Title:
Functional Kinds—A Skeptical Look

Cameron Buckner

Assistant Professor of Philosophy
University of Houston
513 Agnes Arnold Hall
Houston, TX 77204-3004
Phone: 713-743-3010
Fax: 713-743-5162

cjbuckner@uh.edu

Abstract: The functionalist approach to kinds has suffered recently due to its association with law-based approaches to induction and explanation. Philosophers of science increasingly view nomological approaches as inappropriate for the special sciences like psychology and biology, which has led to a surge of interest in approaches to natural kinds that are more obviously compatible with mechanistic and model-based methods, especially homeostatic property cluster theory. But can the functionalist approach to kinds be weaned off its dependency on laws? Dan Weiskopf has recently offered a reboot of the functionalist program by replacing its nomological commitments with a model-based approach more closely derived from practice in psychology. Roughly, Weiskopf holds that the natural kinds of psychology will be the functional properties that feature in many empirically successful cognitive models, and that those properties need not be localized to parts of an underlying mechanism.

I here skeptically examine the three modeling practices that Weiskopf thinks introduce such non-localizable properties: fictionalization, reification, and functional abstraction. In each case, I argue that recognizing functional properties introduced by these practices as autonomous kinds comes at clear cost to those explanations' counterfactual explanatory power. At each step, a tempting functionalist response is parochialism: to hold that the false or omitted counterfactuals fall outside the modeler's explanatory aims, and so should not be counted against functional kinds. I conclude by noting the dangers this attitude poses to scientific disagreement, inviting functionalists to better articulate how the individuation conditions for functional kinds might outstrip the perspective of a single modeler.

I. Introduction

The very label 'functional kind' can seem an oxymoron, implying a curious mix of similarity and dissimilarity. On the one hand, to form a *kind*, a category's members must share a set of common properties, traits, or structures. On the other hand, for a kind to be purely *functional*, these characteristic similarities must consist only in the activities, processes, or roles that category members perform, and those functional profiles must be implementable by significantly dissimilar mechanisms. It may seem a cosmic coincidence that a reliable ability to perform the same function could arise in different systems without that ability being grounded in similar underlying structures. Nevertheless, that the world in which we live is arranged in just this way is, functionalists claim, one of the most important empirical discoveries of the "special sciences" like psychology, biology, and economics.

The classic arguments in favor of functional kinds presuppose a nomological approach to induction and explanation. Briefly, the claim was that the special sciences have discovered many laws holding between functionally-defined categories, so taking these sciences seriously requires us to acknowledge their purely functional kinds. Grounded as it is in an appeal to scientific fidelity, the classic functionalist program has been embarrassed by a growing consensus in philosophy of science that this nomological interpretation of the special sciences is untenable. The special sciences are not in the business of discovering and confirming laws, this emerging consensus holds, but rather models or mechanisms. To address this weakness in the program, Dan Weiskopf (2011a; 2011b; forthcoming) has recently offered an impressive reboot by reformulating functionalism on a more adequate model-based approach to induction and explanation.

Here, I explore some of the challenges facing this "new functionalism" about kinds. The general worry is that models provide a less stable foundation for scientific taxonomy than laws. Laws, if true, are general and eternal, whereas models provide a more partial, provisional, inconsistent, and idiosyncratic purchase on the systems they describe. By contrast, the label 'natural kind' is an honorific reserved for the lynchpins holding together progressive stages of scientific investigation. Epistemic factors can be recognized in the individuation of kinds, but kinds should also float somewhat free from them, and their study—from a variety of epistemic perspectives—should continually reveal more about the world's explanatory structure as their nature is iteratively elaborated by progressive research programs. The more evidence we have that a category will be conserved and regarded as important in future explanations, the stronger should be our conviction that it is a natural kind; and conversely, evidence that it will not be conserved or does not explain as well as alternatives should weaken our belief that it is a natural kind. A key question, then, is whether we have good evidence that the functional properties featuring in successful cognitive models support the relevant sort of explanatory stability.

In this paper, I review reasons to worry that they do not. Whereas Weiskopf (2011b) focuses his defense of functional kinds against reductionists like Polger and Kim, I will here weigh it against mechanistic approaches to kinds as outlined by figures like Boyd (1991, 1999), Machery (2004), and Griffiths (1997). Functionalists and mechanists here share many methodological commonalities; they both think that kinds are typically discovered by breaking the activity of a complex system down into interactions amongst simpler activities. A key contrast, however, concerns the question of an additional causal criterion for kindhood: mechanists require that kind members nonaccidentally exhibit their shared functional profiles due to the operation of a shared underlying mechanism, whereas functionalists deny that members of a kind need to share any underlying structure.[1] Indeed, Weiskopf holds that many functional kinds cannot be located in parts of an underlying mechanism at all, so there would not even be a place where such an influence could be expressed. Thus, if successful explanations in the special sciences routinely featured such nonlocalizable kinds, this would provide strong evidence against the mechanistic approach to kinds.

After providing some conceptual background (Section II), I skeptically examine this evidence by reviewing the three types of modeling activity that Weiskopf thinks introduces nonlocalizable functional categories: fictionalization, reification, and functional abstraction (Section III). Against these three types of modeling activity as a source of natural kinds, I offer two different styles of argument (Section IV). Concerning fictionalization and reification, I concede that these models are distinct from mechanistic models, but argue that interpreting fictions or reifications as natural kinds comes with clear costs in terms of those models' counterfactual power—counterfactual power being an ecumenical "currency" that both functionalists and mechanists value and by which competing explanations can be ranked. Functional abstraction, on the other hand, can be considered a legitimate source of kinds, but only on the condition

[1] In principle, mechanism vs. functionalism about explanation and mechanism vs. functionalism about kinds are independent questions; one could be committed on one dispute and agnostic on the other. But mechanism about kinds fits most naturally with mechanism about explanation, because of their common emphasis on localization in an underlying mechanism.

that the functionally abstract models be interpreted as "templates" that could be elaborated into a more complete mechanistic model.[2]

The conclusion of this latter argument is consistent with the assessments of Craver & Piccinini (2011) and Kaplan & Craver (2011) that functional models are mechanism sketches, but I arrive at this conclusion from within Weiskopf's defense of functional abstraction. I conclude (Section V) by suggesting an alternative emphasis for future work on functional kinds: the need to elaborate their individuation conditions beyond mere functional profiles so that their natures can productively be the subject of sustained empirical disagreement. The failure of these three etiologies for nonlocalizable kinds does not in itself vindicate mechanism about kinds; but it removes a key piece of evidence against it, and in each case the deficiencies of functionalism accentuate the strength of the mechanistic alternative.

II. Historical background: Fodorian Foundations

Natural kinds, compared to other classes, are scientifically precious. Consider some non-natural categories: the set of objects that are grue, non-ravens, or less than two kilometers from the Eiffel Tower. While the definitions of these classes are clear enough to tell us what is in or out of their extensions, they are not suitable categories for scientific research, for they carve up the world in relatively arbitrary ways. Kinds, by contrast, are useful for sciences to track, for their boundaries correspond to important structural divisions in the world. A theory of kinds can be distinguished by its accounts of (1) the nature of these structural breaks and (2) how we come to know whether a proposed class division corresponds to one.

For the latter half of the twentieth century, the notion of a natural kind was closely tied to the notion of a natural law. The connection arose from attempts to address the problems of induction, filtered through Goodman's analysis of projectible predicates. In this tradition, kinds are useful to science due to their inductive potential; if predicates *A* and *B* pick out kinds, and we observe some sample of *A*s causing *B*s, then we are justified in taking these observations to confirm a law of the form $A \rightarrow B$. Quine (1969)

[2] As we will see, we need to distinguish two different notions of "complete" here. In the first sense, an explanation is complete if it is maximally detailed (it omits no relevant specifics). In the second sense, an explanation is complete if it is unlikely to be revised in future iterations of a progressing research program. Both senses may be relevant to kindhood, and both will be considered below.

influentially suggested that natural kinds confirm inductions in this way because members of natural kinds are united by a deep underlying similarity—i.e., if *A* is a natural kind, then the observed *A*s are likely to resemble the unobserved *A*s in scientifically-relevant respects. While defining laws in terms of kinds, and kinds in terms of similarity, Quine worried that he could not reduce similarity to any "less dubious notion" like logic or set theory. He thus left us with a tangle of important and interrelated ideas, but no clearer independent purchase on any of them.

Natural kind theorists after Quine recommend different strategies to untie this knot. Microessentialists appeal to other, "lower-level" sciences, on the assumption that members of natural kinds possess shared essences that serve as necessary and sufficient conditions for kind membership. While microessentialism might have seemed plausible for some paradigm examples like "water = H_2O"[3], it is not a viable approach in the special sciences, since consensus holds that special science kinds lack necessary and sufficient conditions. Homeostatic property cluster (HPC) theorists instead hope to exchange Quine's primitive appeals to similarity for a more precise notion of homeostatic property clustering secured by shared underlying mechanisms. According to HPC theory (most associated with the work of Boyd—1991, 1999), a kind is the maximal class whose members non-accidentally share a large set of scientifically interesting properties due to the operation of at least one shared causal mechanism—but crucially, no property or even subset of properties need be regarded as necessary for kind membership.

Functionalists about kindhood reject an assumption shared by both of these approaches: that an underlying structure or mechanism need be common amongst all members of a kind. They worry that this assumption flies in the face of an obvious truth about kinds in the special sciences: that they can be realized or implemented by significantly different kinds of underlying mechanism. Consider Fodor's classic appeal to Gresham's Law, or the economic principle that "good money drives out bad" (1974, 124). 'Money' is here taken to name a natural kind; but it is, he thinks, wholly implausible that all the

[3] However, without significant complications, microessentialism is probably not even plausible for 'water'; see Van Brakel (2000) and Needham (2011).

various forms of money—dollars, personal checks, wampum, and so on—share any lower-level description.[4] Given the apparent implausibility of lower-level criteria for special science kinds, functionalists instead suppose that the special sciences enjoy taxonomic “autonomy”—that the kinds of a special science do not depend upon the kinds of any other science for their legitimacy, utility, or reality. Multiple realizability and autonomy are today much more controversial than they once were; but since mountains of literature are now devoted to these topics (see Bickle 2010 for a review), I propose to set aside these debates at present and focus more directly on the implications these doctrines hold for the functionalist’s criterion for kindhood, which has received less sustained attention.

The most influential functionalist criterion is also due to Fodor, who suggested that we reverse Quine’s direction of dependency by taking special science laws as primitive and defining special science kinds as those categories that feature in many well-confirmed special science laws. On this view, it is a brute fact about nature—a fact empirically discovered in the special sciences—that types of state or event can feature in macro-level regularities without those regularities being reducible to or secured by any shared micro-level regularities. We may have the metaphysical intuition, Fodor concedes, that there must be some deeper explanation as to why some classes feature in many special science laws and others do not. Fodor’s diagnosis is that this intuition merely reflects the hope that “ontological transparency” will win out over “empirical generality” (1997, 161)—but God could have made the world any way he liked, and the special sciences have revealed that we happen to live in a world where many real, empirical (non-analytic) nomic relations hold amongst multiply realized kinds.

Thankfully, the debate over this interpretation of evidence in the special sciences need not concern us here, for the new functionalists like Weiskopf reject the nomological criterion of kindhood on which it was based. Before moving on, however, it is important to note the elegance of Fodor’s position, which comprises a set of interrelated and mutually-supporting theses:

[4] Viewed by today’s lights, there are obvious problems with this classic example: it narrowly targets the Nagelian bridge-law program, whereas more permissive accounts of interlevel kind criteria (e.g. the HPC view) are more promising; and Fodor requires that all examples of money share intrinsic physical similarities, whereas what it is to be money may depend on extrinsic psychological or institutional relations, which are permitted by some mechanistic accounts of natural kinds (e.g. the HPC view).

1. The autonomy of the special sciences
2. The multiple realizability of special science kinds
3. A nomological approach to induction and explanation
4. A nomological criterion for kindhood

However, other than a few dynamicists (e.g. Walmsley 2008)—unlikely allies for Fodorians—philosophers of the special sciences now consider the nomological approach highlighted in planks 3 and 4 to be dead in the water. As a result, these aspects of the functionalist program must be abandoned. However, jettisoning the nomological criterion of kindhood puts special pressure on the functionalist approach to kinds, for the most popular alternative criteria are variants of the mechanistic approach (e.g., HPC theory). Functionalists cannot simply adopt the mechanist's criterion, for the requirement that kinds share some underlying mechanistic structure appears to be at odds with the remaining planks of the functionalist program, autonomy and multiple realizability. It is thus no easy task to rebuild the functionalist theory of kinds while retaining the core ideas which originally inspired it.

III. Weiskopf's New Functionalism

A concerted campaign to reform the functionalist approach to kinds has been mounted recently by Dan Weiskopf. In a series of papers, Weiskopf (2011a, 2011b, forthcoming) has sketched a new functionalist approach that replaces Fodor's nomological commitments with a model-based foundation more obviously relevant to psychology's actual methodology. Weiskopf's reboot begins with the observation that psychology is characteristically in the business of devising and testing models, rather than laws. Crucially, Weiskopf denies the claims of Craver (2007), Piccinini & Craver (2011), and Kaplan & Craver (2011) that psychological models should always be construed as sketches of neural mechanisms. The cognitive models that Weiskopf takes to be stock-and-trade of psychological explanation are, he concedes, like models of mechanisms in many ways; most notably, they decompose complex capacities into a series of subcapacities that, when interacting together in the right way, can produce the explanandum. However, Weiskopf denies that the functional categories featuring in many successful cognitive models can be localized to parts of an underlying mechanism. Since localization is an essential feature of mechanistic explanation, he concludes that these models cannot be regarded as mechanism sketches.

The models that interest Weiskopf are those that explain psychological capacities in terms of other psychological capacities. In particular, they explain representational capacities in terms of interactions amongst other, usually more basic representational capacities.[5] As Weiskopf puts it, such models specify “the set of representations (primitive and complex) that the system can employ, the relevant stock of operations…, the relevant resources available and how they interact with other operations, [and] how they are organized to take the system from its inputs to its outputs” (2011a, 323). Weiskopf focuses on three examples of such cognitive models: Hummel & Biederman’s (1992) geon-based model of object recognition, Kruschke’s (1992) exemplar-based model of categorization, and Love & Gureckis’ (2004; 2007) cluster-based models of categorization.

For present purposes, the most interesting aspect of Weiskopf’s platform is the space it opens up for a novel functionalist criterion for kindhood. Specifically, just as Fodor holds functional kinds to be classes that stand in many well-confirmed special science laws, Weiskopf holds functional kinds to be “abstractly defined functional categories [that] earn their credentials by participating in a range of models that are themselves empirically validated” (2011b, 251). On this approach, a class *K* becomes a psychological kind not by featuring in many psychological laws, but rather by featuring in many successful cognitive models that apply to many distinct types of cognizer. Weiskopf offers several examples of functional properties that would be considered kinds on his view, including central pattern generators, memory buffers, and analog accumulators. These classes count as natural kinds on Weiskopf’s view because they turn up so often in successful cognitive models, despite the fact that there does not appear to be any common neural mechanism that realizes them in each system that these models correctly describe. (Mechanists about kinds, by contrast, would require the additional constraint that these realizers

[5] In this paper, when I write of “cognitive models” I am using the label as a technical term as defined by Weiskopf; I here take no stand on whether non-representational models should be regarded as ‘cognitive’ in any other sense.

nonaccidentally implement those functional profiles due to the operation of some shared causal mechanism.)

Notably, Weiskopf's reboot—like the nomological approach he aims to supplant—honors both the doctrines of multiple realizability and taxonomic autonomy. Weiskopf's functional kinds are multiply realizable, because the models in which they feature may correctly apply to many different kinds of underlying mechanism. Central pattern generators (CPGs) are his best-worked out example on this point; according to Weiskopf, CPGs can be functionally defined as "units that produce regular oscillations endogenously or in response to input"; but they can be assembled out of such diverse mechanisms as "multi-neuron arrays of varying sizes using inhibitory interneurons, or out of local dendrodentritic connections…[which] differ in their size, location, temporal characteristics, and many other physical/neural properties" (2011b, 247-248). Moreover, on Weiskopf's view classes such as CPGs are functionally-defined and depend for their natural kind status only on the success of the models in which they appear, and so also enjoy taxonomic autonomy from other sciences. The models in which they feature are pitched entirely at the psychological level of description, and their empirical utility can (at least in principle) be confirmed using psychological methods of investigation alone.[6]

Weiskopf thus offers a package of views on kinds and explanation in the special sciences that parallels that of Fodor, but is substantially improved by being more obviously compatible with paradigm examples of inductive and explanatory practice in (at least) psychology. This package includes:

1. The autonomy of special sciences
2. The multiple realizability of special science kinds

3*. A model-based approach to induction and explanation
4*. A model-based criterion for kindhood

A corollary of the package's novel criterion for kindhood is that it will produce a novel list of kinds when applied to putative explanations in the special sciences. A substantial benefit of Weiskopf's approach is that the literature's tired examples of robot and Martian pain can now be set aside, replaced by a set of more relevant categories like CPGs, lateral inhibition, and memory buffers. However, we should subject

[6] Weiskopf concedes that psychology may make use of neural evidence (i.e. he does not endorse a strict reading of *evidential* autonomy—see Weiskopf, forthcoming), but only as a guide to or proxy for psychological findings.

this novel list of kinds to careful scrutiny to determine whether these new functional kinds satisfy traditional constraints on kindhood.

Perhaps the most important of those traditional constraints is that kinds must be explanatory; Weiskopf thus needs to rebut Craver's, Kaplan, and Piccinini's arguments that all functional models are either mechanism sketches or fail to explain. The argument appears in a variety of forms, but for present purposes Weiskopf's summary (2011a, 319) will do:

1. Functional explanations are nonlocalized.[7]
2. Nonlocalized explanations provide only a redescription of the phenomenon or a how-possibly model.
3. Redescriptions and how-possibly models are not explanatory.
4. So [functional] explanations are not explanatory.

In his defense of functional explanation (2011a)—and so, derivatively, his defense of functional kinds (2011b)—Weiskopf challenges the second premise of the argument. He aims to establish that functional kinds can contribute to genuine explanations that do not reduce to either redescriptions of phenomena or how-possibly (i.e. incomplete mechanist) models.

To this end, Weiskopf describes three types of modeling activity that can provide genuinely explanatory but not fully localizable model components: fictionalization, reification, and functional abstraction. First, fictionalization involves "putting components into a model that are known not to correspond to any element of the modeled system, but which serve an essential role in getting the model to operate correctly" (2011a, 331). Second, reification is the "act of positing something with the characteristics of a more or less stable and enduring object, where in fact no such thing exists" (2011a, 328). Third, functional abstraction occurs "when we decompose a modeled system into subsystems and other components on the basis of what they do, rather than their correspondence with organizations and groupings in the target system" (2011a, 329). According to Weiskopf, models containing components introduced by fictionalization, reification, and functional abstraction can satisfy norms of good

[7] Craver and Weiskopf actually use the word 'noncomponential' here, but since Weiskopf later writes about the functional components of models that are noncomponential in this sense (a practice I follow here), I have used the word 'nonlocalized' to avoid confusion.

explanation, and so functional properties picked out by those components should be considered genuinely explanatory kinds even when they do not map onto any parts of a shared underlying mechanism.

The norms to which Weiskopf alludes are derived from Craver's analysis of mechanistic explanations (albeit with some reinterpretation and critique—Weiskopf 2011a, 315-318). First, models should be *well-confirmed*; ceteris paribus, a model that is better supported by the available evidence is to be preferred. Second, models should be *representationally accurate*; ceteris paribus, a model that includes only elements that are real parts of a system, and that describes those parts more specifically—so far as the parts and the level of specificity are relevant to our explanatory purposes—is to be preferred. Third, models should be *genuinely explanatory* (as opposed to merely phenomenologically accurate); ceteris paribus, the model that can answer more counterfactual, what-if-things-had-been-different questions (which may include—but don't on Weiskopf's view, *necessarily* include—questions about the effects of interventions) is to be preferred. Lastly, Weiskopf suggests that models should be *well-integrated*; ceteris paribus, the model that coheres better with general background knowledge is to be preferred.

So interpreted, Weiskopf is correct that functional models are evaluable along all these dimensions; but mere evaluability is not sufficient to establish the taxonomic stability of the relevant functionally-defined properties. Mere evaluability along these dimensions is consistent with functional models always ranking lower than nearby localized alternatives that satisfy them better and thus are to be preferred. Indeed, Craver's "mechanism sketches" arguments (e.g. Craver 2007, p130-131) could be reconstrued to conclude not that functionalists cannot make any of the relevant distinctions between explanations and pseudoexplanations, but rather that for every functionalist decomposition, there will be a more localized alternative that satisfies these norms better. Should every functional decomposition be inferior to a nearby mechanistic elaboration of that decomposition, then the functional properties they depict would lack the stability characteristic of kinds.

IV. To evaluate this possibility, I propose an argument by cases (see Figure 1 below for reference). To wit, the next section focuses on Weiskopf's three types of modeling activity (fictionalization, reification, and functional abstraction) in turn, arguing in each case that there are clear disadvantages to acknowledging functionally-defined kinds introduced by these methods. Should we find each of these three etiologies for non-localizable kinds to be problematic, it will be much less clear that Weiskopf's model-based functionalism offers a viable alternative to mechanistic approaches to kindhood.**Three non-mechanistic origins: fictionalization, reification, and functional abstraction**

Fictionalization

Let us begin with fictionalization. As noted above, fictionalization involves placing a component in a model that is known not to correspond to any element of the system modeled. According to Weiskopf, such fictions can be considered an important part of a successful model, rather than something "clearly intended to be eliminated by [a] better construct in later iterations" (2011a, 331). Fictional components are further to be distinguished from "black box" or "filler" items, in that in addition to providing a true functional description of the system (i.e. an input-output profile that the system in some sense actually performs), fictionalized models posit at least some causal powers or dispositions that the system does not actually possess.

Since natural kinds are traditionally taken to be the real, mind-independent joints of nature, fictional properties may seem non-starters here. However, given the prevalence of fictional components in successful models—and recent arguments that fictions can genuinely explain (e.g. Bokulich 2011)—if they are to be rejected as kinds, it should be based on some principled reason, rather than a blanket bias against unreal entities. On the topic of fictionalization, Weiskopf primarily discusses the Fast Enabling Links (FELs) in Hummel & Biederman's geon-based model (which involve an impossible, infinitely fast transfer of information in the geon-based model of categorization). Weiskopf notes that FELs play an essential role in getting that model to function by synchronizing distant neural regions, which crucially enables the binding of different intermediary representations in object categorization. Since FELs have not been deployed by any other modelers (and so might not count as a natural kind on Weiskopf's own criteria), I will focus on another probably fictional component: backpropagation learning in connectionist networks. This is a better test case for fictionalization than FELs, for the critical discussion surrounding backpropagation is mature, and few other components have featured in as many successful cognitive models.

The (re)discovery of backpropagation learning is largely responsible for the resurgence of connectionist modeling in the 1980s. There are many different varieties and architectures for neural networks, but I will here focus on its use in basic three-layer, feed-forward connectionist networks. Backpropagation is a supervised learning method with two crucial features: sigmoidal (rather than binary) activation functions for nodes and adjustments of link weights computed by iterative, backwards transmission of an error signal. Error signals are computed by comparing a network's actual to desired output in each learning trial and then propagating that error signal backwards from the output to earlier layers of the network across their connections. In other words, in each trial of a connectionist network's training phase, an input vector is presented to the network and activation propagates forward through the network's layers according to each node's sigmoidal activation function and the weights of the links between nodes. Once activation reaches the output layer, the difference between the actual and desired output is computed, generating that node's error value for that trial. For each preceding layer, this error signal is then distributed backwards across the node's input links, and the threshold values of all nodes and weights of all links are then updated (modulo the desired learning rate) to bring the future output closer to the desired output for that item in the training set.

Despite being by far the most common training rule for connectionist networks (crucially featuring in thousands of successful models), backpropagation has been heavily criticized for its purported biologically implausibility. Worries center on the backwards transmission of information across neural synapses, the need for prior knowledge of correct output, and the distinct, individualized error signals used to adjust the thresholds and weights of each node and link in the network. Though connectionist networks are generally thought to be more biologically plausible models of cognition than discrete symbol-based models—due to a closer correspondence of nodes and links to neurons (or neural assemblies) and synapses (or synapse chains)—these features are thought by many to be inconsistent with the operation of real neurons and synapses in the brain.[8] Nevertheless, because it is mathematically well-understood, supported by many software packages, and because no clear favorite has yet emerged from alternative learning rules, backpropagation continues to be commonly used even today.

A potentially surprising feature of current scientific practice, however, is that even modelers who frequently make use of backpropagation attach ontological disclaimers. As backpropagation-using connectionists Gluck & Myers put it, "backpropagation's success as an engineering tool does not necessarily imply anything about its validity as a psychological model of learning" (2001, 109). As further evidence that this unease goes deeper than rhetoric, there is a cottage industry in devising more biologically plausible training rules, including: simulated annealing, neural gas architectures, generalized recirculation, radial basis functions, deep learning, simulations of reinforcement learning, evolutionary methods, particle swarm optimization, and many others (see Haykin 2009 for a review of some of these techniques). This sustained interest in replacements is difficult to explain if backpropagation models are providing fully legitimate, autonomous explanations of learning.

There are good reasons for this caution, for the frequent appearance of useful fictions in models can tell us less about the structure of nature than it does about the representational power of that fictional component. Given enough nodes and a large enough training set, connectionist networks with sigmoidal activation functions are Turing-complete, and so could, in the absence of further constraints, model any computable function. This representational flexibility explains in part how backpropagation is able to feature in so many different successful models—because it is so powerful as to allow modelers to fit nearly any arbitrary data.[9]

[8] To qualify this critical consensus, there are a few interesting arguments in defense of the biological plausibility of backpropagation; some have suggested that backpropagation may be plausible if nodes are regarded not as individual neurons but rather as neural assemblies with recurrent connections (Stork, 1989), and others have concluded on the basis of neuroanatomical studies that something like an error signal—synaptic depression—might be transmitted backwards along individual synapses (though perhaps a time scales inconsistent with backpropagation–Fitsimmonds, Song, & Poo, 1997).

[9] Of course, no actual experiment can be conducted involving an infinite number of nodes and an infinite training set, and the actual neural network implementations of Turing machines have been built by hand (e.g. Siegelmann & Sontag 1991). It is a separate question what neural networks trained with a set number of nodes, a particular learning rule, a plausible learning set, and a fixed learning period can learn *easily*. The point stands, however, that

This worry extends beyond fictional components; other representationally flexible functional categories such as CPGs are subject to similar worries. To consider another relevant example, Prinz, Bucher, & Marder (2004) found that by varying free parameters such as the number, types, and strengths of neural connections in a three-node CPG network, there were over 450,000 possible configurations that produced output patterns consistent with lobster pyloric oscillations. While Weiskopf cites this study as evidence that CPGs are massively multiply realized, Selverston (1980) instead catalogues a number of cases where this enormous space of possible CPG models led researchers down empirical dead ends, given that modelers "usually have enough variable parameters…to produce any rhythm [they] desire" and choices of parameter values not justified by information about actual underlying mechanisms frequently appeared to provide "'confirmatory' evidence for the operation of certain circuits which were subsequently found to be fundamentally incorrect" (1980, 540). In short, representational flexibility in the form of free parameters can allow us to better satisfy one empirical desiderata—retrodiction—but only at the cost of another important desiderata—prediction (see especially Forster & Sober 1994).

Similarly, the representational flexibility provided by fictionalization might help us better predict and explain one aspect of a phenomenon, but only at the cost of a diminished ability to predict and explain another—namely, the aspect that is fictionalized. In the case of backpropagation, let us suppose that fictionalization allows us to establish link weights that will produce correct behavior, and in the case of the geon-based categorization model that the use of FELs can help us make predictions about property inference in object categorization. However, insofar as each of these model components falsely describes an aspect of the systems they are about, these benefits come at a cost. In the case of backpropagation, we cannot offer true predictions or explanations about link weight updating, and in the geon-based model, we cannot offer true explanations or predictions about synchronization. This point can be driven home by consideration of the counterfactuals implied by the fictional components; for if they are indeed fictional, then models using backpropagation imply specific but false counterfactuals about learning curves and

these parameters exhibit a high degree of variability in the literature, and the number of functions that can be approximated by backpropagation-trained neural networks within this space is considerable.

interventions on error signal distribution, and models using FELs imply specific but false counterfactuals about the fine-grained timing of synchronization and interventions on information transfer from one region to another.

However, Bokulich (2008, 2012) has recently defended the legitimacy of model fictions by appealing to their counterfactual power. In particular, she claims that fictions can explain by capturing a "pattern of counterfactual dependence of the relevant features of the target system on the structures represented by the model" (2008a, 226). Bokulich focuses on examples from physics, with the most discussed case being the periodic model of atomic orbits; the idea is that even if subatomic particles are ultimately quantum rather than classical in nature, the periodic model allows us to answer a range of *what-if-things-had-been-different* questions about how quantum wave functions would change in certain circumstances, based on counterfactuals about how the classical periodic orbits might would change. We might think that this defense could generalize to fictional models in the special sciences like psychology—for how do we know that backpropagation learning does not share a similar pattern of counterfactual dependence with the real details of synaptic adjustment in learning, whatever they might actually be?

Two relevant problems with Bokulich's defense of explanatory fictions are noted by Schindler (2014). The first is that Bokulich accepts that not every fictional model which happens to make accurate predictions will be explanatory, so she still needs some way to distinguish genuine patterns of counterfactual dependence from cases where the correspondence between the activity of the false model and the actual system is merely coincidental. The only obvious way to verify this dependence is with an account of model evaluation that is parasitic upon a true explanation of the phenomena, which in the present case would threaten the functionalist's dedication to taxonomic autonomy.[10] The second major problem noted by Schindler is that, since model fictions do not depict actual causes, they still cannot support counterfactuals regarding interventions on the false components. Interventions are standardly recognized in important counterfactual accounts of explanation and explanatory power (Woodward 2005;

[10] Schindler (2014, 1746) notes that it is ultimately the quantum mechanical models, together with a 'translation key', that ends up playing this justificatory role in Bokulich's analysis of periodic orbits in physics.

Ylikoski & Kuorikoski 2010), so it is unclear why the loss of counterfactuals pertaining to interventions should not be viewed as a clear disadvantage when comparing fictional models to more accurate alternatives.[11]

A further functionalist response here might be that these false and omitted counterfactuals may concern some dimension of the phenomena that fall outside the explanatory scope of the models, and so should not be counted as a points against them. Researchers who attach disclaimers or seek alternatives to model fictions might be interested in developing more mechanistically accurate models of these phenomena, but this should not be counted against modelers with purely functionalist aims. Consider Weiskopf's ultimate defense of FELs:

> "We might say: there is something that does what FELs do, but it isn't an entity or a link or anything of that sort. FELs capture the general characteristic of neural systems that they often fire in synchrony. We can model this with FELs and lose nothing of interest." (2011, 332)

Yet FELs imply more than their functional profile—that is what distinguishes fictions from black box terms and functional abstractions—and it is unclear why modelers should be uninterested in the way that real cognitive systems achieve synchrony. The true explanation for synchronization will be of value not only because it provides additional detail at lower levels of description, but also because it will support more counterfactual knowledge at the psychological level of description—especially by helping us distinguish causal counterfactual dependence from coincidence and confound, predict the effects of interventions on the system, and anticipate finer-grained aspects of object categorization data such as response times or learning curves.[12] Moreover, the parochialism behind this functionalist response does not reflect the general attitude or epistemic position of psychologists—who are never in the epistemic position to strictly define and delimit the borders of the phenomenon that they are attempting to predict

[11] Weiskopf (2011, 318) argues that we should distinguish "allowing control and manipulation" from "being able to answer counterfactual questions", recommending a metric of normative assessment for explanations that is neutral between the two. However, counterfactuals about the results of interventions are still counterfactuals, and even a neutral metric would disadvantage models that do not capture the results of interventions on a system's behavior.

[12] For example, Hummel & Biederman (1994, 511) themselves espouse an interest in the way that visual attention may help avoid accidental synchronization, an interest not addressed by the use of FELs.

and explain, even at the psychological level (the attempt to do so being one of the primary mistakes of radical behaviorism—Greenwood 1999).[13]

To summarize, it is not enough to simply state that a functionalist will be uninterested in the costs of the false counterfactuals implied by model fictions or the omitted counterfactuals missed by not having a true explanation of how some phenomena is produced. Interpreting fictional properties as natural kinds reads a metaphysical honorific onto these models, so it is not enough to point out the consistency of functionalist interpretations of these cases. We must also be told why the functionalist interpretation of these models is to be preferred over mechanistic alternatives. The common currency in arbitrating between functionalist and mechanistic interpretations, I have supposed, is counterfactual power, with the interpretation that supports more genuine counterfactuals being preferable, ceteris paribus. The disadvantages of fictional components we have just considered should be counted as reasons to suspect that they lack the relevant forms of kind-making stability.[14] So while FELs and backpropagation do play an important role in getting their respective models to function, they should be regarded as mere conveniences, rather than kinds. Let us thus set aside fictionalization and consider other functional modeling activities with less obvious explanatory costs.

Reification

Compared to fictionalization, reification is prima facie more plausible as a source of functional kinds. According to Weiskopf's account, reification occurs when a modeler introduces a division between model components that does not correspond to a structural division in an underlying mechanism. To contrast with his account of fictionalization, the systems described by the model really do possess the subcapacities attributed to the reified components—those subcapacities just cannot be localized to particular parts of an underlying mechanism. In this section I review general reasons to think that reification is also problematic as an origin for functional kinds by imposing a dilemma: reified

[13] Indeed, it seems an overreach to read this parochialism into Hummel & Biederman, who at times espouse agnosticism regarding the interpretation of FELs, noting that it "remains an open question whether a neuroanatomical analog of FELs will be found to exist" (1992, 510).

[14] Though FELs have been conserved in later iterations of JIM, they have not appeared in any other object categorization models—and indeed have been noted as a weakness of this model by critics (e.g. Robbins 2004).

decompositions will (again) either block us from recognizing important counterfactual generalizations or reduce to a species of the third type of modeling activity, functional abstraction (to be considered in turn).

Though Weiskopf does not draw the distinction, it is important to note that there are at least two different forms of reification that need to be treated separately. I will refer to the first as ‘fissional reification’ and the second as ‘fusional reification’. In fissional reification, we introduce two or more distinct components whose causal capacities are actually possessed by the same underlying part of the system (or the system as a whole). In fusional reification, we introduce a component whose causal capacities are actually distributed amongst distinct parts of the system.

To evaluate fissional reifications, let us (following Weiskopf-2011a, 328-329) continue the central example of the previous subsection and consider work on representation in connectionist networks. After having been trained to solve a categorization task, clustering algorithms can be run over those networks, shared hidden-layer activation patterns can often be located that extensionally correspond to everyday concepts (Shea 2007). These hidden-layer activation patterns are then often treated by modelers as the “representations” learned by the network. Additionally, connectionist networks are often described as modeling “inferences”; trained networks can be fed a series of novel exemplars, with their category values then “inferred” by the propagation of activity through the layers of the network. However, many authors have noted that there is something awkward about describing connectionist networks as both possessing representations and additionally performing inferences, for in connectionist networks both representation and inference are implemented by the same activation vectors (Clark 1991a, 1991b). By describing these networks as both possessing representations and separately performing inferences on those representations, we reify as distinct two entities that are not implemented by distinct mechanisms.

As principle for model evaluation, I offer a general argument against fissional reification called the “*A* without *B*” challenge. This challenge claims that for any two subcapacities *A* and *B*, if the system cannot perform *A* without engaging the very same mechanism that performs *B*, then an explanation that construes *A* and *B* as distinct subcapacities will have less counterfactual power than an otherwise identical model that depicts them as two aspects of the same capacity. The argument is as follows. First, identical mechanisms always have identical causal dispositions or causal powers (i.e., causal powers supervene on mechanistic structure). Second, if two models describe the same mechanism, the depicted aspects of the system will exhibit all the same causal dispositions or powers in the same circumstances. Third, if *A* and *B* cannot be localized to different parts of the same mechanism, then they must be localized to the same whole mechanism, and any time the system *A*'s it could *B*, and vice versa. Therefore, a "holistic" model which captures these counterfactuals (by depicting the unity of *A* and *B*) will be more powerful than a reified decomposition that does not. To summarize, a reified decomposition taken at face value gives us the impression that *A* could be engaged without engaging *B*, or intervened upon without intervening upon *B*; but if *A* and *B* are implemented by the same whole mechanism, this impression is always false.

To draw out this moral in the case connectionist networks, consider the set of counterfactuals that might be inferred from a network model of categorization that reified both representations and inferences. Such a model would imply that inferential resources could be engaged separately from representational resources and vice versa. Where representation and inference are actually implemented by the same activation vectors, these counterfactuals will always be false. For example, the reified model gives us the false impression that we could prime an inference rule without simultaneously priming a set of associated representations, or that we could add representations to the network without subtly altering generalization patterns for the networks' other inferences; but these counterfactuals are all false. Again, the point becomes even more obvious if we expand our interest to counterfactuals pertaining to interventions on the properties depicted by the model components; for a reified decomposition falsely implies that we could, e.g., ablate the resources responsible for a particular inference rule while leaving its associated representations available for other inferences—as we could, for example, disable a particular arithmetic operation on a microprocessor by damaging a particular set of transistors, while leaving other operations and data intact. Moreover, since the subcapacities attributed to both *A* and *B* would both still be included in a unified model, it is not clear that depicting their unity comes with any comparable counterfactual cost.

To adapt a classic functionalist rejoinder, we might argue that fissional reification can compensate us for this loss of counterfactual power by increasing our understanding of *how* the system performs a task. Following Cummins (1977, 1983), let us call this the "analytical strategy"; the idea is that showing how a "sophisticated performance" emerges from "unsophisticated performances in a sophisticated order" (Cummins 1977, 270) makes its own distinctive contribution to explanatory power. To consider a familiar example, we might think that breaking long division down into a series of simpler steps about copying numbers, dividing integers, subtracting, adding, and writing remainders renders an apparently difficult task intelligible in a way that is independent of counterfactual power. Cummins has emphasized that such decompositions offer an entirely different mode of explanation from mechanistic models, and taking analytic intelligibility to contribute a distinctive explanatory desideratum might be thought to support the idea that fissionally reified models should not be considered deficient to mechanistic explanations.

However, it is far from obvious whether such a tradeoff of apparent intelligibility for counterfactual power can be justified—and there are well-known reasons for doubt. Against this bargain, the experience of understanding or intelligibility is a mercurial psychological response that often tracks superficial features of models and explanations. A feeling of understanding may arise merely from familiarity, or from unreliable biases such as hindsight and overconfidence (Trout 2002).[15] At the very least, proponents of reification should explicitly discuss this tradeoff and the benefits that are meant to outweigh this loss of counterfactual power. Furthermore, if this sense of analytic intelligibility cannot be converted to counterfactual currency, then its defense will beg the question against mechanists. There is much more to be said on this subject, but as method of weighing analytic intelligibility against counterfactual power is clearly beyond the scope of the current paper, let us thus move to consider fusional reification instead.

Fusional reification occurs when a model introduces some component whose corresponding operations are actually possessed by a diverse set of resources or widely distributed throughout a system.

[15] Important recent explication of 'explanatory power' have valued the role of familiarity (e.g. see Ylikoski & Kuorikoski 2010 on 'cognitive salience') but only due to the pragmatic benefit that it is easier to infer counterfactuals from a familiar model, and not because it is an explanatory good in its own right.

Fusional reification is well-illustrated by Weiskopf's discussion of Just and Carpenter's (1992) 4CAPS model of sentence comprehension. The 4CAPS model is a hybrid rule-network model that contains a component labeled "activation", a limited resource which attaches to rules and representations, signaling the availability of the component for processing. The more often a component is used in the network, the more activation it consumes, eventually leading to diminished performance. As Weiskopf notes, activation in this model does not correspond to any single entity in the brain; it corresponds rather to "a whole set of resources possessed by neural regions: 'neurotransmitter function and various metabolic support systems, as well as the connectivity and structural integrity of the system'" (Just et al. 1999, p129, quoted on Weiskopf 2011a, p329).

Fusional reification can offer the following rebuttal to the "*A* without *B*" challenge: the theory of kinds is not concerned simply with providing predictions and explanations in particular cases, but also with how well those predictions and explanations *generalize* to a greater number of systems. This appeal has clear value in our ecumenical currency, for counterfactual power is enhanced not only by the number of distinct generalizations we can count, but also by the number of systems subsumed under those generalizations. Specifically, fusional reifications might score higher on the second metric in cases where there are a variety of parts that systematically implement subcapacity *A*, but do not all implement some other subcapacities *B1…Bn*, though in each system that implements *A*, the very same part implements at least one of the other subcapacities *B1…Bn*. Since the "lost" *B1…Bn* counterfactuals exposed by the "A without B" challenge here would not hold across all of these systems across which the *A* counterfactuals hold, the full range systems to which the *A* counterfactuals apply can only be captured by fusional reification.

For now, I merely note that fusional reification so understood is not distinct from functional abstraction. For, as we noted above, truly identical mechanisms have identical causal powers and dispositions, so different systems that implement *A* could not vary in their abilities to implement *B1…Bn* without being different mechanisms. Thus, the various parts that implement *A* must comprise a functionally abstract grouping—one that abstracts away from differences that are responsible for those parts' differential ability to implement *B1…Bn*. In the 4CAPS model, for example, the diverse parts of the system that correspond to "activation"—neurotransmitters, metabolic support systems, and structural features—must share the common (if highly abstract) capacity to influence cognitive processing in the relevant ways, even if they differ in a variety of other ways such as their tendency to be modulated by drugs, dependency upon oxygen and glucose, or connectivity constraints. Let us thus set aside evaluation of fusional reification for now, as it will stand or fall with the broader defense of functional abstraction considered next.

Functional Abstraction

Functional abstraction, finally, is the most traditional of Weiskopf's three etiologies for functional kinds. According to Weiskopf, functional abstraction occurs when we "decompose a modeled system into subsystems and other components on the basis of what they do, rather than their correspondence with organizations and groupings in the target system" (2011a, 329). As initially stated, it is unclear how this approach differs from the mechanistic approach to explanation; indeed, most mechanists think that this kind of strategy is the standard way to decompose a system into its explanatorily-relevant parts (see e.g. Craver 2006). As such, Weiskopf clarifies that functional decompositions can cross-cut other ways of decomposing what the system does, noting that "any system that instantiates functions *that are not highly localized* possesses this feature" (2011a, 329—my emphasis). As for what it means for a function to be "not highly localized," Weiskopf appeals to subcapacities implemented by highly diffuse or distributed parts of the system, as are purportedly found in systems neuroscience, where most cognitive functions are held to be implemented by networks spatially distributed throughout the brain (Anderson 2010; Sporns 2011). To summarize and simplify, a decomposition is functionally abstract in Weiskopf's terms just when it describes subcapacities that 1) the depicted systems really possess, 2) are, at least in principle, implemented by different parts of the system depicted, but, 3) cannot possibly, or can only with difficulty, be localized to particular parts of an underlying mechanism.

Note that there are two versions of this third characteristic suggested by the modifiers "cannot possibly" or "only with difficulty"; these versions must be distinguished, for they lead to importantly distinct positions. The first, stronger version emphasizes metaphysical considerations, holding that there is something about these explanations or the systems they describe that blocks a localized elaboration of the functional model in principle. The second version emphasizes epistemic factors, holding only that a localized elaboration of the functional decomposition is too complex, goes into unnecessary detail, or does not seem to be forthcoming given our current state of knowledge. The problem with the stronger version of this principle is that there is no reason to believe it when evaluating functionally abstract cognitive models; and the problem with the weaker version is that mechanists would not disagree with it, and it makes concessions that would require functionalists to abandon Weiskopf's favored response to Craver's mechanism sketches argument.

The major roadblock for the stronger, "in principle" version of this claim is that, if we take away models containing fictions and fissional reifications (with which we have already dealt), as well as those featuring continuously, reciprocally-coupled components,[16] there is a complete absence of cognitive models that are non-localizable in principle. It is straightforward to see why fictions and fissional reifications resist localization, for the systems described by these models do not actually contain any parts possessing these components' causal capacities. We can also see why models featuring continuous, reciprocally-coupled components resist localization, because the causal dispositions of any one component cannot be isolated from the dispositions of other components (Bechtel & Abrahamsen 2010). Concerning the remaining functionally abstract models, however, a generic sort of standoff develops: functionalists offer a case where we have so far been unsuccessful in showing how apparently explanatory decompositions can be localized to mechanism parts, and mechanists respond by appealing to the general track record in science of overcoming such apparently insurmountable challenges through

[16] The most plausible examples of in-principle non-localizable models in cognitive science are dynamical models from systems neuroscience in which "super- and subordinate levels are indistinct, most interactions are circular, and control is decentralized" (Sporns 2011, 193). However, such models do not easily fit the mold of Weiskopf's cognitive models, for they resist even functional decomposition and their main proponents eschew representational interpretation entirely (e.g. Stepp, Chemero, & Turvey 2011; Silberstein & Chemero 2013). For further arguments that such dynamical models fail to explain if they are non-mechanistic, see Kaplan & Bechtel (2011).

groundbreaking discoveries.[17] The only way to definitively determine whether a localized elaboration of a functionally abstract model can be produced is to *actually do the science.* Nevertheless, there remain no cases of a functional cognitive model with firm evidence that a localized elaboration, no matter how complex or convoluted, will never be forthcoming, so there is currently no reason to believe the stronger, metaphysical version of this tenet.

Much of Weiskopf's defense of functionally abstract kinds, however, is aimed not at establishing *in principle* unlocalizability, but rather at emphasizing the *heuristic utility* of functionally abstract explanations that are merely *difficult* to localize. At times, he concedes that functionally abstract decompositions may be mechanism sketches, while holding that the functional properties featuring in sketches should be granted kind status due to their utility in exploring the space of how-possibly explanations in the search for a structurally localized model. "There is not just one 'functional level'…" he writes,

> "…but rather a whole host of intermediate structures at varying degrees of abstraction from the underlying physical components…It can prove heuristically indispensible, once one has characterized the general function of a cell type or brain region to then propose a range of possible lower level mechanisms that might realize that function, then proceed to rule them out on the basis of side effects, predicted responses to various interventions, predicted anatomical consequences, and so on." (2011b, 254)

On this justification, commonly deployed functionally abstract properties should be counted as natural kinds because they can "serve a crucial heuristic role in discovering mechanisms" (2011b, 243). In short, these abstract model components function as explanatory "templates". These templates may all need to be localized to provide complete or maximally detailed explanations, the thought goes, but the fact that

[17] That such an standoff is not likely to resolve the dispute is evidenced by the number of cases about which philosophers agree on all the details but disagree on their interpretation; e.g. regarding lateral inhibition compare Shapiro (2004, 117-120) to Weiskopf (2011b, 236-239) or on network neuroscience compare (Bechtel 2011, 553) to Silberstein & Chemero (2013, 965-966).

the same templates show up again and again in the intermediate steps of localization should be acknowledged by our special science taxonomy.

The first thing to note about this weaker, ‘heuristic’ defense of functional abstraction is that it no longer offers a clear rebuttal to Craver’s “mechanism sketches” argument. Whereas we noted above that Weiskopf prefers to rebut this argument by denying its second premise, this heuristic defense appears instead to affirm its second disjunct: that functionally abstract models only support *how-possibly* explanations. And indeed, without providing some other reason to think that the functionally abstract model components cannot be localized, it appears to embrace the argument’s conclusion—because we cannot know whether a functional template provides an actual explanation in any particular case until we determine whether it can be localized.

The second (and related) thing to note about this heuristic defense is that it seems to reject functional kinds’ *direct* explanatory import. On this way of understanding matters, non-localized functional kinds are useful not because they themselves provide explanations, but rather because they are essential tools in the search for *how-actually*, localized explanations. This defense diverges from the traditional Fodorian position, which held that functional kinds feature directly in explanations of psychological phenomena through a process of deductive-nomological derivation. While Weiskopf calls this a “significant shift of emphasis” (2011b, 250), this move requires a departure from traditional ideas about kinds, and we should wonder whether any remaining debate with the mechanist is merely terminological—an exercise in relabeling. Perhaps we could call these functionally abstract templates “natural kinds”—but if they do not themselves feature in explanations of psychological phenomena (or do so only when localized), why should we?

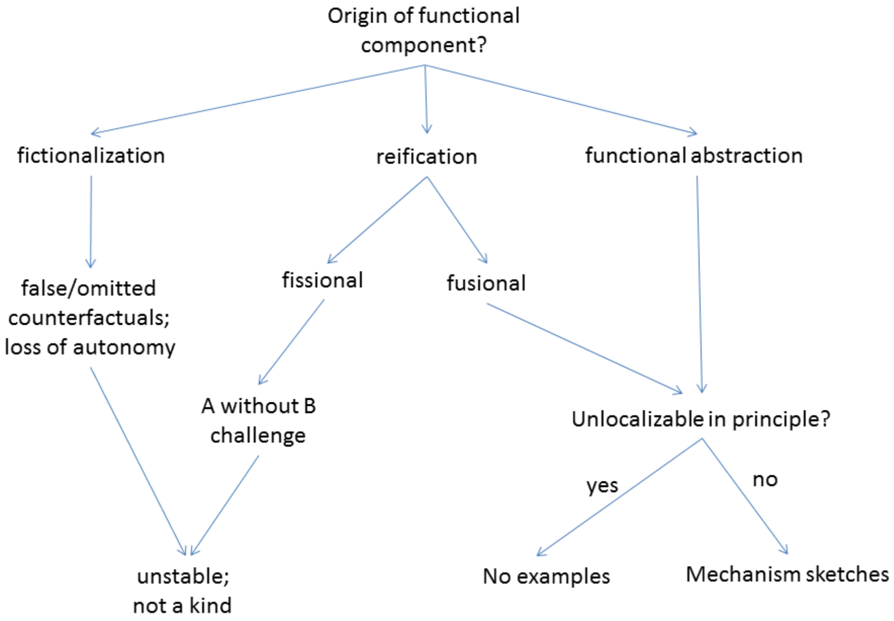

Figure 1. A map of the argument by cases of Section IV.

V. Three functionalist rejoinders

In this section, I will consider a series of functionalist answers to this final question about functionally abstract kinds. I will take it for granted here that fictional, reified, and reciprocally-coupled model components have already been taken off the table, focusing on the heuristic justification for functionally abstract kinds as useful tools in exploring the space of how-possibly explanations. A general theme will be that while these rejoinders are effective against certain varieties of reductionism, they present no challenge to non-reductive mechanistic position on kinds.

A. Not all functional categories are equally useful; and only functionalism about kinds can explain these differences.

Weiskopf challenges all accounts of kinds in the special sciences to explain why some functionally-defined categories are more useful than others in this process of elaboration. Against Shapiro's (2004) reductionist view that "functional concepts [merely] fix a range of 'analytic' truths about things that fall under them", Weiskopf notes that…

> "'…things that can be knocked over with a feather' is a functional category, but I doubt whether it is apt for a special science. Shapiro's view leaves us with no way of explaining why […]. What is *the point* of grouping things together by function if the grouping itself does not pick out an inductively potent kind?" (2011b, 247)

In other words, the thought is that the functional *kinds,* by contrast with non-kind functional categories, possess an inductive potency outstripping their explicit definition. These latter categories grant the models in which they occur a special kind of inductive unity, and this fact must be explained.

The problem with this defense is that the new model-based functionalism about kinds *also* fails to explain the difference between the useful and useless functional categories. Like the older Fodorian view it aims to supplant, it rather takes it as a primitive fact, discovered by the special sciences, that some functional components have the right kind of inductive "oomph". Weiskopf does note that "the sort of causal relations that qualify a grouping as a kind, on this view, are relations that enable a category to play a recurrently useful role in a range of models" (2011b, 253). However, rather than providing a systematic account of these relations that explains how they enhance the counterfactual power of models in which they occur, Weiskopf suggests that they are to be read off of the explanatory success of those models—thereby taking for granted just the thing that was to be explained. By contrast, the mechanistic approach to kinds is built on its answer to this crucial question—that the inductively potent special science categories will be those that robustly possess their characteristic causal powers (and, perhaps, others we have not yet discovered) in virtue of the operation of some shared underlying mechanism(s). Thus, this functionalist rejoinder not only misses its mark, but in so doing demonstrates a key strength of mechanism about kinds.

This is not to say that there are no other functionalist options to explore. The most promising strategy would appeal to some general organizing principles of cognitive systems. For example, expanding on ideas in Bechtel (2007) and Weiskopf (forthcoming), perhaps control theory could provide some answers, the idea being that once a system reaches a certain level of self-organizing complexity, the same types of control systems will tend to develop again and again in diverse substrates to achieve coordination amongst semi-autonomous subsystems (which might perhaps make sense of examples like CPGs). Or, elaborating threads from Burge (2010) and Millikan (2012), perhaps cognitive systems specialize in the detection of invariances across perceptually distinct situations, and this pressure will tend to induce the same functional organization in diverse mechanisms (which might make sense of examples like lateral inhibition and backpropagation learning). At any rate, these suggestions remain largely speculative, and it needs to be determined whether any such strategy could account for the stability of the specific list of kinds offered by Weiskopf without revealing these kinds to be mechanistic in nature.

B. Functionally abstract models can be called 'mechanistic', but only at the cost of stretching the label to meaninglessness.

Let us consider the charge that mechanists can interpret functionally abstract models as mechanistic, but only at the cost of stretching 'mechanistic explanation' past its breaking point. For example, consider Piccinini & Craver's claim that a functionally abstract property in systems neuroscience "may be so [spatially] distributed and diffuse as to defy tidy structural description, though it no doubt has one if we had the time, knowledge, and patience to formulate it" (2011, 291). Weiskopf calls this strategy "mechanism imperialism", worrying that it "[strips] the mechanistic program of any substantial commitment concerning the distinctive ontology of mechanisms" (Weiskopf, forthcoming). In short, the worry is that if mechanists cannot distinguish systems which possess a clear and relevant structural decomposition from those that resist such localized description, the "mechanistic/non-mechanistic" divide will become a distinction without any intelligible difference.

Two responses should be made to this rejoinder. First, the mechanist does not claim that all systems can be mechanistically explained, only that the behavior of systems that truly lack any intelligible mechanistic structure and organization cannot be explained. Perhaps there are some systems—say, large weather systems or liquefying gases—that, despite showing some macro-level regularities, are really chaotic at all other levels of description. The mechanist claims not that there must be some mechanistic decomposition of these systems that explains these macro-level regularities, only that if there is not, then those regularities cannot be explained. The "imperialist" in this case simply thinks that there are no such chaotic systems in cognitive science, and that the brain, described at various levels of abstraction, is ultimately the mechanism that explains cognition.

Second, the claim that mechanists cannot acknowledge the differences between paradigm localizable and only weakly localizable models also rings hollow, for mechanists such as Bechtel & Richardson (2010) have for years articulated the idea of a continuum of localizability. Paradigm examples of mechanistic explanation like long-term potentiation fall to on one end of this spectrum, and paradigm examples of weakly-localizable explanations like spatially distributed neural networks fall to the other. And again some systems—the large weather systems and liquefying gases, perhaps—may drop off the end of this scale entirely. Setting aside the straw-manning assumption that mechanists can only localize with the precision of a 19th century phrenologist, Piccinini & Craver (2011) and Kaplan & Craver (2011) simply argue that explanations in systems neuroscience, given the widespread practice of localizing psychological functions to specific and located distributed brain networks, do not drop off the scale. And while adopting a different attitude towards the explanatory completeness of structurally abstract models, Levy & Bechtel argue that abstract explanations in systems neuroscience with components like motifs are mechanistic because they are derived from mechanisms by abstracting away from structural detail (but crucially "a more concrete description is possible" should our explanatory purposes require it—2013, 242).[18] Though there are important differences between these two mechanist positions, neither abolishes the distinction between localizable and nonlocalizable models.

C. Systems implementing a functionally abstract kind may share a correspondingly abstract mechanistic structure, but these structure types are not "independently certified".

There remains a final arrow in the functionalist quiver, one of the oldest arguments in favor of functional kinds: that while their realizers may share some diffuse, abstract mechanistic structure at lower levels of description, these abstract structures are not "independently certified" as kinds in those lower-

[18] Levy & Bechtel emphasize that network motif models highlight the *organization* of neural mechanisms while omitting structural detail of the parts so organized. Such models are to be distinguished from nonmechanistic decompositions because systems can only be organized in the relevant sense if they "exhibit a certain form of dependency of the whole on its parts" (2013, 244). Components in abstract mechanistic models must at least in principle be localizable, even if such detail is irrelevant to the modeler's current explanatory purposes.

level sciences.[19] It would be ontologically irrelevant, this line of thought goes, that we could gerrymander some kind of gruesome lower-level mechanistic description shared amongst all of a model component's realizers, for this structure would not itself independently count as a natural kind in any lower-level science.

That functionally abstract kinds share a correspondingly abstractly mechanistic description is borne out by a close examination of Weiskopf's examples. In the 4CAPS model, for instance, "activation" is introduced as a disjunction of lower-level kinds ("neurotransmitter function and various metabolic support systems, as well as the connectivity and structural integrity of the system"), localizes to resource utilization in a particular brain region ("considered as a resource pool"), and "is intended to correspond to the amount of brain activation observed with a neuroimaging measure in [that] corresponding area during the corresponding time interval" (Just et al. 1999, 129). So while different forms of resource utilization subsumed by "activation" are diverse at the finest grain of neurophysiological description, they influence processing in the relevant ways because they share abstract structural and organizational properties in the brain. Similarly in the case of CPGs, Prinz et al. note that their computational analysis did not explain *how* the various network models stably achieved the same functional profile, speculating that it would either arise from i) "local stability rules that set the properties of single-neuron excitability and/or synaptic strength" (an abstract structural hypothesis) or ii) "monitors of network performance, such as sensory feedback from target muscles" (which will either involve an external mechanism or reciprocal dynamical coupling) (2004, 1349). While such abstract descriptions are plausibly regarded as mechanistic (e.g. see again Levy & Bechtel 2013), they are indeed not "independently certified" in any lower-level science, in the sense that there would be little reason to group parts together into these categories except for their capacity to secure those higher-level functional profiles.

Though there is much more to be said about this independent certification criterion, it can in the present context be curtly dismissed as an atavism from the older debate between functionalists and

[19] Throughout this section, I use talk of "higher" and "lower" levels to discuss this functionalist rejoinder without ultimately endorsing the intelligibility of such talk. For skepticism about such terminology, see Craver (2007, Ch5).

reductionists. Reduction, as classically conceived by the Nagelian bridge-law program, is indeed a relation between kinds of one science and metaphysically prior kinds of another, lower-level science; but few new mechanists are reductionists, as traditionally conceived or otherwise. Because the mechanist's insistence that model components be localizable is bolstered by concerns about the counterfactual power of explanations rather than assumptions about any ordering or priority amongst different scientific "levels", mechanists can happily concede that the abstract structure underlying a special science kind would not be independently certified without undermining the claim that those abstract structures explain how those systems stably implement their functional profiles.[20] Indeed, prominent mechanists have endorsed just this sort of agnosticism (Craver 2006; Bechtel & Mundale 1999; Boyd 1999), and the mechanist's ability to offer coherent approaches to kinds, explanation, and induction without assuming either autonomy or an ordering amongst the sciences can be seen as an advantage of the position.

VI. Concluding remarks: Kinds and scientific disagreement

To sum up, we began by asking whether functionally-defined model components could stably play the explanatory role in the special sciences characteristic of kinds. We tackled this question by evaluating the modeling practices that introduce functional components: fictionalization, reification, and functional abstraction. Fictional and (fissionally) reified components were found to have clear disadvantages to the counterfactual power of models that contain them, and so these sources of functional kinds were eliminated from consideration. The evaluation of functionally abstract components, however, proved more complex. It threatened to degenerate into a pair of terminological disputes: whether categories that are only indirectly explanatory (by helping us sort through the range of "how possibly" explanations) could properly be called "natural kinds", and whether weakly localizable models could properly be called "mechanistic". Neither of these frames provides a particularly useful way forward in the dispute between functionalists and mechanists about the nature of kinds, so I end by sketching an alternate frame.

[20] A commonly-overlooked issue here is that mechanists about kinds typically concede that the mechanisms securing the homeostatic stability of a kind may be externally located from the system depicted—e.g., constraints on reproduction or predation may ensure that members of a biological species reliably possess their characteristic phenotypic properties (Boyd 1999).

Throughout our discussion, the unwillingness to appeal to information as to whether model components could be localized was repeatedly shown to have disadvantages to the counterfactual power of functional models. Each time, a tempting functionalist response was to adopt the parochial attitude that the lost counterfactual generalizations were somehow not relevant to the modeler's aims, and so "nothing was lost" by ignoring them.

This response poses a danger to the health of empirical debate—for what are we to do when different research groups disagree as to which aspects of a phenomenon are explanatorily relevant? The intention to engage in scientific investigation should imply a willingness to engage in a collaborative search for more powerful explanatory frameworks with other researchers coming from different epistemic perspectives. As Boyd (1999) has repeatedly emphasized, the individuation conditions for mechanistic kinds are not determined just by their functional profiles, but rather by an "accommodation" between these characteristic profiles and underlying mechanisms that could nonaccidentally produce them. Because we have imperfect epistemic access to these underlying structures, searching for them can bring together a range of perspectives in a progressive research program that iteratively elaborates the kind's nature. By contrast, the functionalist attitude encourages each party to suppose that the aspects of the phenomena captured by their models are the definitive or characteristic ones, dismissing those captured by their opponents' models as irrelevant, fictions, or idealizations. Such impasses have been fairly common in cognitive science's short history, and where participants are unwilling to arbitrate their disputes by appeal to underlying mechanistic structure, they can degenerate into scholastic disputes in which "the relevant models are underdetermined, and the modelers themselves perhaps overly determined to win an unwinnable competition" (Bechtel & Abrahamsen 2010, 330).

Much of the difficulty here arises from the partial and provisional way that models depict phenomena. There are several different ways that apparently incommensurable models could relate to one another: they could genuinely disagree about the nature of the same target kinds (with the dispute to be arbitrated empirically), they could describe the same underlying kind at different levels of abstraction or idealization (and so not actually be in conflict), or they might depict different kinds of phenomena (one, perhaps, implementing the other). The mechanist counsels an appeal to underlying structures to arbitrate amongst these different options, but such structure is eschewed by the functionalist. If kinds are individuated only by functional profile, and kindhood determined only by the empirical success of models featuring that kind as a component, then in cases where successful models possess components with different functional profiles, those components must be split into different kinds. While this counsel has the advantage of precluding some ultimately fruitless disagreements, it also precludes other critical engagements that may be productive. The result is a worldview featuring many incommensurable taxonomies, with numerous, cross-cutting kinds whose relationship to one another is ultimately unclear.

By contrast, the search for nature's joints originally began with the hope that locating them could help us decide, for two or more putative categories whose ultimate nature is empirically uncertain, whether science would be more fruitful—across a range of explanatory interests—if they were lumped or split. While the proposed model-based criterion for kindhood excels at splitting, it provides little guidance on lumping—and as such, may not satisfy the hankering for parsimony which inspired the enthusiasm for kinds. Future work on functional kinds might address this dissatisfaction by concentrating on a pair of pressing and difficult issues: how to decide when (1) a parochial attitude towards lost counterfactual power is legitimate and when it is ad hoc flummery, and when (2) two apparently distinct functionally-abstract kinds should be lumped (rather than split).[21] Only with adequate answers to these two questions will we have a theory of functional kinds worthy of the name.

References

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and brain science*s, 33(04), 245-266.

Bechtel, W., & Mundale, J. (1999). Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of Science*, 66 (2), 175-207.

[21] For example, in cases where two researchers from different epistemic perspectives attribute two different functional profiles to the same underlying kind, we might treat those functional profiles as explanatory heuristics that can be revised and improved through collaborative critical interaction (e.g. Hong & Page 2001). What remains to be articulated are the conditions, for the functionalist, where such fusing should be judged the *correct* outcome, as opposed to a mistake (or a changing of the subject).

Bechtel, W. (2007). Biological mechanisms: organized to maintain autonomy. *Systems biology: philosophical foundations*. 269-302. Elsevier, Amsterdam.

Bechtel, W. (2011). “Mechanism and Biological Explanation.” *Philosophy of Science* 78 (4): 533–58.

Bechtel, W. (2010). “Dynamic mechanistic explanation: computational modeling of circadian rhythms as an exemplar for cognitive science”. *Studies in History and Philosophy of Science* 41: 321-333.

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

Bickle, J. (2010). Has the last decade of challenges to the multiple realization argument provided aid and comfort to psychoneural reductionists?. *Synthese,177*(2), 247-260.

Bokulich, A. (2008). Can classical structures explain quantum phenomena?. *The British Journal for the Philosophy of Science*, 59(2), 217-235.

Bokulich, A. (2011). How scientific models can explain. *Synthese*, 180(1), 33-45.

Bokulich, A. (2012). Distinguishing explanatory from nonexplanatory fictions. *Philosophy of Science*, 79(5), 725-737.

Boyd, R.,(1991). Realism, Anti-Foundationalism and the Enthusiasm for Natural Kinds, *Philosophical Studies* 61: 127–148.

Boyd, R. (1999). Kinds, complexity, and multiple realization. *Philosophical Studies*, 95(1), 67-98.

Burge, T. (2010). *Origins of objectivity*. Oxford University Press.

Clark, A. (1991). Systematicity, structured representations and cognitive architecture: A reply to Fodor and Pylyshyn. In *Connectionism and the Philosophy of Mind* (pp. 198-218). Springer Netherlands.

Clark, A. (1991b). *Microcognition: Philosophy, cognitive science, and parallel distributed processing*. MIT Press.

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

Cummins, R. (1977). Programs in the explanation of behavior. *Philosophy of Science*, 269-287.

Cummins, R. C. (1983). *The nature of psychological explanation*. MIT Press.

Fitzsimonds, R. M., Song, H. J., & Poo, M. M. (1997). Propagation of activity-dependent synaptic depression in simple neural networks. *Nature*, 388(6641), 439-448.

Fodor, J. A. (1974). Special sciences (or: the disunity of science as a working hypothesis). *Synthese*, *28*(2), 97-115.

Fodor, J. (1997). Special sciences: Still autonomous after all these years. *Noûs* 31(s11), 149-163.

Forster, M., & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science*, 45(1), 1-35.

Gluck, M. A., & Myers, C. E. (2001). *Gateway to memory: An introduction to neural network modeling of the hippocampus and learning*. MIT Press.

Greenwood, J. D. (1999). Understanding the “cognitive revolution” in psychology. *Journal of the History of the Behavioral Sciences*, 35(1), 1-22.

Griffiths, P. E. (1997). *What emotions really are: The problem of psychological categories* (p. 114). Chicago: University of Chicago Press.

Haykin, S. S., Haykin, S. S., Haykin, S. S., & Haykin, S. S. (2009). *Neural networks and learning machines (Vol. 3)*. Upper Saddle River: Pearson Education.

Hong, L., & Page, S. E. (2001). Problem solving by heterogeneous agents. *Journal of Economic Theory*, 97(1), 123-163.

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

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

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

Kaplan, D. M., & Craver, C. F. (2011). The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective. *Philosophy of Science*, *78*(4), 601-627.

Kaplan, D. M., & Bechtel, W. (2011). Dynamical models: An alternative or complement to mechanistic explanations?. *Topics in Cognitive Science*, 3(2), 438-444.

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

Levy, A., & Bechtel, W. (2013). Abstraction and the organization of mechanisms. *Philosophy of science*, 80(2), 241-261.

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

Machery, E. (2005). Concepts Are Not a Natural Kind. *Philosophy of Science*,*72*(3), 444-467.

Millikan, R. G. (2012). Are there mental indexicals and demonstratives?. *Philosophical Perspectives*, 26(1), 217-234.

Needham, P. (2011). Microessentialism: What is the argument? *Noûs*, 45(1), 1-21.

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

Prinz, A. A., Bucher, D., & Marder, E. (2004). Similar network activity from disparate circuit parameters. *Nature neuroscience*, 7(12), 1345-1352.

Quine, W. V. O. (1969). Natural kinds. *Essays in Honor of Carl G. Hempel: A Tribute on the Occasion of his Sixty-Fifth Birthday* (Vol. 24), N. Rescher (ed.). Springer.

Robbins, S. E. (2004). On time, memory and dynamic form. *Consciousness and Cognition*, 13(4), 762-788.

Selverston, A. I. (1980). Are central pattern generators understandable? *Behavioral and Brain Sciences*, 3(04), 535-540.

Schindler, S. (2014). Explanatory fictions—for real?. *Synthese*, 191(8), 1741-1755.

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

Shea, N. (2007). Content and its vehicles in connectionist systems. *Mind & Language*, 22(3), 246-269.

Siegelmann, H. T., & Sontag, E. D. (1991). Turing computability with neural nets. *Applied Mathematics Letters*, 4(6), 77-80.

Silberstein, M., & Chemero, A. (2013). Constraints on Localization and Decomposition as Explanatory Strategies in the Biological Sciences. *Philosophy of Science*, *80*(5), 958-970.

Sporns, O. (2011). *Networks of the Brain*. MIT press.

Stepp, N., Chemero, A., & Turvey, M. T. (2011). Philosophy for the rest of cognitive science. *Topics in Cognitive Science*, 3(2), 425-437.

Stork, D. G. (1989). Is backpropagation biologically plausible?. In *Proceedings of the International Joint Conference Neural Networks*. IJCNN., (pp. 241-246). IEEE.

Trout, J. D. (2002). Scientific Explanation and the Sense of Understanding. *Philosophy of Science*, 69(2), 212-233.

Van Brakel, J., (2000). *Philosophy of Chemistry*. Leuven: Leuven University Press.

Walmsley, J. (2008). Explanation in Dynamical Cognitive Science. *Minds and Machines* 18(3): 331–48.

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

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

Weisopf, D. (forthcoming). The reality of cognitive models. *Integrating Mind and Brain Science: Mechanistic Perspectives and Beyond*, David Kaplan (ed.). Oxford University Press.

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

Ylikoski, P., & Kuorikoski, J. (2010). Dissecting explanatory power. *Philosophical Studies*, 148(2), 201-219.