Reductive Explanation Between Psychology and Neuroscience

Daniel A. Weiskopf

1. Introduction

Reductionism is one of the most divisive concepts in the popular and philosophical lexicon. Over the past century it has been championed, declared dead, resurrected, and reformed many times over. Its protean character reflects the circumstances of its birth in the polarizing mid-20th century debates over the unity of science. While the totalizing ideal of unified science has lost its luster, localized reductionist projects continue to flourish. In this chapter I sketch the goals and methods of one prominent form of reductionism within the mind-brain sciences and consider the prospects for non-reductionist alternatives.

2. Defining reductive explanation

First, we can distinguish *ontological* and *methodological* reductionism.[1] *Ontological reductionism* centers on the question of whether the kinds appealed to in psychology are ultimately anything "over and above" those appealed to in neuroscience. Reductionists hold that taxonomic distinctions among psychological kinds will align with those made among neuroscientific kinds, so that the way that psychology carves up its domain simply falls out of the way that neuroscience does. Ontological antireductionism maintains that psychological kinds are not necessarily visible using only the classificatory apparatus of neuroscience.

Methodological reductionism claims that the explanatory constructs of psychology are ultimately dispensable in favor of those drawn from neuroscience. Psychological explanations

[1] This distinction is explicated in Schaffner (1993). For discussion of other types of reduction and their relationships, see Bickle (2006; Theurer & Bickle, 2013), Horst (2007), Kaiser (2015), Kim (2005, 2008), Sarkar (1992), Theurer (2013), and Wimsatt (1974, 2006).

are only epistemic stopgaps that will, in the end, turn out to be replaceable by neuroscientific explanations. Methodological antireductionists, by contrast, champion the autonomy of psychological explanation. They hold that psychological explanations either cannot be fully dispensed with in favor of neuroscientific ones, or at least that they *need* not be.

Reductive explanation is an interfield project. A *field* centers on a set of problems, relevant facts and phenomena that bear on their solutions, explanatory goals and norms, and distinctive experimental techniques, materials, and methods (Darden & Maull, 1977). The question is how the problems, phenomena, and explanations generated within one field can be related to those in the other, given that they may have strikingly different ontologies and explanatory frameworks (Poeppel & Embick, 2005). Answering this question requires developing specially tailored interfield theories.

Within the classic Nagelian framework, reduction was an intertheoretic relation: a theory T_1, identified with a systematic body of laws, reduces to another theory T_2 when T_1 can be logically derived from T_2, under certain background conditions. When theories are drawn from different fields or domains that each employ their own specialized vocabulary, a set of connecting ("bridge") principles linking these terms is required to allow the deduction to go through. The condition of *connectability* ensures that the ontology of the two theories can be aligned appropriately, while the condition of *derivability* shows why the laws of the reduced theory *must* hold, given the lower level laws and these connections.

Nagel's model illustrates the constraints that have traditionally governed philosophical accounts of reduction.[2] An adequate interfield reduction should have two characteristics. First, it should preserve the ontology of the reduced field. Ontological conservatism is what separates

[2] Kenneth Schaffner's Generalized Reduction/Replacement model (1993) is the most extensive attempt to preserve the basic insights of Nagelian reduction.

reduction from straightforward elimination. This means that something like Nagel's connectability principle must be part of any non-eliminative reduction. Second, it should preserve (and possibly extend) the reduced field's explanatory insights. Moving to the reducing field should not involve a major loss of explanatory power or generality. Both of these are matters of degree, since pruning, revision, and parameterization of the phenomena and generalizations of both fields is standard in interfield mapping.

An interfield theory aims to show how and why the elements of the participating fields relate systematically to each other. To offer a *reductive* explanation, an interfield theory should show how the ontology and explanatory constructs of the reduced field systematically *depend* on those of the reducing field. That is, it should give a theoretically illuminating account of how the kinds posited in psychology are *realized* or *implemented* by their physical substrate. The dependence condition is crucial because interfield theories exist in many non-reductive contexts of inquiry. We can attempt to integrate two fields or two models, such as general relativity and quantum field theory, by subsuming them in a single framework without assuming that one must be reduced to the other.

What distinguishes reductionist interfield theorizing is that it is both *conservative* and *directional*: the ontology and explanatory content of one field depends on that of another, such that the existence of the higher categories, as well as the explanations that they are part of, can be accounted for solely by the existence and activities of entities within the lower field.

Finally, reductionism is associated with a set of distinctive heuristics and research strategies (Wimsatt, 2006). While these do not strictly define a reductionist project they are strongly indicative of one. They include attempting to articulate a system's microcomponents that serve as the sole causal basis for generating and explaining its macrolevel properties and

behavior (a maneuver Rob Wilson (2004) dubs “smallism”), positing identities between entities and processes in the target and the reducing field (McCauley & Bechtel, 2001), and localizing higher-level functions within discrete microcomponents. Reductionist strategies tend to represent the direction of ontological priority, control, and explanation within a system as being internal and bottom-up. The more it is necessary to draw on external, higher level, and contextually variable factors in explaining the properties of a system, the less traction these heuristics will get.

3. The ontology and methodology of cognitive modeling

Cognitive modeling is one of the main tools used in psychology to describe and understand the mental and behavioral capacities of humans and other organisms. A *cognitive model* includes:

> 1) a characterization of the *target cognitive capacity* itself: its input–output profile, the experimentally derived phenomena associated with it, its distinctive patterns of effects, its normal and abnormal developmental trajectory, and its relationships to other cognitive capacities;
>
> 2) an *ontological inventory*: a set of representational vehicles along with their distinctive properties such as format and informational content, processes that create, combine, store, retrieve, and otherwise transform and operate over them, and resources that can be used in this processing such as memory registers and attentional allocation and processing cycles;

3) an *organization* or *structure*: a specification of the way that these ontological elements are grouped into stable cognitive systems, the regularities and laws that they obey, the rules and paths of influence by which they can interact and influence one another, the control structures that dictate what operations will happen when, and the basic properties of the cognitive architecture in which they are embedded.

Often the role of cognitive modeling is not merely to inferentially bridge the gap between behavior and brain structures (Love, 2016), but to do so by describing real psychological structures. *Realism* is a causal thesis: a realistic model should describe causally active elements and operations of the cognitive system, such that the dynamics of these elements is capable of producing the input–output profile and patterns of phenomena associated with the target capacity. The posited elements should not simply be instrumental fictions, useful in generating predictions but not themselves influencing the system's behavior.[4]

Causally interpreted cognitive models make several commitments with respect to their elements. First, these elements should be the sorts of things that can be manipulated and intervened on to produce specific effects. A model can be regarded as describing part of the total cognitive state that the system can be in. The set of possible states corresponds to possible assignments of values to variables regarding the system's representations, processes, and resources. The model should accurately capture the patterns of manipulation of these state

[4] A caveat: some cognitive modelers aim only for empirical adequacy, and even mechanistic models can have non-realist interpretations (Colombo, Hartmann, & Van Iersel, 2015). Causal interpretation is often partial, meaning that only some elements are treated in a realistic way, while others are fictions or simplifications that facilitate ends such as computational simulation. Little guidance is provided by a model itself as to which of its components should be interpreted causally, hence it is not always straightforward to determine the degree of a model's realistic commitments. The thesis of realism is not the same as Kaplan and Craver's (2011) "3M constraint", since that requires that model elements map onto components of mechanisms, while realism only requires that there be a mapping to causally significant components, without assuming that these must be mechanistically organized. The relevance of this will become clear in later sections.

variables that are possible and the range of effects that varying each element produces. Cognitive models are maps of salient *intervention points* in a psychological system.

Second, these elements should be *robust* (Wimsatt, 1994): there should be a set of independent but converging operations (measurement procedures, experimental protocols, and the like) that can detect and track the state of the system's components. The greater the number of distinct epistemic pathways that exist to detect a component the more confidence we can have that it exists apart from our schemes of modeling and measurement.

Third, the model should successfully generate *predictions* about how the cognitive system will behave under various conditions that conform to the observed phenomena, and also generalize in a natural (non-ad hoc) way to new results. This is a basic criterion of empirical adequacy. Numerous model-fitting techniques can be applied to see whether the model is capable of capturing a dataset.

Cognitive models themselves are neutral with respect to physical or neurobiological structures.[5] The aim of integrating cognitive and neural models is twofold. With respect to a single cognitive model, showing that it can be successfully neurally integrated is thought to provide extra evidence in its favor. With respect to two or more cognitive models that are equivalent in terms of their predictive value, the ability to better integrate one model rather than the others provides some evidential advantage for it. In both scenarios, appeal to the *neural plausibility* of cognitive models provides a field-external constraint on psychological theorizing (Butler, 1994).

Mack, Preston, and Love (2013) offer a nice example of the latter use of neural data in model adjudication. Both prototype and exemplar models of categorization can capture the same

[5] In this sense they are functional kinds, like many others that occur in the special sciences. For discussion and critique of the notion of a functional kind, see Reydon (2009), Weiskopf (2011b), and Buckner (2014).

range of behavioral data, which has led to a stalemate between the two views. However, latent parameters of the two models corresponding to the degree of match between a stimulus and a stored representation can be correlated with global and local patterns of fMRI-measured brain activity in participants who are performing categorization tasks. These correlations suggest that many participants are using exemplar strategies, though a substantial minority use prototype or mixed approaches. These patterns can also be used to isolate regions to be investigated in future studies, meaning that not only can cognitive models be discriminated, hypotheses about their implementations may also be framed. We turn now to one specific proposal about the form these hypotheses might take.

4. Mechanistic integration as a reductive interfield strategy

Many scientific fields center on discovering and elucidating mechanisms (Andersen, 2014a, 2014b, Bechtel, 2008, 2009; Bechtel & Abrahamsen, 2005; Craver, 2007; Craver & Darden, 2013; Darden, 2001; Glennan, 2002; Zednik, 2015). The explanatory target of mechanistic analysis is an entity or system's function—its capacity to carry out a certain sort of activity or to fill a causal role. Humans have the capacity to store and retrieve a limited number of items from memory; the liver has the function of removing toxins from the bloodstream; pyramidal cells have the function of generating action potentials. The question of how each function is carried out is answered by specifying a mechanism.

Mechanisms are organized sets of entities plus their associated activities and processes. Mechanistic analysis is a species of componential causal analysis: it requires decomposing the target system into its component parts, placing them within the overall organization of the system, locating them relative to one another, and understanding their activities and operations:

what they do and how they contribute to the performance of the system's overall function. Wimsatt (2007) explicitly links this form of mechanistic analysis with reduction: "a reductive explanation of a behavior or a property of a system is one that shows it to be mechanistically explicable in terms of the properties of and interactions among the parts of a system" (pp. 670-1).

Mechanistic discovery is often intrafield, as in the case of applying cellular and molecular techniques to understand how the action potential is generated or the process by which neurotransmitters are transported and released into synaptic gaps. When applied as an interfield strategy, mechanistic analysis involves correlating the elements of our cognitive ontology with those of neuroscience, or in Kaplan and Craver's (2011) terms, discovering a *model-to-mechanism mapping*. In practice this implies a strong preference for localizing elements and operations of the cognitive system in discrete, spatially contiguous, "natural"-seeming components of the neural system (Coltheart, 2013). An integration is successful to the extent that this sort of localized mapping of cognitive onto neural structures preserves the explanatory power of the original cognitive model to capture its distinctive phenomena and effects.

When applied to computational models, interfield mapping requires showing how computational states and processes are implemented in neural hardware. Many competing theories of computational implementation exist (Chalmers, 1994, 1996, 2011; Copeland, 1996; Gallistel & King, 2010; Miłkowski, 2011; Rescorla, 2014; Shagrir, 2012). These typically involve imposing some form of structural constraints (normally spelled out in physical or causal terms) on the underlying hardware. One that explicitly draws on mechanistic insights is due to Gualtiero Piccinini (2015). On Piccinini's account, a *computing system* is identified with a kind of mechanism that has the function of carrying out generic computations: "the processing of vehicles by a functional mechanism according to rules that are sensitive solely to differences

between different portions (i.e., spatiotemporal parts) of the vehicles" (p. 121), where these rules conform to a mathematical function from inputs and current states to outputs. The vehicles of computation consist of "spatiotemporal parts or portions", often a concatenated sequence of digits that can take on finitely many discrete states.[6]

Piccinini emphasizes that while computation is "medium independent" (p. 122) in that it only attends to some of the physical properties of the implementing medium and not others, nevertheless implementing a computation places tight structural constraints on the nature of computational realizers. These must be met for it to be true that a system is computing a certain function: "In real systems, structural components lie between the input and the output, and they are organized to exhibit that capacity. Whether the components satisfy the given task analysis depends on whether the components include structures that complete each of the tasks. If no such structures can be identified in the system, then the task analysis must be incorrect" (p. 90).

Such comments reveal a commitment to *componential realism*: in order for a cognitive model to offer a good causal explanation, there need to be real structural components of a neurobiological mechanism that correspond to the elements of that model and that carry out the operations of those elements. Elements of cognitive models, in short, must map onto mechanistic structural components of the brain. In the absence of such a mapping, the explanations that the model offers are simply false, and accepting them would be "to give up on the idea that there is a uniquely correct explanation" (p. 91) of the system's behavior. The idea that distinct computational elements must map onto independent physical components (however those are identified) also plays a key role in Chalmers' (1996) theory of implementation, and it embodies a similar componential realist thought.

[6] Symbols are strings of digits that are potentially semantically interpreted. Though Piccinini doesn't hold that semantics is essential to computational description, models in computational psychology typically treat them as vehicles of thought, i.e., as having content that is relevant to the cognitive task the system is carrying out.

A similar structural condition is proposed by Bechtel and Hamilton (2007) in discussing the role of localization heuristics in mechanistic analysis: “Discovery of an operation that cannot be linked to a part of the structure poses the question of whether that operation is indeed being performed and if so, by what component.” The ability to localize functions in identifiable parts of structures underwrites mappings between entities in different fields. Operations that can’t be tied to working parts of mechanisms are suspect.

A successful mechanistic reduction will map elements of a model onto mechanistic parts in such a way that the workings of those parts causally explains the phenomena that the model does. Insofar as it offers such direct constraints, the componential realism condition offers an advance over previous criteria of neural plausibility, which were grounded in an often-impressionistic sense of similarity between cognitive models and the brain. The element-to-part aspect of the mapping guarantees ontological conservatism. It further satisfies the directional explanatory condition on reduction: the causal operations of neural components explain how they carry out the functions ascribed by the model.

5. Limits of mechanistic reduction

The success of mechanistic reduction turns on the existence of a smooth mapping from cognitive models onto neural mechanisms. However, it is at present an open question how well-aligned these mappings will be. Several writers have raised doubts about this possibility, proposing instead that cognitive and neural models may *cross-classify* the causal structure of the brain (Shapiro, 2015; Stinson, 2016; Weiskopf, 2011a, 2016), meaning that elements of the former often cannot be correlated with elements of the latter. While few have argued that *no*

cognitive elements map onto neural mechanisms, that there might be such failures as a matter of course has been explored seriously.

Stinson, for example, notes that research in the psychology of attention and memory has generated models that do not in any obvious way map onto underlying neural systems. Attentional models use diagrams of information flow through a sequence of layers consisting of perceptual processors, memory stores of varying durations, filters, selectors, and channels. Other elements, such as sequences of encoders for generating representations of particular types of properties, and controllers where higher-level intentional processing can direct the flow of information, are also present. In practice these models are assessed autonomously: while they can be *used* as templates for neural localization, the fact that these elements may not correspond with distinct parts of neural mechanisms has not led to their abandonment within psychology.

Consider a case of interfield taxonomic mismatch. Psychologists distinguish semantic, episodic, and autobiographical forms of declarative memory, and the distinctions among these types has traditionally been made on the basis of behavioral and lesion studies (Tulving, 1983). It was initially suggested that each cognitive system could be localized in a distinct neural region. However, Burianova and colleagues (Burianova & Grady, 2007; Burianova, McIntosh, & Grady, 2010) showed that coordinated activity in a common network of areas including the left lingual gyrus, left hippocampus, and right caudate nucleus is implicated in all three forms of retrieval. The three forms share a substantial, albeit distributed, anatomical basis. If mechanisms are spatially and structurally delimited, this seems prima facie reason to conclude that these are not in fact separate cognitive kinds, since they share a common realizer (Greenberg & Verfaellie, 2010).

Recent work in systems neuroscience suggests there are principled reasons that such mismatches, in which cognitive kinds are split, fused, and intermingled at the level of neural implementation, may be common. The *massive redeployment* (or *neural recycling*) *hypothesis* proposes that the brain consists of a set of interconnected and highly multifunctional processing units that are reused in tasks across many different domains (Anderson, 2010, 2014, 2015; Dehaene, 2011; Dehaene & Cohen, 2007). Successful task performance, particularly in higher cognition, involves coordinating processing across a distributed suite of regions whose activities can be dynamically reconfigured to execute many distinct cognitive functions (Sporns, 2011).

This hypothesis has several consequences. The first is that much of our cognitive activity is neurally *distributed* or *holistic*: particular cognitive operations and entities are spread out across a broad network of brain regions. The second is that neural components are highly *multifunctional*: these regions each participate in and contribute towards the execution of many distinct cognitive functions. The third is that neural activation is *context-sensitive* and *non-local*: the contribution that each component region makes towards carrying out these functions is determined in part by the ongoing behavior of the other components that have been recruited and the overall task being executed (Bechtel, 2012; Bressler & Kelso, 2016; Meehan & Bressler, 2012; Nathan & Del Pinal, 2015).

Points one and two imply a striking consequence, namely that the neural basis for many distinct cognitive functions may be intractably *entangled*. Entanglement occurs when it is difficult or impossible to pull the spatial realization of one function apart from that of another. There are several ways in which spatially overlapping realizers may make distinct functional contributions. There might be rapid physical reconfiguration of the neural wiring within a region, or different properties of a fixed wiring configuration might allow it to execute multiple

functions when put in various activation contexts. In these cases the same region might contain several overlapping physical parts from the standpoint of cognitive realization. The declarative retrieval network mentioned above may be an example. Another possibility is that a single physical structure is inherently multifunctional, so that several cognitive elements genuinely correspond to only a single neural component. In less radical cases, there will still be significant overlap between the network regions implicated in many different functions (Crossley et al., 2013).

To insist that functional differences *must* force us to type physical structures differently is to illegitimately impose a classification scheme that makes the 3M constraint true by fiat. The phenomenon of entangled realizers poses a challenge to reduction because it is often assumed that wholly distinct cognitive elements must be mapped onto wholly distinct neural mechanisms (or parts of mechanisms). This is not strictly implied by componential realism, but it is closely allied with it: what *makes* a cognitive element real is its distinctive mechanistic realization, and to the extent that such failures of fit arise, they are evidence against the posited cognitive ontology (Poldrack, 2010).

The third point poses an additional challenge to many approaches to mechanism. As Woodward (2013) has argued, mechanisms are the class of componential causal explanations in which the behavior of the components obeys principles of stability, modularity and fine-tuning. In many types of dynamical systems, including neural networks and genetic regulatory networks, however, these constraints may not apply.[7] In particular, the processing contribution of each region seems to depend non-locally on the contributions of other regions and on the overall cognitive function that is being carried out: what a part is doing in a context depends on what

[7] The objection from failures of modularity is also pressed by Fagan (2012), who offers a revised mechanistic account in response to it. See Glymour (2007) for a metaphysical picture similar to the one sketched here.

many other parts are also doing, and changing the cognitive context can also change the contribution made by that particular part (Sporns, 2014).

Since explaining the behavior of the parts may require referring not only to causal factors outside of the parts themselves (looking outwards) but also to factors individuated at the cognitive level (looking upwards), a reductionist explanatory strategy that attempts to capture the cognitive properties of the system purely in terms of bottom-up interactions among stable components may not be appropriate. The more it is necessary to look outwards and upwards in describing the contextual parameters that govern the system's behavior—and even the behavior of the parts themselves—the less reductionist the integration becomes.

Some mechanists respond to these worries by claiming that mechanisms need not obey constraints of spatiotemporal contiguity and unity, so that their parts may potentially be widely scattered and intermingled (Piccinini & Craver, 2011). But it is important to resist liberalizing the notion of a mechanistic part or component to include *anything* that an element of a cognitive model can be mapped onto. "Parts" must have greater integrity than this, and such a move trivializes the mechanistic thesis by declaring anything that realizes a cognitive structure to be, de facto, a mechanism.

The mechanist program initially took enormous care to explicate the constraints on mechanisms, and to separate mechanistic explanation from other approaches to complex systems (Bechtel & Richardson, 2010, p. 147).[8] Accordingly, we ought to preserve conceptual space for the possibility of non-mechanistic realization of complex functions. By collapsing the notion of a mechanism into the notion of a realizer, the thesis loses its force as an empirical hypothesis about the best research strategies for understanding systems like the mind/brain.

[8] For related worries about the cogency of the reductionist's notion of a mechanistic part, see Franklin-Hall (2016), Nicholson (2012), and Teller (2010).

6. Towards a realistic antireductionism

The challenges just canvassed can be summed up as follows. First, the network structure of the brain might not resemble the paradigm examples of mechanisms. Second, even if these dynamical brain networks are best understood as mechanisms, it might not be possible to understand how they function in a reductive, bottom-up fashion. Third, even if a bottom-up mechanistic analysis is possible, this analysis might not include spatially circumscribed working parts that correspond neatly to the elements in our best-confirmed cognitive models.

These nested possibilities represent three different failure modes for the program of mechanistic reductive analysis: in one case, there are no mechanisms to map cognitive operations onto, and in the other two cases the mapping fails to be reductive either because the bottom-up directionality assumption fails, or because the mapping is not ontologically conservative. The antireductionist challenge, then, runs as follows. Reductive explanation requires mapping cognitive elements onto structures (working parts) within neural mechanisms. Further, cognitive elements are only real when (or to the extent that) they map onto these structures, as the componential realism claim requires. Failures of model-to-mechanism mapping along any of these lines would not only constitute a failure of reductive explanation, but would pose a daunting challenge to the cognitive ontology that psychological models are committed to. In the remainder of this discussion I will sketch a possible defense of realism from within this antireductionist framework.[9]

Recall that realism is a causal thesis: for a model element to exist requires that it have some sort of causal significance. One proposal for how to understand causal claims is in terms of *interventions* or *manipulations* (Campbell, 2006, 2008, 2010, Woodward, 2003, 2008). On an

[9] Another argument to this conclusion, though one that takes a different route, appears in Egan (2016).

interventionist conception, X causes Y just in case, relative to certain background conditions, if there were a single intervention on X (and only on X), then the value of Y would change. X and Y here are state variables that can take on several possible values, and interventions or manipulations consist in events of making these variables have one such value rather than another. The relationships between variables imply that counterfactuals hold systematically and contrastively between all pairs of values, such that if X were manipulated to have the value x_a, then Y would have the value y_a, and so on. In the circumstances where the system's state changes obey such counterfactuals, it is true to say that changes in the value of X cause changes in the value of Y.

Interventionism as an account of causal explanation is closely linked with experimental procedures that are designed specifically to vary certain conditions and determine what effect, if any, these changes will have on a creature's cognition and behavior. It is therefore tailor made for understanding the cognitive models that psychology generates, since these are developed in response to precisely such interventions (Rescorla, 2016). Within Baddeley's model of working memory, for example, the phonological loop is assumed to be insensitive to semantic relations, so manipulations of these properties should not affect maintenance of information, whereas phonological similarity should increase the confusability of items in memory. The fact that these representational manipulations produce the predicted outcomes is evidence in favor of a component system with the hypothesized properties (Baddeley, 2012).

Similarly, systems analyses of cognitive capacities (boxologies) can generally be interpreted in interventionist terms, since they are visual representations of abstract structures characterized in terms of clusters of intervention points. In models of attention, for instance, there is a distinction between passive and dynamically tunable filters: the former only allow

certain “preset” representations into working memory, while the latter can be adjusted by the current contents of working memory. In practice this comes down to whether intervening on the content of working memory will change the types of information that are allowed in on future trials, particularly whether information on task-irrelevant dimensions will make a difference (Pratt & Hommel, 2003). These diagrams, in short, can be regarded as guides to thinking about patterns of interventions rather than as proto-hypotheses about physical or mechanistic structures.

Interventionism, then, suggests an alternative conception of realism according to which in order for a model element to be causally real, it needs to correspond to system state variables that can be manipulated to alter outcomes in specific ways. This ties model elements fairly directly to experimental procedures in psychology. However, these variables do not need to correspond to structural components of the system’s physical realization base, or more specifically to parts of mechanisms. They may instead be entities (akin to collective variables) that can’t be mapped onto anything that would constitute a neuroanatomically or physiologically recognizable constituent of the brain. This opens up the desired conceptual space by showing how causal realism about psychological kinds can co-exist with their holistic, entangled realization.

7. Objections and replies

The first objection to the antireductionist picture sketched here targets its permissiveness vis-à-vis holistic realization. This objection has been pressed by Peter Godfrey-Smith (2008), who notes that contemporary functionalism is committed to functionally characterized components that are “level-bound”: they are both causally real and capable of supporting explanations of the system’s behavior, but also only visible from the ontological perspective of a

particular level (p. 66). Against this possibility, he argues that entities at the cognitive level must ultimately be discharged in terms of "*bona fide* parts, or states of *bona fide* parts" (p. 68), which alone can underwrite a "literally correct causal description" (p. 70) of the system.

More formally, citing Chalmers' (1996) account of computational implementation, Godfrey-Smith posits a "requirement that each CSA [combinatorial state automaton] substate be mapped onto a *distinct spatial region* of the implementing system… a theory of implementation must exclude a mapping in which each CSA substate is mapped holistically to a partial specification of the physical state of the entire system" (p. 68). When doing computational psychology, we must, in short, move from merely conjectured entities in models to more seriously grounded components of mechanisms.

As we have seen, however, the interventionist conception differs with the componential realist on the conditions for saying that a model element is causally real. It is sufficient that it be the locus of a cluster of interventionist counterfactuals, rather than having any direct mapping onto mechanistic parts. Realization requires that the system be organized in some such way that these counterfactuals are true of it, not that it be organized in a specifically mechanistic way.

A second objection centers on the role of dissociation studies in (dis)confirming cognitive models. Neural interventions can sometimes dissociate cognitive functions that are modeled as being part of the same system, and this is often regarded as evidence against cognitive theories that lump them together. If realizers are routinely widely entangled, it will often be possible to influence several functions at once by selective neural interventions on their common basis regions, and this might suggest that our current cognitive ontologies are mistaken in regarding these functions as separate.

However, cognitive models themselves do not imply any counterfactuals about what would happen if various types of neural interventions were performed. The illusion that they do stems from reading them as componential realists tend to, namely as preliminary sketches or hypotheses about neural structure. From an interventionist standpoint, though, they merely describe relations among manipulable cognitive entities. The fact that a specific neural intervention may disrupt the pattern of counterfactuals that are true of these cognitive entities does not undermine the fact that they nevertheless hold in the routine circumstances when they are assessed against the background of the normal, intact neural system. Putting the point differently, the fact that manipulating X normally results in changes to Y doesn't say anything about whether there is some third thing such that manipulating it might affect them both.

A related objection is that the case for entanglement has been overstated. It might turn out that all psychological constructs *can* be discriminated neurally from one another, even though some can only be discriminated *weakly* (Lenartowicz, Kalar, Congdon, & Poldrack, 2010). (Weak discrimination here means that the brain regions of interest involved in each tasks tapping each construct overlap largely but not entirely.) Therefore, there are no true cases of constructs that are indiscriminably entangled, and no distinct-realizer violations. However, this overlooks an important point: if such small differences are allowed to count against entanglement then they cannot *also* be used as evidence to motivate elimination of constructs from our cognitive ontology. The two claims stand and fall together. A construct might be only weakly neurally discriminable but still be robustly detectable and manipulable using psychological methods.

Finally, the interventionist's criterion of realism may be thought to be overly liberal. Without pinning down cognitive elements to well-behaved mechanistic components, we may be unable to distinguish between models that capture real causal structure and those that are merely

phenomenological, hence not explanatory. It will, in short, be too easy to declare that a psychological model is realized—the view does not make enough discriminations to allow us to separate real from fictional constructs.

This criticism might be fair if interventionism failed to make any principled distinction between phenomenological and explanatory computational models. But it clearly does this. An example of a widely applied and highly predictive model is Latent Semantic Analysis (Landauer & Dumais, 1997). LSA is, in effect, a data-mining technique for analyzing a large corpus of text and generating a high-dimensional representation of the associations among the words and phrases contained therein. These representations can be used to simulate performance on vocabulary tests and the Test of English as a Foreign Language (used as a measure of English proficiency for non-native speakers), and have also been harnessed for automated grading of student essays. However, there are few studies showing that they can be intervened on and manipulated in a way that systematically affects human lexical processing. Since the association matrices that LSA outputs are not plausible targets of intervention, the model's predictive facility is no evidence of their psychological reality—a conclusion that would hold even if LSA were a far more "neurally realistic" model.

8. Conclusion

As the limited success of reductionist interfield strategies indicates, a theory of how computational and other representational elements in psychological models are implemented remains in many ways elusive. The possibility I have sketched and defended here is one on which psychological kinds are neurally realized, but the realization relation might be a hopelessly entangled one. At least some of the empirical evidence currently leans in this

direction. If this situation turns out to be intractable and persistent the mind-body relationship might turn out to be considerably more epistemically opaque than reductionist heuristics have traditionally assumed it to be. In forging a path forward, integrative modeling in the mind/brain sciences may at last be shedding the habits of thought and practice that characterized its reductionist past.

References

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.

Anderson, M. L. (2010). Neural reuse: a fundamental organizational principle of the brain. *The Behavioral and Brain Sciences, 33*(4), 245-66-313.

Anderson, M. L. (2014). *After Phrenology*. Cambridge, MA: MIT Press.

Anderson, M. L. (2015). Précis of *After Phrenology*: Neural Reuse and the Interactive Brain. *Behavioral and Brain Sciences*, 1–22.

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

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

Bechtel, W. (2009). Looking down, around, and up: Mechanistic explanation in psychology. *Philosophical Psychology, 22*(5), 543–564.

Bechtel, W. (2012). Referring to localized cognitive operations in parts of dynamically active brains. In A. Raftopoulos & P. Machamer (Eds.), *Perception, Realism, and the Problem of*

Reference (pp. 262–284). Cambridge: Cambridge University Press.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: a mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(2), 421–41.

Bechtel, W., & Hamilton, A. (2007). Reduction, integration, and the unity of science: Natural, behavioral, and social sciences and the humanities. In T. Kuipers (Ed.), *General Philosophy of Science: Focal Issues* (Vol. 1, pp. 377–430). Amsterdam: North Holland.

Bechtel, W., & Richardson, R. C. (2010). *Discovering Complexity*. Cambridge: MIT Press.

Bickle, J. (2006). Reducing mind to molecular pathways: explicating the reductionism implicit in current cellular and molecular neuroscience. *Synthese*, *151*(3), 411–434.

Bressler, S. L., & Kelso, S. (2016). Coordination dynamics in cognitive neuroscience. *Frontiers in Systems Neuroscience*, *10*(September), 1–7.

Buckner, C. (2014). Functional kinds: a skeptical look. *Synthese, 192*, 3915–3942.

Burianova, H., & Grady, C. L. (2007). Common and unique neural activations in autobiographical, episodic, and semantic retrieval. *Journal of Cognitive Neuroscience*, *19*(9), 1520–34.

Burianova, H., McIntosh, A. R., & Grady, C. L. (2010). A common functional brain network for autobiographical, episodic, and semantic memory retrieval. *NeuroImage*, *49*(1), 865–874.

Butler, K. (1994). Neural constraints in cognitive science. *Minds and Machines*, *4*(2), 129–162.

Campbell, J. (2006). An interventionist approach to causation in psychology. In A. Gopnik & L. E. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 58–66). Oxford, UK: Oxford University Press.

Campbell, J. (2008). Interventionism, control variables and causation in the qualitative world. *Philosophical Issues*, *18*(1), 426–445.

Campbell, J. (2010). Control Variables and Mental Causation. *Proceedings of the Aristotelian Society*, *110*, 15–30.

Chalmers, D. J. (1994). On implementing a computation. *Minds and Machines*, *4*(4), 391–402.

Chalmers, D. J. (1996). Does a Rock Implement Every Finite-State Automaton? *Synthese*, *108*(3), 309–333.

Chalmers, D. J. (2011). A Computational Foundation for the Study of Cognition. *Journal of Cognitive Science*, *12*, 323–357.

Coltheart, M. (2013). How Can Functional Neuroimaging Inform Cognitive Theories? *Perspectives on Psychological Science*, *8*(1), 98–103.

Copeland, B. J. (1996). What is Computation? *Synthese*, *108*(3), 335–359.

Colombo, M., Hartmann, S., & Van Iersel, R. (2015). Models, mechanisms, and coherence. *British Journal for the Philosophy of Science*, *66*(1), 181–212.

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

Craver, C. F., & Darden, L. (2013). *In Search of Mechanisms*. Chicago: University of Chicago Press.

Crossley, N. A., Mechelli, A., Vértes, P. E., Patel, A. X., Ginestet, C. E., McGuire, P., & Bullmore, E. T. (2013). Cognitive relevance of the community structure of the human brain functional coactivation network. *Proceedings of the National Academy of Sciences*, *110*(38), 15502–15502.

Darden, L. (2001). Discovering Mechanisms, 3–15.

Darden, L., & Maull, N. (1977). Interfield theories. *Philosophy of Science*, *44*(1), 43–64.

Dehaene, S. (2011). *The Number Sense*. Oxford: Oxford University Press.

Dehaene, S., & Cohen, L. (2007). Cultural recycling of cortical maps. *Neuron*, *56*(2), 384–398.

Egan, F. (2016). Function-Theoretic Explanation and the Search for Neural Mechanisms.

Fagan, M. B. (2012). The Joint Account of Mechanistic Explanation. *Philosophy of Science, 79*(4), 448–472.

Franklin-Hall, L. R. (2016). New Mechanistic Explanation and the Need for Explanatory Constraints. In *Scientific Composition and Metaphysical Ground* (pp. 1–29).

Gallistel, C. R., & King, A. P. (2010). *Memory and the Computational Brain*. Malden, MA: Wiley-Blackwell.

Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science, 69*(September), 342–354.

Glymour, C. (2007). When is a brain like the planet? *Philosophy of Science, 74*(3), 330–347.

Godfrey-Smith, P. (2008). Reduction in real life. In J. Hohwy & J. Kallestrup (Eds.), *Being Reduced* (pp. 52–74). Oxford: Oxford University Press.

Greenberg, D. L., & Verfaellie, M. (2010). Interdependence of episodic and semantic memory: evidence from neuropsychology. *Journal of the International Neuropsychological Society: JINS, 16*(5), 748–53.

Horst, S. (2007). *Beyond Reduction*. Oxford: Oxford University Press.

Kaiser, M. I. (2015). *Reductive Explanation in the Biological Sciences*. Dordrecht: Springer.

Kaplan, D. M., & Craver, C. F. (2011). The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective. *Philosophy of Science, 78*(Oct.), 601–627.

Kim, J. (2005). *Physicalism, or Something Near Enough*. Princeton, NJ: Princeton University Press.

Kim, J. (2008). Reduction and reductive explanation: Is one possible without the other? In J.

Hohwy & J. Kallestrup (Eds.), *Being Reduced* (pp. 93–114). Oxford: Oxford University Press.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Lenartowicz, A., Kalar, D. J., Congdon, E., & Poldrack, R. A. (2010). Towards an ontology of cognitive control. *Topics in Cognitive Science*, *2*(4), 678–692.

Love, B. C. (2016). Cognitive Models as Bridge between Brain and Behavior. *Trends in Cognitive Sciences*, *20*(4), 247–248.

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.

McCauley, R. N., & Bechtel, W. (2001). Explanatory pluralism and heuristic identity theory. *Theory & Psychology*.

Meehan, T. P., & Bressler, S. L. (2012). Neurocognitive networks: Findings, models, and theory. *Neuroscience and Biobehavioral Reviews*, *36*(10), 2232–2247.

Miłkowski, M. (2011). Beyond Formal Structure: A Mechanistic Perspective on Computation and Implementation. *Journal of Cognitive Science*, *12*(4), 359–379.

Nathan, M. J., & Del Pinal, G. (2015). Mapping the mind: bridge laws and the psycho-neural interface. *Synthese*.

Nicholson, D. J. (2012). The concept of mechanism in biology. *Studies in History and Philosophy of Science Part C*, *43*(1), 152–163.

Piccinini, G. (2015). *Physical Computation*. Oxford: Oxford University Press.

Piccinini, G., & Craver, C. F. (2011). Integrating psychology and neuroscience: functional

analyses as mechanism sketches. *Synthese*, *183*(3), 283–311.

Poeppel, D., & Embick, D. (2005). Defining the relation between linguistics and neuroscience. In A. Cutler (Ed.), *Twenty-First Century Psycholinguistics: Four Cornerstones* (pp. 103–120).

Poldrack, R. A. (2010). Mapping Mental Function to Brain Structure: How Can Cognitive Neuroimaging Succeed? *Perspectives on Psychological Science*, *5*(6), 753–761.

Pratt, J., & Hommel, B. (2003). Symbolic control of visual attention: The role of working memory and attentional control settings. *Journal of Experimental Psychology. Human Perception and Performance*, *29*(5), 835–845.

Rescorla, M. (2014). A theory of computational implementation. *Synthese*, *191*(6), 1277–1307.

Rescorla, M. (2016). An interventionist analysis of psychological explanation, 1–47.

Reydon, T. A. C. (2009). How to Fix Kind Membership: A Problem for HPC Theory and a Solution. *Philosophy of Science*, *76*(5), 724–736.

Sarkar, S. (1992). Models of reduction and categories of reductionism. *Synthese*, 167–194.

Schaffner, K. F. (1993). *Discovery and Explanation in Biology and Medicine*. *Discovery and Explanation in Biology and Medicine*. Chicago: University Of Chicago Press.

Shagrir, O. (2012). Computation, implementation, cognition. *Minds and Machines*, *22*(2), 137–148.

Shapiro, L. A. (2015). Mechanism or Bust? Explanation in Psychology. *The British Journal for the Philosophy of Science*, 1–31.

Sporns, O. (2011). *Networks of the Brain*. Cambridge, MA: MIT Press.

Sporns, O. (2014). Contributions and challenges for network models in cognitive neuroscience. *Nature Neuroscience*, *17*(5), 652–660.

Stinson, C. (2016). Mechanisms in Psychology: Ripping Nature at its Seams. *Synthese*, *193*(5),

1585–1614.

Teller, P. (2010). Mechanism, Reduction, and Emergence in Two Stories of the Human Epistemic Enterprise. *Erkenntnis, 73*(3), 413–425.

Theurer, K. L., & Bickle, J. (2013). What's Old Is New Again: Kemeny-Oppenheim Reduction at Work in Current Molecular Neuroscience, *17*(2), 89–113.

Theurer, K. L. (2013). Seventeenth-Century Mechanism: An Alternative Framework for Reductionism. *Philosophy of Science, 80*(5), 907–918.

Tulving, E. (1983). *Elements of Episodic Memory*. Oxford: Oxford University Press.

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

Weiskopf, D. A. (2011b). The functional unity of special science kinds. *British Journal for the Philosophy of Science, 62*(2), 233–258.

Weiskopf, D. A. (2016). Integrative modeling and the role of neural constraints. *Philosophy of Science*.

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

Wimsatt, W. C. (1974). Reductive explanation: A functional account. *Philosophy of Science*, 671–710.

Wimsatt, W. C. (1994). The ontology of complex systems: levels of organization, perspectives, and causal thickets. *Canadian Journal of Philosophy, 20*, 207–274.

Wimsatt, W. C. (2006). Reductionism and its heuristics: Making methodological reductionism honest. *Synthese, 151*(3), 445–475.

Wimsatt, W. C. (2007). Aggregate, composed, and evolved systems: Reductionistic heuristics as

means to more holistic theories. *Biology & Philosophy*, *21*(5), 667–702.

Woodward, J. (2003). *Making Things Happen*. Oxford: Oxford University Press.

Woodward, J. (2008). Mental causation and neural mechanisms. In *Being Reduced* (pp. 218–262).

Woodward, J. (2013). Mechanistic Explanation: Its Scope and Limits. *Aristotelian Society Supplementary Volume*, *87*(1), 39–65.

Zednik, C. (2015). Heuristics, descriptions, and the scope of mechanistic explanation. In P.-A. Braillard & C. Malaterre (Eds.), *Explanation in Biology* (pp. 295–318). Dordrecht: Springer.