Model-based Cognitive Neuroscience: Multifield Mechanistic Integration in Practice

Mark Povich

Forthcoming in *Theory & Psychology*

Abstract: Autonomist accounts of cognitive science suggest that cognitive model building and theory construction (can or should) proceed independently of findings in neuroscience. Common functionalist justifications of autonomy rely on there being relatively few constraints between neural structure and cognitive function (e.g., Weiskopf, 2011). In contrast, an integrative mechanistic perspective stresses the mutual constraining of structure and function (e.g., Piccinini & Craver, 2011; Povich, 2015). In this paper, I show how model-based cognitive neuroscience (MBCN) epitomizes the integrative mechanistic perspective and concentrates the most revolutionary elements of the cognitive neuroscience revolution (Boone & Piccinini, 2016). I also show how the prominent subset account of functional realization supports the integrative mechanistic perspective I take on MBCN and use it to clarify the intralevel and interlevel components of integration.

1. Introduction

Autonomist accounts of cognitive science, as I will understand them here, suggest that cognitive model building and theory construction (can and/or should) proceed independently of findings in neuroscience. Furthermore, according to autonomists, cognitive models, when they are explanatory, are not explanatory in virtue of representing neural mechanisms. Common functionalist justifications for autonomy have traditionally appealed to multiple realizability and the (putative) fact that there are relatively few constraints between neural structure and cognitive function (e.g., Fodor, 1974; Weiskopf, 2011). In contrast, an integrative multifield mechanistic

perspective stresses the mutual constraining of neural structure and cognitive function (e.g., Boone & Piccinini, 2016; Piccinini & Craver, 2011; Povich, 2015).

The paper is organized as follows. In Section 2, I present the mechanistic account of scientific explanation and the mechanistic perspective on scientific integration. Then, in Section 3, I describe model-based cognitive neuroscience (henceforth MBCN; Turner, Forstmann, Love, Palmeri, & Van Maanen, 2017), presenting it as an interfield theory (Darden & Maull, 1977). In Section 4, I show how MBCN epitomizes multifield mechanistic integration and how a popular philosophical account of the realization relation, the subset account, supports and clarifies my multifield mechanistic perspective on MBCN.

2. Mechanistic Explanation and Multifield Mechanistic Integration

Multifield mechanistic integration is a relation that holds between scientific fields (and between the theories and models therein). An account of that relation was developed as an alternative to the traditional intertheoretic reductionist account (Nagel, 1961; Oppenheim & Putnam, 1958; Schaffner, 1993; see also Craver, 2005; Darden & Maull, 1977; and Maull, 1977). In contrast to that account, the mechanistic perspective provides a more local and descriptively adequate conception of levels (Povich & Craver, 2017), according to which levels are levels of *mechanisms*.

A mechanism, as I will use that term, is any causal system in which parts are organized such that they collectively give rise to the behavior or property of the whole in context. A mechanism is at a higher mechanistic level than its components, and those components are at higher mechanistic levels than their components, and so on, but it makes no sense to say that, in general, one kind of thing (e.g., the atom) is at a higher mechanistic level than another (e.g., an organism). Mechanistic levels are intelligible only with respect to specific (type or token)

mechanisms; that is the sense in which mechanistic levels are *local*, in contrast to the monolithic conception of ontological levels offered by the traditional reductionist account (Povich & Craver, 2017).[1]

The intertheoretic reduction model saw multifield integration as basically an explanatory relationship, explanation being understood according to the deductive-nomological (DN) model (see Salmon, 1989 for a survey of influential critiques of the DN model). The mechanistic account of multifield integration also sees it as an explanatory relationship, explanation being understood according to the mechanistic view of explanation. A mechanistic explanation describes the entities, activities, and organization of the mechanism that produces, underlies, or maintains the explanandum phenomenon (Machamer, Darden, & Craver, 2000).[2]

Mechanists make (at least) two important normative distinctions (or continua). The first is between mechanism schemata and mechanism sketches (Machamer et al., 2000; Craver, 2007). A mechanism schema is an abstract description of a *type* of mechanism, rather than a specific token or instance. Details will inevitably be omitted, but, ideally, only details that are irrelevant to the functioning of the type of mechanism. Details that are specific to tokens of the type can be added as the schema is applied to instances (Machamer et al., 2000, p. 15). Mechanism sketches, on the other hand, are incomplete descriptions of (type or token) mechanisms that contain black boxes and filler terms (Craver, 2007, p. 113). They are still partially explanatory, but they are lacking in relevant detail. More relevant details can be added to the model to fill in the gaps. Of course, no model is ever fully complete, just complete enough for practical purposes (Craver & Darden, 2013). Idealized models qualify as mechanistic explanations to the extent that they capture relevant aspects of mechanisms that produce, underlie, or maintain the explanandum phenomena (Povich, forthcoming).

The second normative distinction (or continuum) is between how-possibly and how-actually models. A how-possibly model describes a merely possible mechanism, whereas a how-actually model describes the mechanism that actually produces, maintains, or underlies the explanandum phenomenon. This distinction is evidential or epistemic: turning a how-possibly model into a how-actually model does not require modifying the model itself in any way; it requires testing the model (Weiskopf, 2011). Turning a mechanism sketch into a more complete mechanism schema, in contrast, requires modifying the model by filling in missing details (Craver & Darden, 2013). These details may be at the same mechanistic level as the rest of the details in the model, or they may be at a lower mechanistic level. The greater the evidential support for a model, the closer it is to a how-actually model. Between how-possibly and how-actually models is a range of how-plausibly models.

Mechanistic explanations, whether they be how-possibly or how-actually models, mechanism sketches or schemata, contrast with merely descriptive or phenomenal models. A merely descriptive or phenomenal model merely describes the explanandum phenomenon itself, usually in a general, concise way. Snell's law is a common example (Craver & Darden, 2013). It accurately and compactly describes the relationship between the angle of incidence and the angle of refraction when light passes between two media, but it does not explain that relationship.

Mechanistic explanations satisfy what are widely considered by mechanists and non-mechanists alike (e.g., Chirimuuta, 2014; Rice, 2015; Weiskopf, 2011) to be the normative constraints on explanation: the ability to answer counterfactual questions about the explanandum phenomenon ('what-if-things-had-been-different' questions or ‘w-questions’), and the ability to manipulate and control the explanandum phenomenon (Craver, 2007). These norms capture what is distinctive about the scientific achievement of explanation, as opposed to other achievements

like prediction, description, or categorization. These norms arguably also provide the basis for explanatory power. A model is more explanatorily powerful, according to mechanists, when and only when it can answer more w-questions about the explanandum phenomenon and afford more opportunities for controlling it (Ylikoski & Kuorikoski, 2010).

The mechanistic account thus sees multifield integration as a process of adding constraints on a multilevel mechanism, whether those constraints be at one level or different levels. Thus, multi*field* integration need not always be multi*level* integration, since different fields can investigate the same mechanism at the same level. Multifield integration is multilevel when different fields add different constraints on different levels of the same mechanism. Multilevel integration is thus aimed at giving embedded or nested multilevel mechanistic explanations of a single explanandum phenomenon, up and down its mechanistic hierarchy. This distinction will be important when I discuss the kind of mechanistic integration provided by MBCN. Though MBCN has both multilevel/interlevel and intralevel aspects, I will argue that its most distinctive achievement is intralevel.

Multifield integration occurs when findings from different fields mutually constrain our understanding of a single mechanism (Craver, 2005, 2007): constraints on its components, their causal interactions, and their spatial, temporal, and hierarchical organization. On this view, scientific integration is relativized to particular explanandum phenomena; it is thus a local process, not one that will or can occur across the whole of science, as on the intertheoretic reduction model. Researchers in different scientific fields work together to bridge mechanistic levels and bring their unique (personal and field-specific) perspectives to bear on one mechanism.

One classic example of multifield mechanistic integration in neuroscience is that of the

mechanistic explanation of the psychological phenomenon of learning and memory by the electrophysiological phenomenon of long-term potentiation (LTP; see Craver, 2005, 2007). Craver (2005, 2007) argues that this example of multifield mechanistic integration involved interlevel and intralevel components. The discovery of the phenomenon of LTP itself, according to Craver (2005), was an example of multifield, intralevel integration: the work of anatomists using Golgi staining combined with the work of electrophysiologists using microelectrodes to reveal aspects of cells at the same mechanistic level. If electrophysiologists were working with psychologists to find the neural correlates of learning and memory, then the discovery of LTP would have been an example of multifield, *inter*level integration. But that is not what electrophysiologists were doing.

When researchers started to think that LTP might be a component in a multilevel mechanism responsible for learning and memory, then multifield, *inter*level integration began taking place. This involved not only clarifying the relation of LTP to learning and memory but discovering the molecular mechanisms responsible for LTP (Craver, 2005).

3. Model-based Cognitive Neuroscience

Model-based cognitive neuroscience (MBCN) is a burgeoning intersection of cognitive neuroscience and cognitive psychology (especially in the form of mathematical psychology) (O'Doherty, Hampton, & Kim, 2007; Gläscher & O'Doherty, 2010; Forstmann, Wagenmakers, Eichele, Brown, & Serences, 2011; Forstmann & Wagenmakers, 2015; Turner, Van Maanen, & Forstmann, 2015; Love, 2015, 2016; Turner et al., 2017; Palmeri, Love, & Turner, 2017). There is now an edited book (Forstmann & Wagenmakers, 2015) and special issue of the *Journal of Mathematical Psychology* (2017, issue 76) dedicated to it.

To make the case that MBCN provides an example of multi*field* integration, I first need to

show that the component disciplines are in fact *fields*. For this purpose I will use Darden and Maull's (1977) account of fields, according to which "a field is an area of science consisting of the following elements: a central problem, a domain consisting of items taken to be facts related to that problem, general explanatory factors and goals providing expectations as to how the problem is to be solved, techniques and methods, and, sometimes, but not always, concepts, laws and theories which are related to the problem and which attempt to realize the explanatory goals" (p. 44).

It is plausible that cognitive neuroscience is such a field. Its central problem is discovering how the brain carries out cognitive tasks; it has a domain of background knowledge consisting of items taken to be facts related to that problem (e.g., facts about the structure and function of certain brain areas, the neuron doctrine, etc.); it has unique neuroimaging and electrophysiological techniques for investigating that problem; and so on.

It is also plausible that cognitive psychology is a field in Darden and Maull's sense. It has as its central problem explaining how the mind functionally carries out cognitive tasks (perception, decision-making, etc.); it has a domain of background knowledge related to that problem (e.g., the guiding ideas of the cognitive revolution, such as that the mind possesses mental representations [in some sense I will not elaborate] or processes information [in some sense I will not elaborate]); it has unique experimental methods (behavioral tasks, protocols, and "paradigms") for investigating that problem; and so on. Investigators in MBCN typically focus on *mathematical* psychology, understood by them to mean that part of cognitive psychology that constructs *formal* models of cognition, because formal cognitive models are more easily and fruitfully combined with neuroimaging and electrophysiological data.[3]

According to Darden and Maull (1977, p. 49), an interfield theory makes explicit and

explains relations between fields. For my purposes, the most relevant such relation that might be the focus of an interfield theory is the following: “A field may investigate the structure of entities or processes, the function of which is investigated in another field” (Ibid.). This is how we should think of the relation between the fields of cognitive neuroscience and cognitive psychology. My account of MBCN also fits with Darden and Maull’s account of when an interfield theory is likely to be generated: “In brief, an interfield theory is likely to be generated when background knowledge indicates that relations already exist between the fields, when the fields share an interest in explaining different aspects of the same phenomenon, and when questions arise about that phenomenon within a field which cannot be answered with the techniques and concepts of that field” (p. 49). MBCN was generated for precisely these reasons, as I explain below.

The component fields of MBCN have historically been distinct. Cognitive psychologists construct (sometimes formal) models of cognition that posit representations and computational processes over them, and they test these cognitive models using behavioral data. Cognitive neuroscientists rely on empirical measurements of neural activity and on statistical models that link that activity to cognitive processes and behavior. However, rarely are these links connected to hypothesized computations of a cognitive model. Because of this neglect of cognitive models, cognitive neuroscience often only tells us the brain areas the activity of which is implicated in or predictive of a cognitive process or behavior.[4] Without a model, cognitive neuroscience is silent on how or why a specific brain area produces a cognitive process or behavior (Turner et al., 2017; Palmeri et al., 2017). In Marrian terms (Marr 1982), cognitive psychologists have been theorizing at the computational and algorithmic levels, while cognitive neuroscientists have been theorizing at the implementational level. But without a cognitive model to guide them, cognitive

neuroscientists are blind to what computations and algorithms might actually be implemented (Ibid.; Love, 2015). Bridging Marr's levels has been called *the integration problem* by MBCN investigators (Turner, Rodriguez, Norcia, McClure, & Steyvers, 2016, p. 21; Turner et al., 2017, p. 66).[5]

It is the explicit goal of MBCN to solve the integration problem (Love, 2016; Turner et al., 2017; Palmeri et al., 2017). This problem motivated the development of new theoretical and statistical frameworks that attempt to bridge mathematical psychology and cognitive neuroscience, the frameworks that define MBCN. The development of these new theoretical and statistical frameworks fits nicely with Darden and Maull's claim that an interfield theory is likely to be generated when questions arise about a phenomenon within a field which cannot be answered with the techniques and concepts of that field. It is also clearly true that background knowledge indicated that relations already exist between cognitive psychology and cognitive neuroscience and that these fields share an interest in explaining different aspects of the same phenomena.

Turner, Forstmann, Love, Palmeri, and Van Maanen (2017, p. 66) helpfully distinguish between three categories of statistical framework for MBCN: those that (1) use neural data to constrain a cognitive model, (2) use a cognitive model to predict neural data, and (3) model both neural and behavioral data simultaneously. Let us refer to the first category as "neural constraint approaches," the second category as "neural prediction approaches," and the third category as "simultaneous approaches". There are subcategories within each kind of approach that will be explained below.

Note that in neural and cognitive model building, respectively, there are neural data (e.g., neuroimaging data or neurophysiological recordings) that are predicted by parameters of a neural

model (e.g., the slope in a general linear model), and there are behavioral data (e.g., reaction or response times) that are predicted by parameters of a cognitive model (e.g., the average rate of evidence accumulation or drift rate of a diffusion decision model of perceptual decision making). The relations between different kinds of datum and parameter help define each statistical framework. I will use the term "cognitive model" synonymously with Turner et al.'s term "behavioral model". I briefly describe each framework in turn.

First, there are neural constraint approaches, one of which Turner et al. (2017, p. 68) call the "theoretical approach". In the theoretical approach, neural considerations inspire the form of a cognitive model. The most prominent example of the theoretical approach is connectionism, where the form of models is intended to approach some level of biological realism, although in the theoretical approach, there is no mathematical link between the neural data and the cognitive model's architecture or parameters. The theoretical approach is not as prominent as other approaches in modern MBCN research, at least partly to avoid philosophical and theoretical questions about the extent to which the implementational level needs to be reflected in the algorithmic and computational levels (Ibid., 69).

The second kind of neural constraint approach is called the "two-stage behavioral approach" (Ibid.). In this approach, neural data are used to set neural model parameters and then some of the cognitive model parameters are replaced by some of the neural parameters. Wong and Wang's (2006) recurrent neural network model of perceptual decision making is an instance of the two-stage behavioral approach. Many of the parameters of that model are set by reference to neural parameters. Approaches like this are also not quite as prominent as others in modern MBCN research at least partly because they are difficult to simulate (Turner et al., 2017, p. 69).

The last neural constraint approach is the direct input approach (Turner et al., 2017). This

approach is like to the two-stage behavioral approach, but instead of replacing some of the cognitive model parameters with neural parameters (which are set by neural data), some of the cognitive model parameters are set directly by the neural data. For example, Palmeri, Schall, and Logan (2015) used neurophysiological data to set the drift rate parameter of a diffusion-type model of perceptual decision making (Turner et al., 2017).

Then there are neural prediction approaches, one of which is the latent input approach (Turner et al., 2017, p. 70). This is the approach that has been used under the heading "model-based fMRI" (O'Doherty et al., 2007; Gläscher & O'Doherty, 2010; Davis, Love, & Preston, 2012; Povich, 2015) though, as Turner et al. note (2017, p. 70), it could be used with any kind of neural data, not just fMRI. In this approach, parameters of a cognitive model are used to generate predicted neural data. These predicted neural data are regressed against observed neural data to look for correlations. This approach can be used to discriminate between competing, equally behaviorally confirmed cognitive models (Mack, Preston, & Love, 2013). For example, Davis, Love, and Preston (2012) used a latent input approach to discriminate between two competing cognitive models of object categorization, SUSTAIN and ALCOVE.[6] Each model predicts different changes in recognition strength and error correction across trials. Davis et al. found that both of SUSTAIN's predicted hemodynamic response functions (HRFs) for recognition strength and error correction were significantly correlated with observed HRFs, whereas only ALCOVE's predicted HRF for error correction was significantly correlated with observed HRFs.

The second kind of neural prediction approach is the two-stage neural approach (Turner et al., 2017, p. 70). The two-stage neural approach is like the latent input approach, but cognitive model parameters are used to predict neural data directly. In other words, instead of using cognitive model parameters to generate predicted neural data, which are then regressed against

observed neural data, cognitive model parameters are themselves regressed against observed neural data. For example, Behrens, Woolrich, Walton, Rushworth (2007) found that the learning rate in a Bayesian model was predictive of anterior cingulate cortex (ACC) activation (Turner et al., 2017, p. 71).

Finally, there are simultaneous modeling approaches, such as the increasingly popular joint modeling approach (Turner et al., 2017, p. 72; see also Turner et al., 2013; Turner, 2015; Turner et al., 2015; Love, 2016; and Palestro et al., 2018). In this approach, the parameters of neural models and the parameters of cognitive models are simultaneously related through a set of hyperparameters, allowing for mutual constraint on all model parameters (Turner et al., 2015, p. 315). Associated with a cognitive model is a probability distribution representing the probability of the behavioral data given the cognitive model parameters. Similarly, associated with a neural model is a probability density function representing the probability of the neural data given the neural model parameters (Turner, 2015, p. 201). The hyperparameters enforce a joint distribution of the parameters in the cognitive and neural models.[7] The kind of joint distribution is up to the modeler. One could, for example, assume that the parameters follow a multivariate normal distribution. This allows predictions to be made about a particular subject's behavioral data given their neural data and vice versa. Turner, Rodriguez, Norcia, McClure, and Steyvers (2016) recently performed a joint modeling study combining a Linear Ballistic Accumulator (LBA; a model of intertemporal choice) with both fMRI and EEG data. They were able to show that LBA combined with both fMRI and EEG data provides better predictions of behavioral data than LBA alone and LBA with either fMRI or EEG data alone (see also Love, 2016).

The second kind of simultaneous modeling approach is the integrative approach (Turner et al., 2017, p. 73; see also Palestro et al., 2018). The integrative approach is like the joint

modeling approach except that hyperparameters do not connect the parameters of the cognitive and neural models. Instead, just one set of parameters predicts both the behavioral data and the neural data. The integrative approach is very difficult technically to implement and computationally intensive (Ibid.). It also requires strong commitments about what cognitive processes are involved and the brain areas that implement them (Ibid.). According to Turner et al. (2017, p. 73), some applications of the cognitive architecture ACT-R have aimed to implement this approach (Borst & Anderson, 2013, 2017), although these applications sometimes more closely resemble the latent input approach. Furthermore, according to Palestro et al. (2018, p. 23), the most accessible implementations of the integrative approach can be viewed simply as instances of the two-stage behavioral approach (which they call the "directed approach").

MBCN represents a major step forward in solving the integration problem, yet its analyses have all been correlational so far. As such, we must be careful not to overstate what such analyses show. They do not, for example, show that any given cognitive model is implemented in the brain, though they provide steps for reaching that conclusion eventually. Correlations between parameters of cognitive models and neuroimaging or neural recording data may provide some, but not much, evidence that the correlated brain region implements the representation or performs the computation specified by the cognitive model parameter. Ritchie, Kaplan, and Klein's (2017; see also Carlson, Goddard, Kaplan, Klein, & Ritchie, 2018) critique of multivariate pattern analysis (MVPA) and neural decoding algorithms echoes similar concerns. To establish that specific neural activity implements a specific computation, the usual causal interventions and manipulations are required, such as activation and inhibition experiments (Craver, 2007; Thomson & Piccinini, 2018). Advances in optogenetics (Deisseroth, 2011) and designer receptors exclusively activated by designer drugs (DREADDs) (Roth, 2016)

have greatly increased the specificity of such causal interventions.

4. MBCN Exemplifies Multifield Mechanistic Integration

Although MBCN does not, and currently cannot, demonstrate which specific neural mechanisms underlie a cognitive process, it does place constraints on what the neural mechanism might be. The mutual constraining of cognitive psychology and cognitive neuroscience that characterizes MBCN is definitive of multifield mechanistic integration. Each of the statistical frameworks and examples of them described above represents a different way of achieving that mutual constraint.

Researchers in MBCN are quite explicit that their studies illustrate mutual constraint and that that is the intended goal of the development of these new statistical frameworks. For example, Turner et al. (2017, pp. 66-68) write that the category of neural constraint approaches "uses neural data as auxiliary information that guides or constrains a behavioral model. … [T]he neural data are considered important, but only in the sense that they inform the mechanisms in the behavioral model". For Forstmann and Wagenmakers (2015, p. 153), the goals of MBCN are to show "how mathematical models can advance cognitive neuroscience, and how cognitive neuroscience can provide constraint for mathematical models". According to Palmeri, Love, and Turner (2017, p. 60), "One [of the factors that drove the development of MBCN] was the recognition on the part of cognitive modelers and mathematical psychologists interested in understanding the mechanisms that brain data is simply additional data by which to constrain and contrast models". Turner, Van Maanen, and Forstmann (2015, pp. 331-332) write that "our goal was to develop a model that could use prestimulus measures of brain activity to better constrain and inform the mechanisms assumed by a cognitive model of choice response time". Quotations such as these are replete in MBCN studies.

By providing mutual constraint from different scientific fields on the mechanisms that underlie and are responsible for cognitive processes, MBCN represents multifield mechanistic integration. This integration has multilevel and intralevel aspects. Building cognitive models that explain cognitive processes is a multilevel endeavor: the cognitive model posits representations and computational processes that are components of a mechanism responsible for the explanandum phenomenon.[8] Cognitive modeling is the province of cognitive psychology, a component field of MBCN. Finding *implementations* of those representations and computational processes is an *intra*level endeavor. For example, in the study mentioned above (Davis, Love, and Preston, 2012), statistically significant correlations were found between SUSTAIN's recognition strength parameter and medial temporal lobe (MTL) regions (call this set of regions "the MTL network"). The relation between the representations and computations SUSTAIN posits and what SUSTAIN is meant to explain (i.e., object categorization) is multilevel – those representations are components of a mechanism responsible for object recognition. But the representation of recognition strength and the neural activity that implements it are not at different mechanistic levels. MBCN realizes (or at least approximates) the mechanistic ideal of explanation by helping to identify potential realizers for the components of cognitive models that are themselves mechanistic explanations of some cognitive capacity. That is, via cognitive models, MBCN identifies potential components of mechanisms for cognitive capacities. Testing whether these potential mechanistic components really are components – whether these potential realizers really are realizers – requires causal experiment.

It is true generally that the representations and computational processes over them that are posited by a cognitive model and the neural mechanisms that implement them are not at different mechanistic levels. Marr's levels are levels of realization, description, or analysis, not

mechanistic levels (Povich & Craver, 2017). Mechanistic levels are compositional, and the implementational level is not literally a part or component of the algorithmic level or computational level. Similarly, functional roles and the structures that play them are not at different mechanistic levels. This is easiest to see if we consider a concrete account of the realization relation.

4.1. Flat Realization and Intralevel Mechanistic Integration

A philosophical account of the realization relation aims to explicate the relation that holds when one thing (object, property, set of properties, etc.) is realized by another (object, property, set of properties, etc.; the relata of the realization relation are themselves controversial). This is the asymmetric relation that holds between functional roles and the structures that play them. It is also the relation that holds between Marr's levels. The two most prominent philosophical accounts of realization are the "flat" (i.e., intralevel) subset account (Kim, 1998; Polger & Shapiro, 2008; Shoemaker, 2007) and the dimensioned (i.e., interlevel) account (Gillett, 2002a, 2002b).

According to a flat account of realization, realizee and realizer are not at different mechanistic levels. A realizer is just something that meets the criteria that serve to identify instances of the realizee (Polger & Shapiro, 2008, p. 217). In metaphysical terms, this is usually cashed out in terms of "higher-level properties": the realizee is the property of having some property that plays a specific functional role and the realizer plays that functional role. Thus, to use an old example, being in pain is possessing the property of having some property that plays the functional role of pain and the firing of C-fibers plays that role.[9] Shoemaker (2007) has cashed this out in different metaphysical terms.[10] On his subset account of realization, property P realizes property Q if and only if the causal powers conferred by Q are a subset of causal powers

conferred by P. According to Shoemaker, this is what it means, metaphysically speaking, to say that P plays the functional Q-role. P may do things that Q does not do (hence the appeal to subsets, rather than identity, of causal powers), but it does everything that Q does and, so, counts as playing the functional role of Q.

According to the dimensioned account (Gillett, 2002a, 2002b), the realization relation is an interlevel relation. It is a relation that holds between a property of an object and the properties of its parts. Specifically, a collection of properties P1-Pn of constituents of an object realize property G in that object if and only if the causal powers conferred by G to that object are conferred in virtue of the causal powers conferred by P1-Pn. For example, the hardness of a diamond is realized by the bonding properties of its atoms because the powers conferred by hardness to the diamond are conferred in virtue of the causal powers conferred by the bonding properties of its atoms (Ibid.).

Although these accounts of realization have generated heated debate (Gillett, 2002a, 2002b; Polger & Shapiro, 2008; Polger, 2004; Shapiro, 2004), Endicott (2011) helpfully suggests that they are not competing accounts. He distinguishes between the 'what' and the 'how' of realization. The intralevel subset account is an account of the former, and the interlevel dimensioned account is an account of the latter. Dimensioned realization is essentially mechanistic explanation – it explains how a structure is able to perform some function in virtue of the properties of its parts and their organization. In contrast, the subset view says in virtue of what that structure *counts* as performing some function in the first place. Thus, the subset view identifies what structure realizes some function and the dimensioned account shows how it realizes that function.[11]

To put this abstract metaphysics in the context of MBCN, MBCN provides suggestive

evidence that the causal powers associated with recognition strength in SUSTAIN are a subset of the causal powers of the MTL network. MBCN thus provides suggestive evidence about *what* in the brain realizes recognition strength. The flat subset account adequately describes the relation between cognitive model parameters (and representations, computations, etc.) and the neural areas that might realize them.[12] Thus, the novel statistical frameworks distinctive of MBCN specifically allow a form of *intra*level integration,[13] though, as I explained earlier, the brain regions identified are components of mechanisms for cognitive capacities. The mechanistic multifield integration of MBCN is thus a complex interplay of intralevel and interlevel integrations.

How does my argument square with claims that MBCN "provides insight into 'how' a particular cognitive function might be implemented in the brain, not only 'where' it is implemented" (O'Doherty et al., 2007, p. 39) and that traditional non-model-based cognitive neuroscience "can tell us *which* brain regions are predictive of a particular behavior and even *by how much*, but they say nothing about neither *how* nor *why* particular brain regions produce said behavior" (Turner et al., 2017, p. 66; emphasis in original)? Here it is important to note that the use of a *cognitive model* is what provides insight into the 'how' and 'why'. The cognitive model has components and they and their interactions explain how some cognitive ability is performed. Those cognitive model components can be mapped onto brain regions. In that way, MBCN can provide insight into how the *cognitive ability* is performed, but not how the *cognitive model components* are implemented. It provides evidence about *what* implements the cognitive model components. To continue with my previous example, by identifying the MTL network as a potential realizer for recognition strength, MBCN can provide insight into how object categorization is performed, but it has not provided any insight into *how* the MTL network

realizes recognition strength.

The subset view of realization also makes explicit why adding implementational details does not constitute *inter*level mechanistic integration: to add implementational details is to extend the set of causal powers possessed at a single mechanistic level beyond the set picked out by a functional concept. The causal powers contained in the proper superset (i.e., the causal powers in the superset not contained in the subset), while they may be functionally irrelevant, can often be used to identify the potential realizer. For example, SUSTAIN's concept of *recognition strength* picks out certain causal powers, and the concept *MTL network* picks out certain causal powers. The former are a subset of the latter. That is what it means for the MTL network to play the recognition strength role. The causal powers in the proper superset include all those powers of the MTL network that are irrelevant to it playing the recognition strength role. However, we can often use those powers (e.g., the power to look a certain way in a microscope or neuroimager) to identify the region on which we should intervene to test the hypothesis that it has the causal powers associated with recognition strength.[14] Interestingly, while it is unlikely that MBCN researchers have thought about multifield integration using these concepts, they come close to endorsing a view similar to the one just presented when they write, "The central assumption of these analyses is that information obtained from either source of data [neural or behavioral] can tell a similar story – albeit in different languages – about some aspect of cognition, and the integration of the these measures assimilates the differences in languages across data modalities" (Turner et al., 2017, p. 66). Here I have presented a philosophical defense and elaboration of that claim.

5. Conclusion

What is revolutionary about cognitive neuroscience is its rejection of the traditional

cognitive science view of autonomous metaphysical and methodological levels: functional/cognitive/computational vs. neural/mechanistic/implementation (Boone & Piccinini, 2016).[15] As a prime example of the integrative multifield mechanistic perspective with the explicit aim of linking cognitive models to the brain, MBCN thus concentrates the most revolutionary elements of cognitive neuroscience. It is the vanguard of the cognitive neuroscience revolution. It has invented new statistical frameworks that allow cognitive psychology and cognitive neuroscience mutually to constrain each other. This multifield mechanistic integration has interlevel and intralevel components. The interlevel component is illustrated by cognitive modeling and the intralevel component is illustrated by efforts to find implementations of those cognitive models. The latter is the distinctive goal of the new statistical frameworks of MBCN.

Endnotes

1. Eronen (2015) argues that the main ideas of mechanists can be cashed out in a deflationary sense of levels that simply combines composition and scale. I think it's fine if one wants to cash out my claims similarly. For example, later I argue that, according to the mechanistic sense of level, a structure and the function it performs are not at different levels; but this is true in Eronen's deflationary sense of level too. Eronen also argues that Craver's same-level criterion, which gives a sufficient condition for sameness of mechanistic level, leads to a branching structure of levels that makes it difficult to align scientific models. However, Povich and Craver basically accept Eronen's critique of this criterion. After arguing that there is only a necessary condition for sameness of mechanistic level, they write: "But one might just as easily say on this basis that sameness of level has no deep conceptual significance for mechanistic levels. … For some, this is tantamount to abandoning the idea of levels (e.g., Eronen 2015). We see our account

rather as a distillate of the ordinary scientific concept, an extraction of an explanatorily and metaphysically central idea, leaving behind as residue the problematic commitments inherent in our inchoate, folk talk of 'levels'" (Povich & Craver, 2017, p. 188). I thank the editors for pressing me to clarify this.

2. Sometimes causal-etiological and contextual explanations are also considered species of mechanistic explanation (Bechtel, 2011). I will not be concerned with those here.

3. As the editors point out, it is not as plausible that mathematical psychology is itself a distinct field in Darden and Maull's sense. I thank them for helping me clarify this section.

4. An early exception (and an early refutation of Coltheart, 2006) that nicely illustrates multifield integration is Petersen, Fox, Posner, Mintun, and Raichle's (1988) PET study of single-word processing. In that study, PET data were used as additional constraints on cognitive models of lexical processing. Their results favored a parallel, rather than serial, processing model.

5. The development MBCN arguably fits nicely with Bickle's (2016) account of tool-driven scientific revolution.

6. These models are the topic of Weiskopf's (2011) criticisms of mechanistic explanation. See Povich (2015) for a response.

7. Because this procedure treats cognitive and neural model parameters as covariates, Palestro et al. (2018) call it the "covariation approach" to joint modeling and use "joint modeling" as a broader term that includes the two-stage behavioral approach (which they call the "directed approach") and the integrative approach, which I discuss next.

8. The claim that cognitive models provide mechanistic explanations is controversial. For a defense of it, see Piccinini & Craver (2011) and Povich (2015).

9. This talk of properties and properties of having properties is meant to accommodate multiple

realizability. If pain is multiply realizable, it cannot be strictly identified with, for example, C-fibers firing. Appealing to "higher-level" properties of having properties is supposed to respect this intuition while providing a way to remain metaphysically materialist.

10. Or he uses the same metaphysical terms but cashes out "higher-level properties" in a specific way (see Shoemaker, 2007).

11. Piccinini and Maley (2014) appear confusingly to run the what and the how together in their account.

12. The subset view thus aligns nicely with the idea that "levels of realization" are levels of abstraction, an idea espoused by earlier mechanists (Bechtel & Mundale, 1999). Computational models are more abstract descriptions and neural models are less abstract descriptions of the same reality.

13. Boone & Piccinini (2016) mistakenly characterize MBCN's mechanistic integration as exclusively multilevel, but the kind of integration on which MBCN is distinctively focused is intralevel.

14. In other words, it is no part of the function of recognition strength to look a certain way to a neuroscientist, though a neuroscientist can only find a potential realizer of recognition strength by its effect on neuroimaging or electrophysiological recording devices, after statistical analysis. I suspect this is a general feature of the structure-function relation, sans any structures whose function just is to be perceived by us in a certain way. For example, it is no part of the function of a corkscrew to look a certain way to me, though that is how I find one when I open the kitchen drawer. Actually to test whether what looks like a corkscrew is a corkscrew, I intervene on relevant properties. I do not wish here to defend this more general claim though – I only wish to claim that irrelevant properties can be and sometimes are used to find potential realizers.

Actually testing whether they are realizers requires intervening on (putatively) relevant properties.

15. The extent to which autonomy held sway among early working cognitive scientists can be debated. For example, Marr, although he approached his three levels in a relatively top-down fashion, seems to have been against autonomy in the sense at issue. However, autonomy was certainly present in, for example, Pylyshyn (1984), who was influenced by Fodor. See also the relevant references listed by Boone and Piccinini (2016, p. 1513). Autonomy is also how "early cognitive science" is portrayed in contemporary cognitive science textbooks, such as Bermudez's (2014, section 3.1) influential textbook. I thank an anonymous referee for pressing me here.

References

Bechtel, W. (2009). Looking down, around, and up: Mechanistic explanation in psychology. *Philosophical Psychology*, *22*, 543-564. https://doi.org/10.1080/09515080903238948

Bechtel, W., & Mundale, J. (1999). Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of science*, *66*(2), 175-207. https://doi.org/10.1086/392683

Behrens, T., Woolrich, M., Walton, M., & Rushworth, M. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*, 1214–1221. https://doi.org/10.1038/nn1954

Bermúdez, J. L. (2014). *Cognitive science: An introduction to the science of the mind*. Cambridge, MA: Cambridge University Press. https://doi.org/10.1017/cbo9781107279889

Bickle, J. (2016). Revolutions in neuroscience: Tool development. *Frontiers in Systems Neuroscience*, *10*, 24. https://doi.org/10.3389/fnsys.2016.00024

Boone, W., & Piccinini, G. (2016). The cognitive neuroscience revolution. *Synthese*, *193*, 1509-

1534. https://doi.org/10.1007/s11229-015-0783-4

Borst, J. P., & Anderson, J. R. (2013). Using model-based functional MRI to locate working memory updates and declarative memory retrievals in the frontoparietal network. *Proceedings of the National Academy of Sciences of the United States*, *110*, 1628–1633. https://doi.org/10.1073/pnas.1221572110

Borst, J. P., & Anderson, J. R. (2017). A step-by-step tutorial on using the cognitive architecture ACT-R in combination with fMRI data. *Journal of Mathematical Psychology*, *76*, 94-103. https://doi.org/10.1016/j.jmp.2016.05.005

Carlson, T., Goddard, E., Kaplan, D. M., Klein, C., & Ritchie, J. B. (2018). Ghosts in machine learning for cognitive neuroscience: moving from data to theory. *NeuroImage*, *180*, 88-100. https://doi.org/10.1016/j.neuroimage.2017.08.019

Coltheart, M. (2006). What has functional neuroimaging told us about the mind (so far)? *Cortex*, *42*(3), 323-331. https://doi.org/10.1016/s0010-9452(08)70358-7

Chirimuuta, M. (2014). Minimal models and canonical neural computations: The distinctness of computational explanation in neuroscience. *Synthese*, *191*, 127–53. https://doi.org/10.1007/s11229-013-0369-y

Craver, C. F. (2005). Beyond reduction: mechanisms, multifield integration and the unity of neuroscience. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*, 373-395. https://doi.org/10.1016/j.shpsc.2005.03.008

Craver, C. F. (2007). *Explaining the brain*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199299317.001.0001

Craver, C. F., & Darden, L. (2013). *In search of mechanisms: Discoveries across the life*

sciences. Chicago: University of Chicago Press. https://doi.org/10.7208/chicago/9780226039824.001.0001

Darden, L., & Maull, N. (1977). Interfield theories. *Philosophy of Science*, *44*, 43–64. https://doi.org/10.1086/288723

Davis, T., Love, B. & Preston, A. (2012). Learning the exception to the rule: Model-based fMRI reveals specialized representations for surprising category members. *Cerebral Cortex*, *22*, 260–73. https://doi.org/10.1093/cercor/bhr036

Deisseroth, K. (2011). Optogenetics. *Nature Methods*, *8*, 26-29. https://doi.org/10.1038/nmeth.f.324

Endicott, R. P. (2011). Flat versus dimensioned: The what and the how of functional realization. *Journal of Philosophical Research*, *36*, 191–208. https://doi.org/10.5840/jpr_2011_13

Eronen, M. I. (2015). Levels of organization: A deflationary account. *Biology & Philosophy*, *30*(1), 39–58. https://doi.org/10.1007/s10539-014-9461-z

Forstmann, B. U., Wagenmakers, E.-J., Eichele, T., Brown, S., & Serences, J. T. (2011). Reciprocal relations between cognitive neuroscience and formal cognitive models: opposites attract? *Trends in Cognitive Sciences*, *15*, 272-279. https://doi.org/10.1016/j.tics.2011.04.002

Forstmann, B. U., & Wagenmakers, E.-J. (2015). Model-based cognitive neuroscience: A conceptual introduction. In B. U. Forstmann & E.-J. Wagenmakers (Eds.), *An introduction to model-based cognitive neuroscience* (pp. 139-156). New York, NY: Springer. https://doi.org/10.1007/978-1-4939-2236-9_7

Gillett, C. (2002a). The dimensions of realization: A critique of the standard view. *Analysis*, *62*(276), 316-323. https://doi.org/10.1111/1467-8284.00377

Gillett, C. (2002b). The metaphysics of realization, multiple realizability, and the special sciences. *The Journal of Philosophy*, *100*(11), 591-603. Retrieved from http://www.jstor.org/stable/3655746

Gläscher, J. P., & O'Doherty, J. P. (2010). Model-based approaches to neuroimaging: combining reinforcement learning theory with fMRI data. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(4) 501-510. https://doi.org/10.1002/wcs.57

Hempel, C. (1965). *Aspects of scientific explanation*. New York, NY: Free Press.

Kim, J. (1998). *Mind in a physical world*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/4629.001.0001

Love, B. C. (2015). The algorithmic level is the bridge between computation and brain. *Topics in Cognitive Science*, *7*(2), 230-242. https://doi.org/10.1111/tops.12131

Love, B. C. (2016). Cognitive models as bridge between brain and behavior. *Trends in Cognitive Sciences*, *20*(4), 247-248. https://doi.org/10.1016/j.tics.2016.02.006

Machamer, P., Darden, L. & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*, 1–25. https://doi.org/10.1086/392759

Mack, M. L., Preston, A. R. & Love, B. C. (2013). Decoding the brain's algorithm for categorization from its neural implementation. *Current Biology*, *23*(20), 2023-2027. https://doi.org/10.1016/j.cub.2013.08.035

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: Freeman.

Maull, N. (1977). Unifying science without reduction. *Studies in History and Philosophy of Science*, *8*, 143–162. https://doi.org/10.1016/0039-3681(77)90012-7

Nagel, E. (1961). *The structure of science*. New York: Harcourt, Brace and World.

Newell, A. (1980). Physical symbol systems. *Cognitive Science*, *4*, 135–183. https://doi.org/10.1207/s15516709cog0402_2

O'Doherty, J. P., Hampton, A. & Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, *1104*(1), 35-53. https://doi.org/10.1196/annals.1390.022

Oppenheim, P. & Putnam, H. (1958). Unity of science as a working hypothesis. In H. Feigl, M. Scriven, & G. Maxwell (Eds.), *Concepts, theories, and the mind-body problem* (pp. 3–36). Minneapolis, MN: University of Minnesota Press.

Palestro, J. J., Bahg, G., Sederberg, P. B., Lu, Z.-L., Steyvers, M., & Turner, B. M. (2018). A tutorial on joint models of neural and behavioral measures of cognition. *Journal of Mathematical Psychology*, *84*, 20-48. https://doi.org/10.1016/j.jmp.2018.03.003

Palmeri, T., Schall, J., & Logan, G. (2015). Neurocognitive modelling of perceptual decisions. In J. R. Busemeyer, J. Townsend, Z. J. Wang, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 320-340). Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199957996.013.15

Palmeri, T. J., Love, B. C. & Turner, B. M. (2017). Model-based cognitive neuroscience. *Journal of Mathematical Psychology*, *76*, 59-64. https://doi.org/10.1016/j.jmp.2016.10.010

Petersen, S. E., Fox, P. T., Posner, M. I., Mintun, M., & Raichle, M. E. (1988). Positron emission tomographic studies of the cortical anatomy of single-word processing. *Nature*, *331*(6157), 585-589. https://doi.org/10.1038/331585a0

Piccinini, G., & Craver, C. F. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183*(3), 283-311. https://doi.org/10.1007/s11229-011-9898-4

Piccinini, G., & Maley, C. J. (2014). The metaphysics of mind and the multiple sources of multiple realizability. In M. Sprevak and J. Kallestrup (Eds.), *New waves in philosophy of mind* (pp. 125-152). Palgrave Macmillan, London. https://doi.org/10.1057/9781137286734_7

Polger, T. W. (2006). *Natural minds*. Cambridge: MIT Press. https://doi.org/10.7551/mitpress/4863.001.0001

Polger, T. W., & Shapiro, L. A. (2008). Understanding the dimensions of realization. *The Journal of Philosophy*, *105*(4), 213-222. https://doi.org/10.5840/jphil2008105415

Povich, M. (2015). Mechanisms and model-based functional magnetic resonance imaging. *Philosophy of Science*, *82*(5), 1035–46. https://doi.org/10.1086/683438

Povich, M. (Forthcoming). Mechanistic explanation in psychology. In H. Stam & H. Looren de Jong (Eds.), *The Sage handbook of theoretical psychology*.

Povich, M., & Craver, C. F. (2018). Mechanistic levels, reduction, and emergence. In S. Glennan & P. Illari (Eds.), *The Routledge handbook of mechanisms and mechanical philosophy* (pp. 185-197). London: Routledge. https://doi.org/10.4324/9781315731544-14

Pylyshyn, Z. W. (1984). *Computation and cognition*. Cambridge, MA: MIT Press.

Rice, C. (2015). Moving beyond causes: Optimality models and scientific explanation. *Noûs*, *49*(3): 589–615.

Ritchie, J. B., Kaplan, D. M., & Klein, C. (2017). Decoding the brain: Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *The British Journal for the Philosophy of Science*. https://doi.org/10.1101/127233

Roth, B. L. (2016). DREADDs for neuroscientists. *Neuron*, *89*(4), 683-694. https://doi.org/10.1016/j.neuron.2016.01.040

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton: Princeton University Press.

Salmon, W. C. (1989). Four decades of scientific explanation. In P. Kitcher, & W. C. Salmon (Eds.), *Scientific explanation* (pp. 3–219). Minnesota Studies in the Philosophy of Science, XVIII. Minneapolis: University of Minnesota Press.

Schaffner, K. F. (1993). *Discovery and explanation in biology and medicine*. Chicago: University of Chicago Press.

Shapiro, L. A. (2004). *The mind incarnate*. Cambridge: MIT Press.

Shoemaker, S. (2007). *Physical realization*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199214396.001.0001

Thomson, E. & Piccinini, G. (2018). Neural representations observed. *Minds and Machines, 28*(1), 191-235. https://doi.org/10.1007/s11023-018-9459-4

Turner, B. M., (2015). Constraining cognitive abstractions through Bayesian modeling. In B. U. Forstmann & E.-J. Wagenmakers (Eds.), *An introduction to model-based cognitive neuroscience* (pp. 199-220). New York, NY: Springer. https://doi.org/10.1007/978-1-4939-2236-9_10

Turner, B. M., Forstmann, B. U., Love, B. C., Palmeri, T. J., & Van Maanen, L. (2017). Approaches to analysis in model-based cognitive neuroscience. *Journal of Mathematical Psychology*, *76*, 65-79. https://doi.org/10.1016/j.jmp.2016.01.001

Turner, B. M., Forstmann, B. U., Wagenmakers, E. J., Brown, S. D., Sederberg, P. B., & Steyvers, M. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. *NeuroImage*, *72*, 193-206. https://doi.org/10.1016/j.neuroimage.2013.01.048

Turner, B. M., Rodriguez, C. A., Norcia, T. M., McClure, S. M., & Steyvers, M. (2016). Why more is better: Simultaneous modeling of EEG, fMRI, and behavioral data. *Neuroimage*, *128*, 96-115. https://doi.org/10.1016/j.neuroimage.2015.12.030

Turner, B. M., Van Maanen, L., & Forstmann, B. U. (2015). Informing cognitive abstractions through neuroimaging: The neural drift diffusion model. *Psychological Review*, *122*(2), 312. https://doi.org/10.1037/a0038894

Weiskopf, D. A. (2011). Models and mechanisms in psychological explanation. *Synthese*, *183*(3), 313–38. https://doi.org/10.1007/s11229-011-9958-9

Wong, K.-F., & Wang, X.-J. (2006). A recurrent network mechanism of time integration in perceptual decisions. *The Journal of Neuroscience*, *26*, 1314–1328. https://doi.org/10.1523/jneurosci.3733-05.2006

Ylikoski, P., & Kuorikoski, J. (2010). Dissecting explanatory power. *Philosophical Studies*, *148*(2), 201-219. https://doi.org/10.1007/s11098-008-9324-z