cognitive model needs revision or rejection, but its rejection should not be a foregone conclusion. This sort of situation is not an adequate reason for overthrowing the standards by which cognitive models are judged. Fit with cognate fields is among the criteria used for judging scientific models, but it is not a trump card. Arguments as to why neuroscientific findings should trump other considerations either assume a particular kind of integration as a normative ideal (Piccinini and Craver 2011), assume that low-level mechanistic explanations are the only true explanations (Kaplan and Craver 2011; Kaplan 2011), or assume a metaphysical picture where mechanisms at higher and lower levels must fit together into a neat part-whole hierarchy (Craver 2007).

There is another option. Instead of throwing out cognitive models whose components do not map neatly onto neural working parts, we could heed cognitive psychologists' repeated pleas that their field does have a legitimate subject matter, and that their models do track robust regularities in the world.

5 Integration with seams

Instead of constructing philosophical justifications for neuro-chauvinism, we can accept a more limited, seam-filled kind of integration. It is still possible to find neural explanations of cognitive phenomena without the cognitive and neural decompositions neatly matching up. These explanations may be partial ones, we may need many dif-

ferent explanations for different explanatory contexts, and there may be a lot of loose threads left hanging. Just as what we once hoped would be full reductions in physics turn out to be "fragmentary patchy explanations" that fail at the margins, as Schaffner (2006) describes, so too are cognitive–neural integrations likely to be patchy. In the biological sciences, these patches tend to be rather more numerous, and the margins rather wider than in physics, but the problem is the same.

Just as we don't abandon electromagnetic theory because it fails to entirely account for optical phenomena, we needn't abandon cognitive psychology's model of memory because it fails to entirely account for that phenomenon. For some explanations in physics, we need to invoke quantum mechanics or special relativity in addition to electromagnetic theory. For some explanations of memory we need to invoke particular neural circuits or cellular dynamics in addition to the cognitive model. Although we sometimes need lower-level theories to get the details right, working exclusively with a low-level theory like quantum mechanics would be unwieldy; electromagnetic theory still plays an important role, because it describes robust regularities. Those regularities don't show up at the lower level. Likewise, although we sometimes need to invoke neural mechanisms, working exclusively with them would be unwieldy and would miss some phenomena. Cognitive models still play an important role, because they too describe robust regularities. Integrative explanations should not replace the one kind of explanation with the other, they should combine them.

The sort of explanatory texture we're left with is an uneven one, where for any given phenomenon, we might need to combine models or theories at multiple scales, and to explain different phenomena, we might need to patch together a different collection of models. The mosaic picture Craver (2007) describes supposes that mechanisms will neatly fit together like the pieces of a puzzle. My suggestion is that the picture is more like a garment so patched and mended that it's no longer clear what is original and what is patch.[12] The whole thing may indeed not hang together perfectly. It may have some holes that can't be mended.

Obviously something more than just a new metaphor would be helpful. Below I work towards a positive account by specifying a few of the characteristics that a more realistic solution to the problem of cognitive–neural integration should have, differentiating my account from that of others, and giving a few examples.

5.1 Cognitive mechanisms

The first characteristic an account of integration should have is that the richness of cognitive phenomena should not be whitewashed in the name of decomposability. At least one popular account of mechanistic explanation is compatible with this desideratum. Glennan defines mechanisms in terms of the functional relations between parts, and notes that when functional decompositions do not match up with spatially localizable parts, it is still the functional structure that is constitutive of the mechanism (Glennan

[12] I have in mind something like the Barbour coat the Prince of Wales's wore while mending hedges on TV, which was more patch than original, and led to a great hullabaloo in the British press over his habit of repairing old clothes belonging to long-dead kings, and having shoes made of leather dredged up from shipwrecks (Wallop 2013).

2005, p. 447). This means that multiple models can all be 'correct.' Glennan endorses Giere's view of models as being like maps, in that they represent only some aspects of the system modeled.

I agree with Glennan that models typically only account for some aspects of the system they model, and thus there may be many models, even apparently contradictory ones, all of which are 'correct.'[13] Where I disagree with Glennan is on the role of representation. The map metaphor is a nice metaphor, and models are indeed partial, but I do not think they are partial representations the way maps are. Instead they are partial because they instantiate just some of the types to which the system belongs, or put another way, they identify just some of the causes at play. Just about any system will have causes working at various levels, which means that there will be correct (but incomplete) models at various levels.

Treating models as map-like representations fails to constrain models sufficiently. A second characteristic an account of integration should have is that integrative models serve as explanations. A good explanatory model should pick out at least some of the causally-relevant entities in the system it describes, and instantiate analogues of them. Componential accounts of mechanistic explanation call these "working parts," but causes needn't strictly be parts. A good explanatory model should also instantiate at least some of the causal relations among its entities, or what MDC call activities. Calling a representation map-like makes clear that not all grains of detail are included, but fails to emphasize the important criterion that causal relations, beyond spatial ones, are instantiated.

Weiskopf (2011) argues in his critique of Piccinini and Craver (2011) that functional analyses can be non-componential (for instance if the whole rather than a part performs a task), while still fulfilling the norms of good explanation. I think he's right that an important category of psychological explanations are non-componential, and I fully agree that functional analyses can be good explanations regardless of whether they are fine-grained. Weiskopf also points out that functional and structural decompositions can cross-cut one another, which is a point that much of the current literature on mechanisms glosses over. He has a stricter view of mechanistic models than Glennan. He says a mechanistic 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" (Weiskopf 2011). Weiskopf argues on these grounds that cognitive models are not in general mechanistic.

My motivation for allowing cognitive mechanisms is not just for the sake of proliferating mechanism talk, but because one of the main uses for the term in science is to describe more generic ways in which functions are achieved. Information-processing mechanisms needn't refer to the specific entities that pass the information around, but can still explain things like bottlenecks and bandwidth. Statistical mechanisms can explain why ions pass through porous membranes with a particular gradient, or why populations become homozygous for beneficial traits, without having to refer to the particulars. That the specifics aren't essential to understanding how the mechanisms work does not mean that these aren't real-world mechanisms. There are many real-

[13] See Mitchell (2000) for a detailed elaboration of this sort of pluralist view.

world information processors, chemical gradients, and populations that instantiate these generic mechanisms. Cognitive mechanisms are, in the same way, mechanisms that operate the way they do because of the coarser grained entities involved, and the activities of those. There are, of course, interesting and important details at a finer grain, but this does not erase the fact that coarser grained mechanisms are also at work. For a complete picture, one needs to know about all the mechanisms at work, and how all of these interact.

The main thesis of Bechtel (2008) and Piccinini and Craver (2011), that integration can be achieved via mechanistic explanation, I think is correct. This is the third characteristic that we should expect an account of integration to have: that it be a mechanistic account (simply because there are currently no other viable options on the table). Weiskopf's contention that cognitive models are not in general mechanistic suggests that such an integration is not possible, but I disagree. For there to be non-componential mechanistic explanations, what we need is for causal powers to sometimes be held by constituents of mechanisms that aren't parts. Intuitively this shouldn't be a problem: running causes a person's heart-rate to increase, roundness causes balls to roll, massiveness causes planets to exert gravitational force, growing up in Ottawa causes one to pronounce "about" with a vowel sound amusing to Americans, and so on. All of these non-components are difference-makers, so on an account of causation like Woodward's (2005), they are causes.[14]

A non-componential view of mechanisms would not rule out cognitive mechanisms. Relaxing the requirement that mechanisms form a unified hierarchy allows for partial integrations between cognitive and neural mechanisms. Instead of one unified hierarchy, we could have many local, partial hierarchies, with some entities participating in more than one. These partial hierarchies are how different models connect. Unified explanations are a tempting goal, but probably not a realistic one, at least in complex fields like cognitive neuroscience. What we currently have are collections of models, each of which explains some small part of the phenomenon of interest: whether it be some subset of the system's behaviour, a limited range of parameter values, or how it acts in a particular context. There is no reason to believe that it should always be possible or even desirable to merge these collections of models into a unified hierarchy of mechanisms.

The remaining problem is to figure out how to combine different causally-relevant factors into integrated explanations. This is not a problem specific to non-componential, non-unified mechanistic explanations. Even with many causes operating on the same level, figuring out how they all combine is usually not as simple as just summing their effects, as one might do with the forces of colliding billiard balls. Where there are many causes at work, each with its own domain of applicability and set of interactions with other causes, empirical knowledge or computer simulation may be necessary to work out what the outcome will be in a given case. There may not be a neat story about how the causes interact; it may be context-dependent. One of the strengths of mechanistic explanation compared to covering law models of explanation is its honesty about where the real work of science is to be done. Just as theories do

[14] Even if one objects to top-down causation, there remain several examples in this list that are not top-down.

not give all the answers, knowing which mechanisms are at work in a given situation does not tell you everything, because mechanisms don't always work in exactly the same way in different circumstances.

In the case of memory research, we now have a collection of models each covering a different aspect of what the hippocampus and related areas do. There are models of associative memory, models of pattern completion and separation, models of cognitive mapping, models of path integration in grid cells, and models of synaptic plasticity in the fly's olfactory system, among others. Each of these provides a piece of the story about how memory works, highlighting one or more of its important features, but none gives a complete picture.

Even two models of the same phenomenon, such as Fuhs and Touretzky (2006) and Burgess et al. (2007) are not strictly competitors. Fuhs and Touretzky (2006) suggest that the hexagonal patterns found in grid cell activations result from closely packed cells fully connected to their neighbours, as in a spin glass network. Burgess et al. (2007) provide an alternative mechanism based on interference patterns between dendritic oscillations. Both models point out structures that exist in and around these cells and both highlight causal processes that operate over those structures which may be relevant for explaining the way grid cells work. The dendritic oscillation model is preferred, but there may be some truth to both. Burgess et al. (2007) state that the mechanism Fuhs and Touretzky (2006) suggested might be added to their model "to maintain the relative location of grids and enhance their stability and precision" (Burgess et al. 2007, p. 810). Multiple models of memory each describe part of the picture, but they don't combine the way hierarchies of mechanisms are supposed to. Although they may overlap significantly, they have different boundaries, and include different sets of entities.

This is the sort of partial overlap between models that should be expected in integrative explanations. Integration does not have to mean travelling up and down a unique hierarchy identifying mereological relationships between mechanisms. Instead, integration can mean figuring out how several partial models of the same or related phenomenon connect. This might mean figuring out the range of conditions under which one mechanism is dominant, under what circumstances another mechanism's effects interfere with the dominant effect, the ways in which one mechanism helps another one do its job, by making its effects more robust or stable, etc. Research on grid cells is by no means an isolated example of this.

Overall, my conclusions on the main points are similar to Weiskopf's: Piccinini and Craver are too quick in assimilating psychological explanation to mechanistic explanation, because some cognitive models can not be usefully decomposed further into neural parts, and functional and structural decompositions can cross-cut one another such that their constituents do not neatly map onto one another. Where our views depart is over whether lack of decomposability and/or cross-cutting decompositions means that cognitive models can't be mechanistic. Insisting that mechanistic explanations must form a unified hierarchy is an unrealistic requirement. It is an extra assumption beyond what makes a model a mechanistic model, and I suspect that cognitive mechanisms are not the only ones it would rule out. If what we expect from integration are the partial connections that are found in real examples from cognitive neuroscience,

there is no reason to deny that cognitive models could be integrated with neural models into multi-level mechanistic explanations.

6 Conclusion

Much of cognitive psychology does not seem to be concerned with developing models that lend themselves easily to being used as sketches of neural mechanisms. Instead of increasingly detailed accounts of which parts of the brain correspond to the entities appearing in cognitive models, cognitive psychology is primarily concerned with providing good explanations of cognitive-level phenomena. Piccinini and Craver (2011) insist that psychologists ought to pay attention to neural constraints when building their models, because they want these models to be seamlessly integrated with neural mechanisms.

I think it's perfectly true that what some people in cognitive neuroscience are doing, for example with fMRI subtraction studies, is using cognitive models as sketches of neural mechanisms. As a descriptive account of a popular heuristic strategy, this is correct. I have argued that it is not an appropriate norm for constraining what counts as a good cognitive model.

Unlike some other defenders of psychological explanations, my point is not to support autonomy, nor to argue that psychological and neuroscientific explanations are necessarily of distinct types. Rather my point is to mark the ways they may fail to fit neatly together so as to understand the work that needs to be done if the goal of integration is to be achieved. Glossing over the differences and treating psychological explanations as mere sketches of neural mechanisms seems like an approach likely to alienate psychologists and unlikely to provide much helpful direction.

The tricky work to be done is figuring out how to go about doing psychological work while keeping in mind constraints from neuroscience, without this interfering with the main goal of forming good cognitive models; how to go about doing good neuroscience while keeping in mind richer descriptive definitions of behavior; and most importantly, what to do when the models we arrive at using these two sets of constraints conflict.

I've argued that the integrations we can realistically hope for are connections between multiple, partially overlapping mechanistic models, not a unified hierarchy of mechanisms stretching upward from neural to cognitive parts. It makes for a complicated story, not a seamless one, but one that fits better with the realities of scientific research.

Acknowledgments The first draft of this paper was written with the support of a predoctoral fellowship at the Max Planck Institute for the History of Science in Berlin in 2010–2011. Thanks are especially due to their wonderful library services. An earlier version formed part of Chapter 2 of my PhD Dissertation, "Cognitive Mechanisms and Computational Models: Explanation in Cognitive Neuroscience" at the University of Pittsburgh, 2013. Thanks to Peter Machamer, Ken Schaffner, Jim Bogen, Floh Thiels, and Boris Hennig for their helpful comments on the chapter. The final drafts were written with the support of a predoctoral fellowship at the Centre for Integrative Neuroscience, Eberhard Karls Universität Tübingen. Thanks to the participants at the workshop, *Explaining Mental Phenomena*, held in Tübingen on 24 July 2012, where I presented the paper, especially to Uljana Feest, who provided extended commentary.

References

Aizawa, K., & Gillett, C. (2011). The autonomy of psychology in the age of neuroscience. In P. M. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences*. New York: Oxford University Press.

Andersen, H. (2014a). A field guide to mechanisms: Part I. *Philosophy Compass*, *9*(4), 274–283.

Andersen, H. (2014b). A field guide to mechanisms: Part II. *Philosophy Compass*, *9*(4), 284–293.

Bechtel, W. (1986). The nature of scientific integration. In W. Bechtel (Ed.), *Integrating scientific disciplines* (pp. 3–52). Dordrecht: Martinus Nijhoff.

Bechtel, W. (1993). Integrating sciences by creating new disciplines: The case of cell biology. *Biology and Philosophy*, *8*(3), 277–299.

Bechtel, W. (2001). Cognitive neuroscience: Relating neural mechanisms and cognition. In P. K. Machamer, R. Grush, & P. McLaughlin (Eds.), *Theory and method in the neurosciences*. Pittsburgh, PA: University of Pittsburgh Press.

Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. London: Routledge.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in the History and Philosophy of Science, Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(2), 421–441.

Bechtel, W., & Richardson, R. C. (1993). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Princeton, NJ: Princeton University Press.

Bechtel, W., & Richardson, R. C. (2010). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Cambridge, MA: MIT Press.

Broadbent, D. E. (1958). *Perception and communication*. London: Pergamon Press.

Burgess, N., Barry, C., & O'Keefe, J. (2007). An oscillatory interference model of grid cell firing. *Hippocampus*, *17*, 801–812.

Craver, C. F. (2006). When mechanistic models explain. *Synthese*, *153*(3), 355–376.

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford: Oxford University Press.

Doyle, D. A. (2004). Structural changes during ion channel gating. *Trends in Neurosciences*, *27*(6), 298–302.

Doyle, D. A., Cabral, J. M., Pfuetzner, R. A., Kuo, A., Gulbis, J. M., Cohen, S. L., et al. (1998). The structure of the potassium channel: Molecular basis of K^+ conduction and selectivity. *Science*, *280*(3 April), 69–77.

Eichenbaum, H. (2004). Hippocampus: Cognitive processes and neural representations that underlie declarative memory. *Neuron*, *44*(1), 109–120.

Eichenbaum, H., & Cohen, N. J. (2014). Can we reconcile the declarative memory and spatial navigation views on hippocampal function? *Neuron*, *83*(4), 764–770.

Fuhs, M. C., & Touretzky, D. S. (2006). A spin glass model of path integration in rat medial entorhinal cortex. *The Journal of Neuroscience*, *26*(16), 4266–4276.

Gazzaniga, M. S., Ivry, R. B., & Mangun, G. R. (2009). *Cognitive neuroscience: The biology of the mind* (3rd ed.). New York: W. W. Norton and Company.

Glennan, S. (1996). Mechanisms and the nature of causation. *Erkenntnis*, *44*(1), 49–71.

Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science*, *69*(S3), 342–353.

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

Greenberg, A. S., Verstynen, T., Chiu, Y. C., Yantis, S., Schneider, W., & Behrmann, M. (2012). Visuotopic cortical connectivity underlying attention revealed with white-matter tractography. *The Journal of Neuroscience*, *32*(8), 2773–2782.

Harvey, A. G., Lee, J., Williams, J., Hollon, S. D., Walker, M. P., Thompson, M. A., et al. (2014). Improving outcome of psychosocial treatments by enhancing memory and learning. *Perspectives on Psychological Science*, *9*(2), 161–179.

Hille, B. (2001). *Ion channels of excitable membranes* (3rd ed.). Sunderland, MA: Sinauer Associates.

Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese*, *183*, 339–373.

Kaplan, D. M., & Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science*, *78*(4), 601–627.

Lipkind, G. M., & Fozzard, H. A. (2008). Voltage-gated NA channel selectivity: The role of the conserved domain III lysine residue. *The Journal of General Physiology*, *131*(6), 523–529.

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

Machamer, P. K., & Sullivan, J. A. (2001). Leveling reductionism. http://philsci-archive.pitt.edu/id/eprint/386.

Mitchell, S. D. (2000). Dimensions of scientific law. *Philosophy of Science*, *67*(2), 242–265.

Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183*(3), 283–311.

Revonsuo, A. (2001). On the nature of explanation in the neurosciences. In P. K. Machamer, R. Grush, & P. McLaughlin (Eds.), *Theory and method in the neurosciences* (pp. 45–69). Pittsburgh, PA: University of Pittsburgh Press.

Schaffner, K. F. (2006). Reduction: The Cheshire cat problem and a return to roots. *Synthese*, *151*(3), 377–402.

Schneider, W., & Chein, J. M. (2003). Controlled and automatic processing: Behavior, theory, and biological mechanisms. *Cognitive Science*, *27*, 525–559.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*(1), 1–66.

Shepherd, G. M. (1983). *Neurobiology*. New York: Oxford University Press.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, *84*(2), 127–190.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica*, *30*, 276–315.

Stinson, C. (2013) *Cognitive mechanisms and computational models: Explanation in cognitive neuroscience*. PhD thesis, University of Pittsburgh.

Sullivan, J. A. (2009). The multiplicity of experimental protocols: A challenge to reductionist and non-reductionist models of the unity of neuroscience. *Synthese*, *167*, 511–539.

Treisman, A. (1960). Contextual cues in selective listening. *The Quarterly Journal of Experimental Psychology*, *12*(4), 242–248.

Vul, E., & Kanwisher, N. (2010). Begging the question: The nonindependence error in fMRI data analysis. In S. J. Hanson & M. Bunzl (Eds.), *Foundational issues in human brain mapping* (pp. 71–92). Cambridge, MA: MIT Press.

Wallop, H. (2013, March 11). The makings of a patchwork Prince of Wales. *The Telegraph*.

Weiskopf, D. A. (2011). Models and mechanisms in psychological explanation. *Synthese*, *183*(3), 313–338.

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. New York: Oxford University Press.

Yu, F. H., & Catterall, W. A. (2003). Overview of the voltage-gated sodium channel family. *Genome Biology*, *4*(3), 207.

Zars, T., Fischer, M., Schulz, R., & Heisenberg, M. (2000). Localization of a short-term memory in *Drosophila*. *Science*, *288*(28 April), 672–675.