European Journal for Philosophy of Science (2021) 11: 103
https://doi.org/10.1007/s13194-021-00420-2

PAPER IN GENERAL PHILOSOPHY OF SCIENCE

New functionalism and the social and behavioral sciences

Lukas Beck[1] · **James D. Grayot**[2]

Received: 7 January 2021 / Accepted: 1 October 2021 / Published online: 2 November 2021
© The Author(s) 2021

Abstract
Functionalism about kinds is still the dominant style of thought in the special sciences, like economics, psychology, and biology. Generally construed, functionalism is the view that states or processes can be individuated based on what role they play rather than what they are constituted of or realized by. Recently, Weiskopf (2011a, 2011b) has posited a reformulation of functionalism on the model-based approach to explanation. We refer to this reformulation as 'new functionalism'. In this paper, we seek to defend new functionalism and to recast it in light of the concrete explanatory aims of the special sciences. In particular, we argue that the assessment of the explanatory legitimacy of a functional kind needs to take into account the explanatory purpose of the model in which the functional kind is employed. We aim at demonstrating this by appealing to model-based explanations from the social and behavioral sciences. Specifically, we focus on preferences and signals as functional kinds. Our argument is intended to have the double impact of deflecting criticisms against new functionalism from the perspective of mechanistic decomposition while also expanding the scope of new functionalism to encompass the social and behavioral sciences.

Keywords Functionalism · Mechanism · Model-based explanations · Choice-theory

✉ Lukas Beck
lb760@cam.ac.uk

James D. Grayot
james.grayot@gmail.com

[1] Department of History and Philosophy of Science, University of Cambridge, Free School Lane, Cambridge CB2 3RH, UK

[2] Department of Theoretical Philosophy, University of Groningen, Oude Boteringestraat 52, 9712 GL Groninge, The Netherlands

1 Introduction

Whether functionalism offers a legitimate basis for providing explanations in the special sciences is an on-going debate in philosophy of science. The standard view of functionalism holds that some states and processes can be individuated based on what *role* they play rather than on what they are strictly *constituted of*.[1] Functionalists maintain that such functionally individuated states and processes can legitimately figure into explanations in the special sciences.

While functionalism is still dominant in many domains of the biological, behavioral, and psychological sciences, evolving debates in philosophy of science indicate problems with traditional arguments in support of functionalism. On the one hand, early defenders of functionalism, like Fodor (1974), argued that the existence of well-supported laws involving functional kinds vindicates functionalism as an explanatory strategy. The main problem with this defense is that it has become doubtful whether there are any such laws in the special sciences (Cartwright, 1999; cf. Kincaid, 1996, 2004). On the other hand, recent growing support for mechanistic explanation in the special sciences suggests that functionalism's emphasis on roles can be problematic *if* it prevents scientists from decomposing systems under investigation into mechanisms (Craver & Bechtel, 2006; Bechtel & Richardson, 2010). According to mechanists, it is through mechanistic decomposition that the special sciences ultimately achieve their explanatory aims. This leaves functionalism in a precarious position regarding whether and how functionally individuated states and processes can legitimately figure into explanations.

To address these concerns, Weiskopf (2011a, 2011b, 2017) has posited a reformulation of functionalism on the model-based approach to explanation—we refer to this reformulation as new functionalism.[2] Roughly, new functionalism holds that functionally individuated states and processes constitute *kinds* if they figure into a range of successful models instead of well-supported laws. Moreover, according to Weiskopf, functional kinds can be individuated via one of three different strategies: *fictionalization*, *reification*, or *abstraction*. Each of these strategies indicates why a particular functional kind is not amenable to mechanistic decomposition. Taken together, both of these elements are intended to show that functional kinds can serve as central explanatory units in the special sciences even if they are not amenable to mechanistic decomposition.

However, new functionalism has also received some critical reactions. In a recent iteration of the debate between functionalists and mechanists, Buckner (2015) introduces a dilemma for Weiskopf's account. Buckner argues that functional kinds are either still amenable to mechanistic decomposition (this holds for abstractions) or that the models involving functional kinds incur a loss of counterfactual power (this holds for fictions and reifications). He takes this dilemma to pose a serious challenge to new functionalism.

[1] The term 'individuation' refers to the description by which states and processes are distinguished—it is what countenances their role in explaining a phenomenon.

[2] We owe this term to Buckner (2015).

In this paper, we aim at defending new functionalism by recasting it in light of the concrete explanatory aims of the special sciences, more broadly construed. We argue that the assessment of the explanatory legitimacy of functional kinds also needs to consider the *purpose* of the model in which a functional kind is employed. In this respect, we hold that Weiskopf's account neglects the diversity of explanatory purposes found in the special sciences. However, and more importantly, we show that once these explanatory purposes are taken into account, the horns of Buckner's dilemma will often turn out to be dull. We demonstrate this by appealing to model-based explanations from the social and behavioral sciences. Specifically, we make the case that *preferences* and *signals*—understood as functional kinds—are typically not affected by the dilemma given the explanatory purposes of the choice theoretic models in which they are employed. We take our argument to have the double impact of deflecting Buckner's dilemma while also expanding the scope of new functionalism to encompass the social and behavioral sciences.

In what follows, section 2 situates Weiskopf's new functionalism in contrast to traditional defenses of functionalism, and section 3 spells out Buckner's mechanistic critique in response. Section 4 closely analyzes preferences and signals to highlight the shortcomings of both Weiskopf's account and Buckner's rejoinder. Section 5 repackages new functionalism in light of our analysis and addresses the charge that our defense leads to parochialism or anything-goes pluralism. Section 6 concludes.

2 Functionalism in the special sciences

Fodor's (1974, 1997) seminal argument for functional kinds relies on postulating the existence of law-like generalizations in the special sciences. The main target of his argument is physicalist reductionism, specifically the view that all special science laws are reducible to physical laws.[3] He argues that a special science law $S_1x \rightarrow S_2x$ is reducible to a physical science law $P_1x \rightarrow P_2x$ *iff* there exist bridge laws connecting special science predicates (that figure into a special science law) to physical predicates (that figure into a physical science law). Bridge Laws, e.g., $S_1x \leftrightarrow P_1x$, express contingent event identities stating that every event consisting of some *x* satisfying *S* is identical with some event consisting of *x* satisfying *P*. They are contingent in the sense that they cannot be established a priori. According to Fodor, what the reductionist needs to establish is that there are, in fact, bridge laws that connect each and every special science *natural kind* predicate with a physical science natural predicate. Here, natural kind predicates are those predicates that figure into the laws of (completed or ideal) science. Yet, Fodor argues that the natural kind predicates in a special science will most likely correspond to a *heterogeneous disjunction* of physical predicates that do not constitute natural kinds, i.e. $S_1x \leftrightarrow P_1x \vee P_2x \vee P_3x \vee \ldots . P_nx$, where $P_1x \vee P_2x \vee P_3x \vee \ldots . P_nx$ is not a natural kind predicate. Hence, reductionism is unlikely to succeed.

[3] Fodor clarifies that this version of reductionism is "a stronger one than many philosophers of science hold" (114, f.n. 2). Though, in his follow-up (1997) article, he attributes this view to Kim (1992).

Yet, if the natural kind predicates of the special sciences correspond to a diverse set of physical predicates, how should we individuate them? Functionalists like Fodor argue that we can individuate special science predicates via the *role* they play. For example, in order for something to be money, it is important that it functions as a medium of exchange independent of what it is constituted of (gold, silver, copper, etc.). If this argument is correct, the special sciences are unlikely to reduce to physics. Therefore, they maintain autonomy with respect to their taxonomy.

To be clear, what's important for Fodor's argument is not just that some predicates may not correspond to physical natural kind predicates, but more specifically, that the natural kind predicates of the special sciences do not correspond to physical natural kind predicates. If natural kind predicates are those predicates that figure into the laws of (completed or ideal) science, we need laws of special science in order to have special science natural kind predicates. However, according to some, there is an emerging consensus in the philosophy of science that the special sciences are not primarily in the business of discovering laws (Cartwright, 1999; cf. Kincaid, 1996, 2004). Rather, they are in the business of devising and testing *models* which serve various explanatory functions with regard to their target phenomena. If this is the case, then it seems philosophers of science cannot secure the autonomy of the special sciences, and the legitimacy of using functionally individuated kinds in explanations by appealing (solely) to the existence of laws involving such kinds, for the simple reason that it is contested whether there are any such laws in the special sciences.[4]

There are, however, a growing number of alternatives to law-based defenses of functionalism in the special sciences. One alternative can be found in the work of Ross & Spurrett (2004, see also, Ladyman et al., 2007, Ladyman, 2008), which aims at exposing the limitations of reductionist approaches to causal explanation (i.e., Kim, 1992). In particular, they argue that non-reductionistic approaches to explanation are better suited to identify and track properties and dispositions of macro-level entities given that these are primarily epistemic properties and not necessarily causal ones. Moreover, Ross (2005, 2006) has expanded on this idea, offering a comprehensive account of functionalism in the context of economics. This account builds on Dennett's notions of intentional systems and 'real patterns' ontology (Dennett, 1989, 1991). While we are sympathetic to the anti-reductionist stance of Ross and Spurrett, and find Ross's intentional-stance functionalism to be a promising take for microeconomic research, we here would like to focus on a more general formulation of functionalism, one that offers a direct replacement to the law-based arguments of Fodor.

This brings us to the second alternative: the *model-based approach to functionalism*. Weiskopf (2011a, 2011b) argues that legitimate functional kinds of the special sciences are those states or properties that feature in many coherently

[4] We do not take a stance on whether there are any special science laws as the answer to this question clearly depends on the particular notion of a law that one accepts. Though, we accept that, on some of the less demanding notions of laws, having a model that fulfills Weiskopf's criteria outlined below may already suffice for having a law (cf. Kincaid, 2004).

integrated and empirically successful models. In particular, he argues that a class of models needs to satisfy three conditions—these are:

1. *The models need to be well confirmed*. This implies that if there is a set of models that is, ceteris paribus, better supported by the available evidence, then this set is to be preferred and we should only consider the functional properties of these models to constitute natural kinds. (2011a, 336; 2011b, 252)
2. *Models should be representationally accurate*. That is, if there is a set of models that includes, ceteris paribus, more elements that are real parts of its targets or describes these elements in more (explanatory relevant) detail, we should only consider the functional properties of the more *representationally accurate* set of models to be natural kinds. (2011a, 336; cf. Gierre, 1988)
3. *Models should be genuinely explanatory.* This suggests that the set of models that can, ceteris paribus, answer more "what-if-things-had-been-different" questions is to be preferred and we should only consider the functional properties of these models to constitute natural kinds. (2011a, 320; 2011b, 250, 252)

Weiskopf also suggests that the models should be a good fit with our "general background knowledge" (2011a, 336). If we find a set of models that meets these criteria, he argues, we can take their functional categories (denoting the relevant states, processes, and properties) as providing legitimate explanations for the biological, behavioral, or psychological capacities under investigation.

But, for this argument to go through, we must accept two things: first, that when models explain, they explain in virtue of the functional categories, i.e., kinds, they posit; second, that what determines how functional kinds explain depends on how exactly they are supported by models, as opposed to laws. To this end, Weiskopf claims the following:

> Much of the work of building theories, models, and simulations in the biological, behavioral, and psychological sciences involves finding the appropriate concepts to use in analyzing a system. On the view I propose, functional categories are kinds when they are appropriate and useful for constructing explanations of how a system comes to exercise particular psychological and behavioral capacities. (2011b, 247)

The key shift in emphasis here is from requiring categories to play a role in law-like empirical generalizations to requiring that they play a role in well-supported models. This shift in what constitutes a kind meshes particularly well with the recent emphasis on mechanistic and model-based explanation, as opposed to nomic or law-based explanation in the sciences (2011b, 251–52).

It seems clear, then, that the goal of new functionalism is to preserve the taxonomic autonomy of the special sciences by stipulating the conditions under which functionally individuated states and processes can legitimately figure into explanations. Though, it is somewhat ironic that Weiskopf emphasizes the alliance between mechanistic and model-based approaches in their break from nomic and law-based explanations for, as we show in the next section, it is the mechanist

who poses a real challenge to the explanatory legitimacy of functional kinds. Below we say a bit more about how Weiskopf envisions the individuation of functional kinds and present Buckner's (2015) dilemma in response.

3 The mechanistic challenge to new functionalism

In contrast to the reductionists, whom Fodor argued against, recent challenges against functionalism have been raised by non-reductive mechanists. Proponents of causal-mechanistic approaches to explanation tend to agree that explanations need not be reductive to be legitimate (Machamer et al., 2000); moreover, they also tend to agree that the special sciences are not in the business of finding laws, and they accept that the main business of the special sciences is to construct models that aim at explaining their target phenomena. Yet, they argue that the way in which models in the special sciences achieve their explanatory aim is, ultimately, by offering mechanistic decompositions (Bechtel & Richardson, 2010). Mechanistic decomposition enables special scientists to see how some target phenomenon can be decomposed into different parts and thus to see how these parts interact — i.e., how operations can be *localized* to *working parts* — in order to produce the capacities of the target that we are interested in.

But what is a mechanism? Glennan, for example, defines a mechanism as follows: "A mechanism for a behavior is a complex system that produces that behavior by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations" (2002, S344). Similarly, Bechtel and Abrahamsen define mechanisms as structures producing their effects "in virtue of [their] component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena" (2005, 423).[5]

Building on such definitions, Craver (2006, 2007; Piccinini & Craver, 2011) suggests that explanations employing functional kinds are explanatory only insofar as they provide *mechanism sketches*. Mechanism sketches provide incomplete albeit informative descriptions of how components and operations of a potential mechanism are responsible for bringing about a certain phenomenon. To put this into perspective, proponents of the mechanistic approach to explanation might say that it is permissible to employ the functional kind *neurotransmitter* to describe what a particular molecule in the human brain does provided that neuroscientists have not *yet* been able to further specify the (relevant) parts of the mechanisms that involve the molecule. By further specifying what happens with(in) the molecule, i.e., localizing its operations to component parts, they can then begin to fill in the explanatory gaps and get a better understanding of why the system behaves in the way it does. Hence, according to what we here are calling arguments from mechanistic decomposition

[5] For a more comprehensive analysis of different accounts of mechanistic explanation in cognitive science and the social and behavioral sciences, see Illari and Williamson (2012).

(AMD), functionally individuated states and processes can only explain by virtue of offering mechanism sketches, otherwise they don't explain at all.

In contrast to AMD, Weiskopf's account (2011a, 2011b) elevates the status of certain functional properties above their role as mere mechanism sketches—it maintains that such functional kinds do not need to be decomposable or localizable to parts of an underlying mechanism or collection of mechanisms in order to be explanatory (e.g., such as when we localize what happens within a molecule that fits the functional description of being a neurotransmitter). More specifically, Weiskopf argues that many functionally individuated states and processes that figure into coherently integrated and empirically successful models do not lend themselves to mechanistic decomposition. This is meant to block AMD for those functional states and processes that qualify as legitimate under new functionalism.

3.1 Individuating functional kinds—three strategies

In contributing to the debate between new functionalism and the mechanistic approach to explanation, Buckner (2015) considers Weiskopf's three strategies for individuating functional kinds.[6] We briefly revisit each below:

> **Fictionalization** is the process of "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" (Weiskopf, 2011a, 331).

To demonstrate that functional kinds can be individuated via fictionalization, Weiskopf discusses the case of "fast enabling links" (FELs) in the brain. FELs are thought to play an essential role in neuroscientific models of vision by depicting the synchronization of distant neural regions in the brain. Weiskopf argues that although FELs possess "physically impossible characteristics" (331), they provide a description of the process by which the binding of intermediary visual representations takes place, a process critically involved in the categorization of spatial objects. Important to this understanding of FELs as *fictions* is that they describe a process that is not carried out by any specific biological mechanisms currently known to neuroscience. Yet, their explanatory force rests in the fact that they are not merely convenient labels for an unknown process but are a necessary part of models explaining certain aspects of visual processing.

> **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" (Weiskopf, 2011a, 328).

Reification, like fictionalization, also involves positing descriptions of phenomena whose actual components or structures may be quite different from modeled systems (construed mechanistically). To illustrate this, Weiskopf turns to one of the

[6] Buckner refers to these individuation strategies as "etiologies" (2015, 3917).

core concepts of cognitive science, *representation*. Understood as the vehicles of content for classical computational systems, representations exemplify the strategy of reification in that they are commonly posited as static and stable entities when, in fact, the process of representation depends upon multiple operations that are inextricably linked at the level of neural realization, such as the dynamics of spike trains, excitatory and inhibitory potentials, and other electro-chemical events. So, like fictionalization, reifications do not directly correspond to any known mechanistic components of the system.[7]

> **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" (Weiskopf, 2011a, 329).

Abstractions are perhaps the most common method of functional individuation — they decompose modeled systems on the basis of *what they do*, not on how they are actually structured. Simply put, the strategy of abstraction applies to any system that instantiates functions, and hence functional kinds, that are not highly localized, but whose roles capture the essential operations of a system. Invariably, this involves discarding or ignoring details of the modeled system in favor of "coarse-grained" or "black-box" descriptions (2011b, 329). Weiskopf characterizes the power of functional abstraction by again appealing to the neural basis of object recognition in vision, suggesting that the act of parsing neural processing into "layer-like" stages helps vision scientists gain epistemic traction on processes whose actual organizing structures possess greater internal organization which may not be directly relevant to the primary function of interest. However, unlike fictionalization and reification, a description is deemed functionally abstract when it describes subcapacities that the depicted systems *really possess* (meaning that nothing new is being posited—they are literal subcapacities of the system). Moreover, these subcapacities are (in principle) implemented, or implementable, by different parts of the system depicted, but cannot possibly, or can only with difficulty, be localized to particular parts of an underlying mechanism.

In response, Bucker argues that none of these three strategies can ensure the explanatory legitimacy of functional kinds. In rebutting the method of fictionalization, Buckner argues that because cases of fictionalization run the risk of losing counterfactual power, this method of individuation is not a viable justification for functional kinds. In particular, his claim is that although fictionalization might help us to account for certain aspects of a phenomenon, it does so "at the cost of a diminished ability to predict and explain another—namely, the aspect that is fictionalized" (2015, 3927). In other words, the counterfactuals implied by the fictionalized components are likely to be false. Bucker takes this to be a good reason to think that the kinds picked out by fictionalization do not meet the normative constraints of good explanation.

[7] But unlike fictionalization, reification makes use of our knowledge of the organization of the underlying components (those that can be known). It is these components that determine the characteristics of the functional capacity being reified. This is what distinguishes FELs from neural representations. In light of this, it might be interesting to ask whether this distinction can be understood in purely epistemic terms.

In rebutting the method of reification, Buckner distinguishes two forms of reification, which he terms "fissional" and "fusional" reification (2015, 3929–3232). As the names suggest, fissional and fusional reification refer, respectively, to cases in which functional kinds are individuated via partitioning component operations (fission) or via aggregating component operations (fusion). In fissional reification, two or more distinct components are introduced whose causal capacities are actually possessed by the same underlying part of the system (or the system as a whole), whereas in fusional reification, one introduces a component whose causal capacities are actually distributed amongst distinct parts of the system. Buckner surmises that attributing kindhood to neural representations is an act of fissional reification since neural representations do not refer to discrete components of a discrete mechanism but instead depict a process by which encoded information enables perceptual inference. Against fissional reification, he invokes what he calls the "A without B" challenge. In brief, the challenge is as follows: "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" (2015, 3929). Hence, fissional reifications, like fictions, are not explanatory because they lead to losses in counterfactual power. He likewise suggests that similar concerns about counterfactual power can be put forward against fusional reifications (though, he also suggests that fusional reification, if it doesn't suffer from counterfactual power, ought to be seen as just another form of abstraction).

Finally, in addressing abstractions, Buckner considers separate reasons why one might appeal to the explanatoriness of abstractions. The first consideration is a metaphysical concern that functional kinds are, or seem to be, unlocalizable *in principle*. This consideration is often paired with arguments suggesting that functional abstractions are serviceable for the purposes of dealing with emergent properties and events which are ontologically irreducible to base components. Buckner's reaction to this consideration is surprisingly simple: there's little reason to think that functional kinds in cognitive science are, in principle, 'metaphysically' complex and thereby resistant to mechanistic decomposition. This is because, once we rid ourselves of models that involve fictions and reifications, as he argues we ought to do, there's no reason to think that anything complex remains.

The second consideration is that, even if functional kinds, construed as abstractions, are localizable in principle, they nevertheless have unique *epistemic utility*. This consideration is often paired with arguments suggesting that the complexity of certain systems may render attempts at decomposition too costly. In this regard, Buckner argues that whereas fictionalization and reification both involve unfortunate tradeoffs in counterfactual power, abstraction *could* count as a legitimate reason to posit functional kinds. However, to be legitimate, the abstraction must be interpreted as a sketch that could be elaborated into a more complete mechanistic model. In other words, abstraction, unlike fictionalization and reification, only counts as legitimate when it is known to pick out underlying mechanisms. Hence, Bucker affirms the view that functional models are ultimately just mechanism sketches. We can thus distill Buckner's worries about new functionalism to the following:

Buckner's dilemma. Special science models that employ functional kinds *either* suffer from a loss of counterfactual power (as compared to mechanistic explanations), *or* they must be interpreted as mechanism sketches that should ideally be elaborated into full mechanistic explanations.[8]

3.2 Deflecting Buckner's dilemma

Buckner's analysis raises a number of important issues for Weiskopf's account, many of which we find compelling, e.g., we agree that functional kinds—understood as fictions and/or reifications—can, in some contexts, lead to considerable losses in counterfactual power; moreover, we agree that functional kinds—understood as abstractions—can aid in mechanism discovery (for further clarification on the heuristic value of mechanism sketches, see, e.g., Craver, 2006; Piccinini & Craver, 2011). Yet, we find two short-comings with Buckner's argument:

First, Buckner's criticisms of functional kinds are limited to models in biology and cognitive science. They are, therefore, far from grounding the general skepticism about functional kinds that he takes them to establish. This motivates our turn to functional kinds that are employed in models in the social and behavioral sciences.[9] Second, Buckner's dilemma presumes that model-based functionalism needs to block or avoid AMD otherwise functional kinds fail to be explanatory. However, this charge overlooks the fact that functional kinds can be employed for explanatory purposes for which offering mechanism sketches seems inappropriate. That is to say, even if individuation by fictionalization or reification leads to losses in counterfactual power and individuation by abstraction suggests that mechanistic decomposition is possible, this doesn't entail that models which employ functional kinds don't satisfy "norms of good explanation" (Buckner, 2015, 3922). As we will now argue, this is because functional kinds may be invoked for many different explanatory purposes—these different purposes are, so far, not taken seriously by either Buckner or Weiskopf.

In the next section, we present and defend two examples from the social and behavioral sciences which reveal that models involving functional kinds aim at a diverse set of explanatory purposes. Not only do we take this to present a challenge to Buckner's dilemma, but we also note that, for each purpose, the associated kind's

[8] To be clear, Buckner's dilemma is meant to apply at the level of individual functional states or processes and not at the level of the model itself. For instance, a model could posit multiple functionally individuated states and processes (e.g., abstractions and fictions) and so could in principle be plagued by both horns of the dilemma simultaneously. In such cases, the strategy for deflecting the dilemma — which we propose in section 4— will need to be investigated for each of those states and processes. While it is, of course, an interesting question how we should proceed *if* deflecting the dilemma only works for some (but not all) posited functional kinds, we ultimately think this needs to be considered on a case-by-case basis.

[9] To be fair to Buckner, we recognize that Weiskopf's original defense of model-based functionalism was primarily aimed at promoting functional explanations in the cognitive sciences, not necessarily the social and behavioral sciences. It is thus an open question whether Buckner's dilemma applies differently across disciplinary boundaries. For now, we simply wish to point out that, in light of certain explanatory purposes in the social and behavioral sciences—which we discuss in section 4 and 5—Buckner's dilemma can be deflected. While we remain optimistic that this can also be done for different explanatory purposes in the cognitive sciences, further investigation is required.

explanatory role is determined *independently* of how it can be individuated according to any of Weiskopf's three individuation strategies. For reasons discussed in section 5, we think this points to an improvement for model-based functionalism.

4 Unlocalizable functional entities in the social and behavioral sciences

To motivate the importance of a model's purpose for assessing the explanatory legitimacy of particular functional kinds, this section examines the use of *preferences* and *signals* in the social and behavioral sciences. We suggest that each are candidate functional kinds. Moreover, we illustrate the explanatory purposes of the models in which these functional kinds *are* frequently employed (this is a descriptive claim). We then show that different instances of signals and preferences are amenable to different parsings according to Weiskopf's taxonomy. Finally, we show how being employed in the context of the explanatory purposes of their models allows these functional kinds to evade both horns of Buckner's dilemma. That is, they evade the dilemma independently of whether we understand them as abstractions, reifications, or fictions.

4.1 The case of preferences

One class of explanatory projects that seems to be predestined for involving functional kinds are those which aim at comparing various heterogeneous systems or studying the interplay of those systems. We take the application of choice theory to the behavior of various entities, e.g., firms, households, humans, and animals, to be a prime example of this. *Preferences* are a fundamental concept in choice theoretic disciplines like microeconomics and game theory. In this regard, preferences are frequently employed for the purpose of explaining via uncovering the same patterns across mechanistically heterogeneous systems. In this subsection we will argue that this can allow them to evade Buckner's dilemma.

4.1.1 How do preferences explain?

Let us illustrate preferences in more detail. While economics is very explicit about the structural assumptions that preferences are supposed to satisfy, there is virtually no explicit definition of the concept in economic textbooks. Nevertheless, substantial views about what preferences in economics are meant to refer to usually identify them as functional kinds. This holds true for most prominent accounts of preferences, of which there are basically three. Mentalists (Dietrich & List, 2016; Okasha, 2016) hold that we should view preferences as referring to mental states. As such, proponents of mentalism typically defend functionalism about mental states, often in opposition to reductive, neurocentric interpretations of mental states. In contrast, the most common non-mentalistic position is behaviorism (Gul & Pesendorfer, 2008). Behaviorism is sometimes characterized as holding that preferences in economics are merely short-hand descriptions for summarizing patterns in choice behavior (Clarke, 2016). In other words, preferences are just the

choices that agents make. Yet, even in this case, what counts as a choice is arguably individuated on functional grounds (Clarke, 2020). There are also more subtle positions that sometimes get classified as behavioristic. For instance, Ross has argued for interpreting preferences as "real patterns" (Ross, 2005, 2006). Following Dennett, he makes the case that those patterns have a special ontological status in virtue of how we attribute beliefs and intentions to the agents who exhibit them. Like mentalism, this view also resists reductive interpretations of preferences; but unlike mentalism, it takes preferences to be contingent upon the actual behavioral profile of agents, which is bound by various sociological and institutional constraints. Finally, dispositionalism (Guala, 2019) maintains that preferences are more adequately viewed as multiply realizable and information-dependent choice-dispositions. What is important for dispositionalism is that the causal base of a preference can be realized in multiple and diverse ways, i.e., different sets of causal properties can give rise to the same set of choice-dispositions (cf. Prior et al., 1982). Hence, entities with quite different causal properties can all exhibit the same preferences. So, under dispositionalism, preferences are also functionally individuated entities.[10]

One might then ask, why are preferences in choice theory typically viewed as functionally individuated entities? First, choice theoretic models involving preferences are used to describe the behaviors of very different entities (Herfeld, 2018; Guala, 2019). For instance, there are choice theoretic models of firms and households (Mas-Colell et al., 1995), mice (Holm et al., 2007), and even hermit crabs (Elwood & Appel, 2009). To give a simple example, two firms can share the same demand curve for copper even though the mechanisms in virtue of which they make their decisions differ substantially. Similarly, if food options require work to be obtained, preferences can also enable us to compare the demand curves of two different species of animals for those food options (cf. Elwood & Appel, 2009; Holm et al., 2007).

What is more, not only do choice theoretic models allow us to spot commonalities between different entities, but they also enable us to explain what would result from the interaction of those entities. Just consider that various types of households regularly interact with various types of firms in basic economic models (Mas-Colell et al., 1995). One could, of course, try to explain the interaction of these different systems by decomposing the mechanism by which they make their decisions and interact. However, this would likely involve the study of quite heterogeneous mechanisms, invoking multiple scientific disciplines. It would, thereby, miss that the interaction of quite dissimilar entities is often driven by the same or similar principles, such as information asymmetries or loss aversion (Herfeld, 2018; Guala, 2019).

To give a concrete but simple example of how economists explain with preferences, consider the prisoner's dilemma. In this game, when both players decide to cooperate (c), they both receive the reward R for cooperating. If they both decide to defect (d), they both receive punishment P. If one of them cooperates, while the other one defects, the cooperator receives the sucker payoff S and the defector the

[10] Even though there are important on-going debates about the proper interpretation of preferences—see, e.g., Dowding (2002), Camerer (2008), Hausman (2008, 2012), Guala (2012), Thoma (2020), Vredenburgh (2020), and Clarke (2020)—we bracket these debates for the sake of space. What is important for us is that preferences in choice theory are broadly and uncontroversially construed as functionally individuated entities.

Table 1 The prisoner's dilemma

	P2		
P1		c	d
	c	R/R	S/T
	d	T/S	P/P

P1 denotes player 1, P2 denotes player 2, c and d stand for the actions the players can perform. The table denotes the outcomes that result from the combinations of those actions

temptation payoff T. Each of the two players *prefers* T > R > P > S (see Table 1 for an illustration of the prisoner's dilemma).

With the help of this information, we can already offer an explanation for why both agents will play 'defect'. Each of the agent's prefers T > R. Consequently, if they believed that the other player would play 'cooperate', they would play defect. Similarly, both players prefer P > S. Hence, if they believed that the other player would play 'defect', they would play 'defect'. So, independently of what they believe the other player will do, they will always choose to defect. In a prisoner's dilemma, we can simply refer to an agent's preference structure to explain why both agents will choose to defect.

What is important to note about this explanation is that it does not matter how exactly the agent's preferences are constituted, e.g., what their causal base is. That's not to say that the actual implementation of the agent's preference structure doesn't depend on facts about the causal composition of their base or realizer; rather, we don't need to know what comprises them in order to explain *why* the players choose to defect (when they do). The 'why' question being answered here concerns a behavioral-level phenomenon, on which one can provide explanations for the behavior of humans as well as non-human agents. Moreover, many of the entities to which economists apply choice theoretic models will likely turn out to make their choices in virtue of quite different mechanisms. Consequently, modeling them with the help of the toolbox of preferences (and beliefs) allows one to spot and explain patterns that would have otherwise been likely missed. For this reason, we take the use of preferences in choice theory to indicate a class of explanatory purposes that can be facilitated by involving functional kinds in one's models but are unlikely to be accomplished by purely mechanistic models: *Functional kinds enable comparisons between mechanistically heterogeneous systems and help to account for the interaction of such systems.*[11]

[11] To be clear, what we say here about rational choice models involving preferences is not new. For example, Ross (2005, 2014; see also Ross & Spurrett, 2004) argues that non-reductive preference-based models in economics and social science can and do provide rich causal explanations using modeling tools like utility functions, budget constraints, and relative prices. These causal explanations are not amenable to mechanistic decomposition (in Buckner's sense of the term) precisely because they emerge from real behavioral patterns. While we are sympathetic to Ross' argument, our aim here is slightly more general. In particular, we think that one can endorse the view that functional kinds enable comparisons between mechanistically heterogeneous systems without assenting to the view that preferences are real patterns. Moreover, our aim in defending model-based functionalism is meant to go beyond the example of preferences discussed here.

4.1.2 Individuating preferences

From the perspective of the debate between Weiskopf and Buckner, it would seem natural to ask how, exactly, preferences are to be individuated in order to assess whether we should really grant them the explanatory power that choice theorists take them to have. For Weiskopf, this matters because defending the legitimacy of functional kinds against the mechanist depends on their being countenanced by one of his three individuation strategies. Whereas, for Buckner, this matters because the efficacy of his dilemma depends on showing that each individuation strategy fails to justify functional kinds figuring in special science explanations. The issue we want to raise here is that it often has no bearing on the explanatory power of preferences which of the three strategies one happens to accept.

Preferences-as-abstractions Preferences could be thought of as abstractions if it turns out that one and the same mechanism is always responsible for a certain choice, e.g., a over b. If it would be the case that every time the agent is presented with a choice between two options, *a* and *b*, the same causal pathway would lead the agent to choose *a*, then the preference '*a over b*' would be an abstraction insofar as it provides a "coarse-grained" or "black-boxed" description of the relevant mechanisms. Following Bucker's line of argument, preferences construed as abstractions could only be seen as legitimately explanatory in cases where they provide mechanism sketches that could (and ideally should) be elaborated into more complete mechanistic models.

However, we take it that this judgment cannot be easily maintained once we take into account the concrete explanatory purposes in which preferences are usually involved. In following Guala (2019), we have to distinguish between *explaining preferences* and *explaining with preferences*. In the former case, one tries to explain the details why a particular agent has the preferences she exhibits. This explanatory project could, for instance, be pursued by decomposing a preference that counts as an abstraction into the individual parts that comprise the underlying mechanisms (this is one of the goals of neuroeconomics—Camerer et al., 2005). Yet, when it comes to explaining with preferences, economists are usually not interested in explaining why a particular agent has the particular preferences she has; instead, they are interested in, among other things, modeling the interaction of different types of agents in a common framework and/or explaining why the same behavioral patterns are found across a range of different types of agent. For instance, Herfeld (2018) outlines how Aklerlof's (1970) market-for-lemons model allows us to see that information asymmetries can account for the same behavioral patterns occurring across a wide range of settings involving agents that differ significantly when it comes to the mechanisms by which they make their choices (e.g., agents in the used-car market in Zurich and those that buy second-hand goods in the Ecuadorian rainforest).

Given that explaining preferences and explaining *with* preferences are different explanatory projects, the claim that preferences, as abstractions, are just mechanism sketches misses the mark. There are two reasons for this. First, providing additional details on the causal compositions of an agent's preferences does not necessarily improve the explanation we are pursuing when explaining *with* preferences. For instance, if we already know how an agent will behave in a prisoner's dilemma-type situation given their preferences, information about the neural basis of those

preferences will not have added value for the relevant explanation. Second, mechanistic decomposition may even threaten to complicate achieving the aim we are after when explaining *with* preferences. This is because we risk losing the abilities to facilitate comparisons between mechanistically heterogeneous systems and to account for the interaction of such systems. For example, explaining the interaction of the agents in the prisoner's dilemma purely in mechanistic terms may require gathering a lot of details about the potentially different causal compositions of the agents' preferences.[12] If doing all this work does not improve the explanation we already get from knowing their preference orderings, mechanistic decomposition becomes a futile task. Hence, claiming that preferences are mechanism sketches that *should* be expanded upon seems highly misleading. Consequently, the horn of Buckner's dilemma that applies to abstractions turns out to be quite dull if we consider preferences in the context of the explanatory purposes of the models in which they are employed.

Preferences-as-reifications* and *preferences-as-fictions On the one hand, it is plausible to view some preferences as reifications. This can, for instance, be the case when it turns out that it is not always the same mechanism, or mechanisms, responsible for a choice between two options. For example, imagine that an agent in the morning could choose '*coffee over tea*' because she needs the caffeine kick to start her work. However, the same choice between *tea* and *coffee* in the afternoon may be influenced by the fact that the agent desires the taste of coffee more than that of tea while having a break. For many explanatory projects in choice theory and microeconomics, all that matters is that the agent has a stable preference for coffee over tea; it matters little that different mechanisms are at work in realizing the same preference in dissimilar contexts.

On the other hand, certain preferences must be construed as fictions. For instance, consider the assumption of *continuous preferences* that is frequently employed in economic models. The assumption states that whenever an agent prefers A to B to C, there is a probability p such that the agent is indifferent between $p*A+(1-p)*C$ and B. Roughly, this assumption can be understood as agents having infinitely fine-grained preferences. For example, not only do agents like 2 apples more than 1, if they have continuous preferences over apples, they also prefer having 33.33333333% of an apple over 33.33333332% of the same apple. As a result of this, some preferences that certain models ascribe to the agents might be too fine grained to be psychologically plausible. Yet, continuity often facilitates the derivation of the results of a certain model (cf. Reiss, 2012). Hence, economists frequently put components into their models that are known not to correspond to any element of the modeled system but are nevertheless important for the model to function.

This brings us to the horn of Buckner's dilemma dealing with reifications and fictions. Could he still claim that if preferences are construed as reifications or fictions,

[12] As we mentioned already, arguing that preferences are merely mechanism sketches overlooks the fact that preferences can be multiply realized. For this reason, we are very sympathetic to Ross (2014, 218-28) who criticizes the neuroeconomics program defended by Camerer et al. (2005), known as "behavioral economics in the scanner", which attempts to reduce agents' preferences to neural processes and mechanisms. This program epitomizes the kinds of reductive approaches to explaining preferences that we here are trying to avoid by emphasizing that some explanatory purposes call for explaining *with* preferences.

they fail to explain? Let us first consider reifications. As with representations in cognitive science, it is plausible to expect that some preferences and beliefs in choice theory will turn out to be fissional reifications. Buckner's main worry here is that fissional reifications involve a loss of counterfactual power as the causal capacities attributed to beliefs and preferences, respectively, are actually possessed by the same underlying component of the system in question (recall the "A without B" challenge). Hence, construing preferences as distinct entities can lead to false counterfactual statements—that is, it can lead us to make false statements about what will happen when we change an agent's beliefs because we mistakenly assume that this can be done while holding her preferences constant.

However, what becomes relevant once we take the explanatory purposes for which preferences are usually employed into account, is whether this particular loss of counterfactual power matters in the context of those purposes. For instance, it is not that clear that by assigning a particular set of preferences to an agent (e.g., $T > R > P > S$) we also make claims about the plasticity of those preferences. That is, while choice theoretic models can explain by showing how particular patterns in behavior result from different types of agents having certain preferences in often highly different circumstances, it is far from obvious that they, thereby, are also meant to serve the purpose of informing us about what would happen if an agent's beliefs or preferences were changed (cf. Clarke, 2020). Hence, the loss of counterfactual power that comes with reification may not affect many of the models in which preferences are frequently employed given their explanatory purpose.

Finally let us consider fictions. Above we said that at least some preferences, viz. those that seem to be too fine-grained to be psychologically plausible and applicable to ordinary choice behavior, are fictions. Could Buckner at least maintain that preferences, construed in this way, are problematic even if we account for the explanatory purposes for which they are employed? The short answer is no. While some preferences are clearly fictions, e.g., those that we assign on the basis of continuity assumptions, we rarely find choice theoretic models in which all preferences are fictions. Moreover, even if preferences are introduced as fictions, this is typically for the purposes of mathematical convenience. This is not to say that idealizations, such as assuming continuity, always yield explanatory power in choice theoretic models. It is just to say that the fictions that we introduce as a result of those idealizations do not introduce additional worries beyond those about idealizations that one may already have (cf. Reiss, 2012; Hausman, 2013). Whether we accept the resulting loss of counterfactual power as a price for the mathematical convenience that is gained by introducing preferences as fictions into our models, and how this tradeoff bears on the explanatoriness of the relevant model, is a question that needs to be assessed on a case-by-case basis. Importantly, for the debate at hand, this means that we cannot simply undermine the legitimacy of particular choice theoretic models by pointing out that some of the preferences they assume agents to have are ultimately fictions.

All in all, it seems that regardless of whether preferences are construed as abstractions, reifications, or fictions, neither of the horns of Buckner's Dilemma seem to affect them in the context of the explanatory purpose of many choice theoretic models in which they are employed. We take the upshot of all of this to be that many

special science models that aim at showing commonalities between mechanistically heterogeneous entities—and which try to explain the interaction of those entities based on such commonalities—stand to incur great benefits and no harmful side-effects from employing functional kinds in their models.

4.2 The case of signals

Often functional kinds are also involved in models whose explanatory purposes are intricately linked to the presence of functional kinds—this is because the properties that make these models explanatory in the light of the model's purpose would disappear at the mechanistic level. This can be true even for cases where decomposition is quite easy and also for those cases where we are not interested in detecting patterns across different systems or studying their interaction. To illustrate this kind of explanatory project, we consider the case of signals in the Lewis-Skyrms approach to the study of meaning. This approach investigates how meaning emerges and how different signals acquire their meaning. It thereby relies heavily on game-theoretic models (see, e.g., Lewis, 1969, Skyrms, 2010). Central to this approach is the notion of signaling games. A signaling game is a coordination game in which one agent (the sender) sends a signal to another agent (the receiver), who then performs an action depending on the signal. A signal can be anything that serves the function of inciting a certain response in the receiver. In a signaling game, agents try to realize mutual benefit. In order to realize this benefit, the receiver has to perform the correct action, whereby the correctness of an action depends on the state of the world which only the sender can observe. The main idea behind the Lewis-Skyrms approach is that the meaning of a signal is constituted by the *function* the signal plays in a signaling system (see, e.g., Lewis, 1969; Godfrey-Smith, 2017; Harms, 2004). Hence, we take this approach to be a promising case study in the context of model-based functionalism.

4.2.1 How do signals explain?

Before we will illustrate how the Lewis-Skyrms approach explains meaning, we will briefly expand on some of the details of the approach. A signaling system consists of a set of contingency plans, i.e., the contingency plans of the sender(s) and receiver(s). The contingency plan of the sender(s) is a mapping from states of the world to signals. The contingency plan of the receiver is a mapping from signals to actions. In this regard, we can think of contingency plans as behavioral dispositions of the sender and the receiver.

Following Lewis (1969), we only speak of a signaling system if the contingency plans of the speaker and the receiver combine in such a way that they constitute a coordination equilibrium in a coordination game. More precisely, if contingency plans are plans of actions over all possible situations—i.e., what game theorists refer to as *strategies*—they constitute a Nash equilibrium *iff* no agent could do strictly

better by changing her contingency plans unilaterally. A coordination equilibrium is a Nash equilibrium in which every player prefers that everyone else conforms to a set of contingency plans if at least all but one conforms.

For these reasons, according to the Lewis-Skyrms approach, the meaning which the signal is carrying, i.e., what kind of signal it is, is determined by its functional properties. More precisely, it is determined by the role it plays in the signaling system. Accordingly, explaining meaning requires us to look at signals at the functional level (see Table 2 for an illustration of two signaling systems).

To illustrate how this approach intends to explain meaning, consider one of Lewis' (1969) favorite examples. Lewis envisioned a signaling game in which two agents have to exchange messages about the onslaught of the British on the continental army. The Sexton of the old new church (the sender) has knowledge about the British, and Paul Revere (the receiver) is in a position to warn the continental army. Both aim at giving the correct warning to the continental army, i.e., they give no warning if the British stay at home, they give warning that the British are coming by sea if the British are coming by sea, and so on. Yet, the Sexton is only able to signal the state of the world to Paul Revere by hanging lanterns into the window of the church. Hence, they aim at matching their respective contingency plans in such a way that they always reach the best possible outcome. For example, the Sexton can choose to hang no lantern in the window if the British are staying at home, one lantern if they are coming by land, and so on. Paul Revere can choose to give no warning if there is no lantern, warn that they are coming by sea if there is one lantern, and so on. Now imagine that Paul and the Sexton have to repeatedly engage in the described situation. Once the actual contingency plans of both agents match in such a way that they form a signaling system (i.e., that they will always give the correct warning), the resulting conventional solution to the coordination problem constitutes the meaning of the signal. In other words, the meaning of the signals (the lanterns in the window) are constituted by their role in such signaling systems. A real-world case that is similar to Lewis' example can be found in the system of warning cries of Campbell's monkeys, who give an alarm cry when they spot a predator (Zollman, 2011). According to the Lewis-Skyrms model, the meaning of these warning cries is determined by the function they play in the monkey's warning system (see Table 3 for an illustration of Lewis' lantern game).

Table 2 Two signaling systems

Sender	Receiver
S1 ⇒ *SigA*	SigA ⇒ *A*1
S2 ⇒ *SigB*	SigB ⇒ *A*2
S2 ⇒ *SigA*	SigA ⇒ *A*2
S1 ⇒ *SigB*	SigB ⇒ *A*1

Two signaling systems in a signaling game where successful coordination depends on action A1 being performed if S1 is the state of the world and A2 being performed if S2 is the case. Signals SigA and SigB play different roles in each system. Hence, their meaning differs between the systems (cf. Zollman, 2011, 161)

Table 3 Lewis' lantern game

	C1	C2	C3	…
L1	1/1	0/ 0	0,5/0,5	…
L2	0/ 0	1/1	0,5/0,5	…
L3	0,5/0,5	0,5/0,5	1/1	…
⋮	⋮	⋮	⋮	

C1, C2, C3 … denote the contingency plans of Paul. L1, L2, L3 … denote the contingency plans of the sexton. Players get payoffs 1/1 if their contingency plans form a signaling system; 0,5/0,5 if their contingency plans match such that they lead sometimes to the correct outcome; 0/0 if they never lead to the correct outcome (see Lewis, 1969, 124)

Many of Lewis' successors argued for different equilibrium concepts that can underlie a signaling system and investigated the conditions under which such systems exhibit certain stability properties (see, e.g., Sugden, 2004). Moreover, others (most notably Skyrms, 2010) have shown how various selection processes—like biological evolution, reinforcement learning, imitation of success, and rational choice—can shape agents' contingency plans in such a way that they can constitute a signaling system. Despite all of this, the main message of the Lewis-Skyrms approach to the study of meaning is that meaning is explained by a signal's function in a signaling system. We take this to indicate a second class of explanatory purposes for models involving functional kinds: *Functional kinds enable the tracking of phenomena that aren't visible at the level of the components of the relevant mechanisms.*[13]

4.2.2 Individuating signals

Let's again consider what Weiskopf's individuation strategies can tell us about signals. We take it that one can view signals as either abstractions or reifications.[14]

Signals-as-abstractions To return to Lewis' example, we could interpret signals as abstractions of the lantern mechanism. For instance, the signal that the British are coming by sea, could be broken down into the Sexton's act of putting two lanterns into the window and Paul receiving a certain visual input, which, in turn, leads him to send a certain warning (e.g., the British are coming by sea). We can interpret this whole process as a signal. This would offer a very coarse-grained description of the relevant mechanisms. Moreover, the whole process is only individuated as a

[13] Unlike preferences, the case of signals demonstrates that functional kinds do more than facilitate cross-system comparisons and explain interactions; it indicates how functional kinds can provide knowledge and understanding of patterns which *cannot* be tracked, and therefore not explained, via mechanistic decomposition—this holds even for individual systems. However, we think this doesn't rule out that mechanistic intervention through decomposition can help to explain other aspects of the system, for example, when the systems in question break down or fail to function. But this would complement, not replace, either of the here mentioned explanatory purposes.

[14] It seems unlikely that signals could be construed as fictions, so we don't consider this strategy.

signal because of its function. Understood like this, signals seem to fit the notion of abstraction that Buckner and Weiskopf rely upon.

How would Bruckner's dilemma affect this interpretation? It seems obvious that arguing that signals, as abstractions, would still be amenable to mechanistic decomposition would not undermine their explanatory legitimacy in the context of the Lewis-Skyrms approach. The reason is simply that decomposing the relevant signals would lead us away from the functional level at which meaning is ultimately to be explained. Once we have specified the role of a signal in a signaling system, further details about the underlying mechanisms can do little to explain the signal's meaning. Hence, we hold that signals as abstractions do not succumb to the charge of being mere mechanism sketches.

Signals-as-reifications Consider that it may not only be necessary that Paul receives a certain visual input in order for him to send a certain warning, but it may also be required that Paul classifies this input the right way. If Paul would suffer from occasional hallucinations that lead him to see additional spots of light in the church window, the signals in Lewis's example could not function in the relevant signaling system as they are intended to do. Hence, in order for the relevant signals to occupy their specific role in the signaling system, it seems to be important that Paul's brain mechanisms responsible for classifying visual inputs work in a certain (perceptually correct) way. Therefore, it could also be argued that the signals in Lewis's example are reifications that fuse the lantern mechanisms with Paul's brain mechanisms.

How then would the horn of the dilemma that is supposed to apply to reifications affect signals under the Lewis-Skyrms approach? As illustrated above, according to the Lewis-Skyrms approach, meaning is located at the functional level of signaling systems. Any attempt to decompose the signal into its constituent parts would, therefore, be unable to explain why a particular signal has its particular meaning. In other words, under the Lewis-Skyrms approach, functional kinds located at the level of signaling systems are indispensable when it comes to the explanations of meaning. Hence, no matter what general explanatory drawback we identify for reifications—such as losses in counterfactual power—these drawbacks in and of themselves cannot undermine the legitimacy of using signals as functional kinds for explaining meaning because, within the Lewis-Skyrms framework, the meanings that signals encode can only be explained at the functional level.

Given all of this, we hold that arguing about whether the signal is an abstraction, or a reification, would not be very illuminating with respect to the explanatory role that signals play within the Lewis-Skyrms approach because the phenomenon those models intend to account for is simply not visible at the level of the components of the relevant mechanisms. In cases like this, Buckner's dilemma cannot undermine the explanatory legitimacy of the relevant functional kinds.

Against this, however, one might argue that relying on the Lewis-Skyrms approach in the context of new functionalism is misguided as it does not attempt to give a *causal* explanation, but rather tries to explain in what meaning is *grounded*, i.e., the role of a signal in a signaling system. More generally, one could hold that new functionalism should be restricted to causal explanations. If this were correct, it would undermine our signal example and we would instead have to identify an

attempt at giving a causal explanation where mechanistic decomposition would make the phenomena to be explained vanish.

First, we hold that new functionalism should not be restricted in this way. Second, there are also causal explanations that help us to make our case. To see this, just consider that the Sexton in Lewis' example could use different types of 'lanterns' to communicate, e.g., he could use flaming torches or kerosine lamps. Now consider that Paul sees the lights that were produced by the kerosine lamps. Should we say that his warnings were caused by the specific light-emitting technology, *or* say that it was caused by seeing a specific number of lights in the church window? It seems that only the latter statement would be explanatory in this particular context, while the first statement would threaten to obscure the behavioral pattern exhibited by Paul (see also, List & Menzies, 2009). Consequently, we take the Lewis-Skyrms approach to be a straightforward example of a model-based research program involving functional kinds that are unaffected by Buckner's dilemma. Moreover, the worry that the phenomena to be explained will not be visible at the level of the components of mechanisms seems to generalize to models whose purpose it is to offer causal explanations.

5 New functionalism, restated

Given the considerations of the previous section, we can now see that the explanatory purposes in which functional kinds are involved cross-cuts the ways in which they are, or could be, individuated. In some cases, knowing how a functional kind fits into Weiskopf's taxonomy may not even be necessary to defend its legitimacy against Buckner's dilemma once we take its explanatory purposes into account.

This indicates to us that an assessment of the model-based approach to functional explanation requires *more* than an account of how functional kinds are individuated—it needs to also recognize that models utilizing functional kinds often serve quite different explanatory purposes. In particular, we focused on functional kinds that (a) enable comparisons between mechanistically heterogeneous systems and help to account for the interaction of such systems, and those that (b) enable the tracking of phenomena that aren't visible at the level of the components of the relevant mechanisms. In such cases, one cannot simply ask how a particular model involving functional kinds would compare with some mechanistic explanations; one has to ask how it serves the particular explanatory purpose of the model in which it is employed. To put it differently, the legitimacy of a particular functional kind does not only depend on how it is individuated, but also on the explanatory purposes in which the models that employ those functional kinds are involved.

The aim of this final section is to outline how our arguments offer a defense of new functionalism. First, we characterize our arguments in terms of a reformulation of Weiskopf's original conditions for model-based functional explanations. More specifically, we suggest that his third condition, the *genuine explanatoriness* condition, needs to be broadened if we want to evade Buckner's dilemma. Second, we

argue that broadening this condition does not lead to parochialism or anything-goes pluralism.

5.1 Genuine explanatoriness does not depend only on counterfactual power

We have illustrated different classes of explanatory purposes for which functional kinds may be legitimately employed. How does this relate back to Weiskopf's original vision regarding model-based functionalism? Recall that according to his formulation, explanations employing functional kinds have to be (i) *well-confirmed*, (ii) *representationally accurate*, and (iii) *genuinely explanatory*. Nothing in our discussion above shows (i) and (ii) to be problematic.

However, if we take a closer look at how Weiskopf spells out genuine explanatoriness, some may worry that our cases above do not exemplify this condition. Recall that Weiskopf states that condition (iii) entails that a set of models that can, ceteris paribus, answer more 'what-if-things-had-been-different' questions is to be preferred, and further, that we should only consider the functional properties of these models to constitute kinds. For example, one might now be tempted to think that models containing preferences are unable to answer 'what-if-things-had-been-different' questions about the source and structure of those preferences. This, we agree, may happen if one takes preferences to be reifications or fictions.[15] Similarly, one may think that signaling models do not lend themselves to making counterfactual statements about how changing the realizers may affect the meaning of the signals and hence the behavior of the senders/receivers. However, as we have also suggested, this should not be a major worry as demanding that the relevant models answer those kinds of questions would amount to neglecting their specific explanatory purposes.

This brings us back to the second horn of Buckner's dilemma—namely, that models that do not answer as many counterfactual questions as possible fail to be explanatory. The argumentative thrust of this horn comes from the presumption that, ideally, the relevant model should provide as many counterfactual statements about its target as possible—this means that, given two competing models, the one which can provide more answers to counterfactual questions about the target phenomena is the better model. In this regard, Buckner states that:

> We also must be told why the functionalist interpretation of [their] 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*. (2015, 3928)

[15] To be clear, this worry really only applies to *non-behavioristic* interpretations of preferences (e.g., Dietrich & List, 2016; Guala, 2019). But for *behavioristic* interpretations, there should be no worry about the counterfactual power because nothing is being reified or fictionalized--a preference just is a choice. So, we can ask and answer 'what-if-things-had-been-different' questions about preferences understood as choice-behavior (e.g., asking how a change in price will affect a change in choice).

However, our discussion suggests that it should not primarily matter how many counterfactual statements the models in which functional kinds are employed can support, but whether they support the kind of counterfactual statements *that are relevant for the explanatory purpose of the model.* Everything else would amount to comparing apples with oranges. Consequently, while we do not think that Buckner's dilemma is the devastating objection to new functionalism that he takes it to be, it is nonetheless a very useful device for pointing out weaknesses in Weiskopf's specific construal of new functionalism.

We, therefore, hold that genuine explanatoriness does not imply giving as many counterfactual statements as possible; it means answering as many questions as possible that are relevant to the purpose of the model—this may of course include answering some 'what-if things-has-been-different' questions. Hence, we suggest that condition (iii) should be construed as a *purpose-adequacy condition*, meaning that a model's explanatory power should be examined and judged in terms of the model's contribution to answering questions relevant to its explanatory purpose.

5.2 Explanatory monism, pluralism, and holism

We suggested that Weiskopf's condition (iii) for model-based functionalism should be amended so as to allow for different explanatory purposes. As will become clear below, we hold that this adjustment is necessary to account for the trade-off in explanatory purposes one can pursue with any given model. However, our amendment may raise the concern that we are defending a form of 'anything-goes' pluralism about scientific explanation. Buckner (2015) argues that appealing to various explanatory purposes, as we have, can lead to a form of model *parochialism*. The worry is that appealing to various different purposes allows one to evade answering questions about counterfactual power by restricting the aims of the model in an ad hoc manner. Yet, we will now make the case that this worry is exaggerated.

Let us, therefore, briefly expand on the two most common normative positions one can take with respect to the explanatory aims of science. One can be *monist* and hold that there is one central, context-independent aim in science and only models that satisfy this aim are legitimately explanatory. We take it that many mechanists are, at least implicitly, monists. However, as our discussion so far has indicated, we reject this position. By contrast, one can be *pluralist* and hold that there are various purposes a model can pursue to be legitimately explanatory. Most often for the pluralist, pragmatic considerations determine which models are explanatory in a given context (see e.g., Potochnik, 2017). This view is usually rejected by mechanists, like Buckner, as they worry that it makes it too easy for scientists to insulate their models from critique by redefining their explanatory purposes in an ad-hoc manner.

We take the functional kinds and models discussed in section 4 to already suggest that appealing to different explanatory purposes is, by far, not always an ad hoc move. In fact, by looking at the wider modeling literature it is easy to see that model building is context-sensitive in the sense that modelers construct their models with specific purposes in mind. That is, models are never meant to be the kind of universal purpose tools that monists might take them to be. In this regard, Mäki

(2009) argues that models are usually accompanied by a *commentary* that identifies the specific purposes of the model. This stands in stark contrast to Buckner's view that appealing to the aims of a model to defend the inclusion of functional kinds is usually an ad hoc defense. Mäki demonstrates this by appealing to the commentary that accompanies the famous Schelling model (Schelling, 1971), saying it "explicitly states that the rationale of his models is *not* the ambitious one of serving as first approximations that can be elaborated to simulate with higher fidelity the real situations we want to examine" (2009, 38). In other words, the model does not aim at offering a mechanism sketch. In a similar vein, Potochnik (2017, 67) notes with respect to game theoretic models that study the emergence of cooperation in evolutionary biology (e.g., Axelrod & Hamilton, 1981), that it is clear that "their aim is to depict patterns in how natural selection can, in general, causally contribute to the emergence of cooperation. Because these models of cooperation ignore so many other causal influences in order to focus on the role of natural selection, they have a limited range of application and limited accuracy of most any evolved trait." In both of these examples, the commentary accompanying the models clarifies the purpose and scope of the model. In neither case can this additional information be construed as an ad hoc defense of the model's explanatory limits—it is, rather, a specification about how to understand the aims of the model. Consequently, the wider modelling literature suggests that pluralism should be preferred to monism and that pluralism does not need to give way to the ad hoc defenses that Buckner is worried about.

Nevertheless, this may not address the issue of parochialism fully as one can still disagree with the guiding motivations which set a particular modeling agenda. In other words, one may simply be unconvinced that a given explanatory purpose that can be used to legitimize functional kinds is *itself* legitimate. This raises the worry that, if we are correct in our argument so far, the debate about new functionalism is potentially deadlocked due to contrasting commitments about the legitimate aims of science. Consequently, one may worry that our attempt to deflect Buckner's dilemma has not gotten us very far because one could simply respond by insisting on the illegitimacy of explanatory purposes that would allow us to legitimize functional kinds. Even though we cannot fully address this here, we briefly respond to this worry in two ways.

Our first response is that we take the burden of proof to lie with mechanists. This is because they need to argue that explanatory purposes that are commonly pursued in the special sciences are, in fact, illegitimate. However, we anticipate that this will not be enough to persuade the mechanists. Our second response is, therefore, to point at a potential way out of this deadlock. This is the view from *explanatory holism*, which is advantageous to different scientific aims but also blocks anything-goes pluralism or parochialism.

According to Hochstein (2017), explanatory pluralism does not offer a viable alternative to monism because both positions treat the assessment of explanatory legitimacy of models in the abstract—that is, independently of other models with which they may be informed and interconnected. He takes this to be a crucial flaw for both monism and pluralism. On the one hand, he agrees with the wider modeling literature that individual scientific models cannot simultaneously satisfy all the scientific goals typically associated with explanation—that is, "a given model's ability

to satisfy some goals must always come at the expense of satisfying others" (2017, 1105). On the other hand, any model that sacrifices some explanatory goals to attain others will always "necessarily undermine its own explanatory power in the process" from the perspective of the monist. Hochstein's response is that, in assessing the legitimacy of a model's purposes, we must appeal to *collections of models* and assess how much attaining the explanatory purposes of one model can help us to satisfy other explanatory purposes.

Taking inspiration from Hochstein, we now propose that one way to break the deadlock and to block the charge of parochialism is to ask in how far the explanatory purposes we pursue with any given class of models are able to support other explanatory purposes pursued by other classes of models. In asking this question, we are committed to the view that the special sciences are ultimately a cooperative endeavor. Consequently, we take the perspective that a model's explanatory purpose should be evaluated also on the basis of how well it contributes to the broader network of explanatory aims. That is, of course, not to say that any individual model actually has to pursue all of these aims. We agree with Hochstein, and the wider modelling literature, that it cannot. Instead, the relevant issue is whether achieving one explanatory goal can also help us to achieve others. Asking this question introduces a criterion for demarcating explanatory purposes—in particular, it allows us to distinguish purposes which support merely parochial projects from purposes which also contribute to other explanatory aims. For instance, one way to defend the explanatory purposes of, say, Akerlof's market-for-lemons model, against the charge of parochialism would be to show that having knowledge of the general patterns that arise in markets due to information asymmetries makes it easier to search for the particular mechanism that give rise to this pattern in a particular instance (cf. Herfeld, 2018). Consequently, we view explanatory holism as offering a promising way out of this deadlock.

Of course, one may retort that we fall short of showing that the explanatory purposes we have pointed out here always or necessarily contribute to some broader network of explanatory aims. Moreover, it remains an open question whether *all* legitime explanatory purposes contribute to such a network. These two issues relate to a broader concern regarding our appeal to holism, which is that it allows for the possibility that models involving functional kinds will remain explanatorily incomplete unless properly integrated into a larger model-network, which presumably also includes mechanistic models.

On the one hand, this would grant the mechanist room to argue that mechanistic explanations are still required at some level. In this regard, we agree that it is an important question how networks of models can and should come together to provide holistic explanations. We suspect that there is not one type of holistic explanation, but various types, whose compositions may depend on different, local sociological factors. But how to identify these types is a question for another paper. Whether each and every holistic explanation requires mechanistic models is, therefore, an open question. In any case, our argument here suggests that some explanatory purposes can only be achieved by invoking functional kinds (which are not amenable to mechanistic decomposition).

On the other hand, one could argue that if genuine explanations are only offered by networks of models, this would undermine the explanatory purposes mentioned above. In other words, one could hold that only holistic explanations should count as genuine (i.e., capital 'E') explanations. However, while one may hold this view, it would not change any of the substantial points we made in this paper. That is, our strategy for deflecting Buckner's dilemma would not be affected by it.[16] In any case, for now, our aim was just to point out that holism provides one promising strategy for getting out of the deadlock and blocking the charge of parochialism.

6 Conclusion

Our aim in this paper has been (a) to demonstrate that there *are* diverse explanatory aims in the social and behavioral sciences, and (b) that we can appeal to these aims in a non-ad hoc manner which indicates the limitations of Buckner's dilemma for criticizing new functionalism. To this end, we have defended an approach to model-based explanation by proposing a reformulation of Weiskopf's conditions under which functional kinds legitimately explain. We have pointed out that simply disagreeing with the legitimacy of the purposes we introduced would be a potential way to reassert the validity of Buckner's critique; and, that the resulting disagreement about normative commitments concerning the explanatory aims of science could lead to a deadlock. To demonstrate that there is a way out of this deadlock, we sketched how holism can provide us with the basis for demarcating explanatory aims.

Acknowledgments We thank Anna Alexandrova, O. Calgar Dede, Melissa Vergara Fernández, Marta Halina and Marcel Jahn, for very valuable feedback on earlier drafts of this paper.

Funding Lukas Beck is currently doing a PhD that is supported by a doctoral scholarship of the German Academic Scholarship Foundation (Studienstiftung des deutschen Volkes).

Declarations

Disclosures There are no potential conflicts of interest to report.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

[16] We are thankful to an anonymous referee for bringing up these points.

References

Akerlof, G. (1970). The market for 'lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics, 84*(3), 488–500.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science, 211*(4489), 1390–1396.

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

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

Buckner, C. (2015). Functional kinds: A skeptical look. *Synthese, 192*(12), 3915–3942.

Camerer, C. (2008). The case for mindful economics. In A. Caplin & A. Schotter (Eds.), *The foundations of positive and normative economics: A handbook (pp. 43–69)*. Oxford University Press.

Camerer, C., Loewenstein, G., & Prelec, D. (2005). Neuroeconomics: How neuroscience can inform economics. *Journal of Economic Literature, 43*(1), 9–64.

Cartwright, N. (1999). *The dappled world: A study of the boundaries of science*. Cambridge University Press.

Clarke, C. (2016). Preferences and positivist methodology in economics. *Philosophy of Science, 83*(2), 192–212.

Clarke, C. (2020). Functionalism and the role of psychology in economics. *Journal of Economic Methodology*, 1–19.

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

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

Craver, C. F., & Bechtel, W. (2006). Mechanism. In S. Sarkar & J. Pfeifer (Eds.), *Philosophy of science:An encyclopedia* (pp. 469–478). Routledge.

Dennett, D. C. (1989). *The intentional stance*. MIT Press.

Dennett, D. C. (1991). Real patterns. *The Journal of Philosophy, 88*(1), 27–51.

Dietrich, F., & List, C. (2016). Mentalism versus behaviourism in economics: A philosophy-of-science perspective. *Economics & Philosophy, 32*(2), 249–281.

Dowding, K. (2002). Revealed preference and external reference. *Rationality and Society, 14*, 259–284.

Elwood, R. W., & Appel, M. (2009). Pain experience in hermit crabs? *Animal Behaviour, 77*(5), 1243–1246.

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

Fodor, J. (1997). Special sciences: Still autonomous after all these years. *Philosophical Perspectives, 11*, 149–163.

Gierre, R. N. (1988). *Explaining science: A cognitive approach*. University of Chicago Press.

Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science, 69*(S3), S342–S353.

Godfrey-Smith, P. (2017). Senders, receivers, and symbolic artifacts. *Biological Theory, 12*(4), 275–286.

Guala, F. (2012). Are preferences for real? Choice theory, folk psychology, and the hard case for commonsensible realism. In Economics for real (pp. 151-169). Routledge.

Guala, F. (2019). Preferences: Neither behavioural nor mental. *Economics & Philosophy, 35*(3), 383–401.

Gul, F., & Pesendorfer, W. (2008). The case for mindless economics. In A. Caplin & A. Schotter (Eds.), *The foundations of positive and normative economics: A handbook* (pp. 3–42). Oxford University Press.

Harms, W. F. (2004). Primitive content, translation, and the emergence of meaning in animal communication. Evolution of communication systems: A comparative approach, 31–48.

Hausman, D. M. (2008). Mindless or mindful economics: A methodological evaluation. In A. Caplin & A. Schotter (Eds.), *The foundations of positive and normative economics: A handbook* (pp. 125–151). Oxford University Press.

Hausman, D. M. (2012). *Preference, value, choice and welfare*. Cambridge University Press.

Hausman, D. M. (2013). Paradox postponed. *Journal of Economic Methodology, 20*(3), 250–254.

Herfeld, C. (2018). Explaining patterns, not details: Reevaluating rational choice models in light of their explananda. *Journal of Economic Methodology, 25*(2), 179–209.

Hochstein, E. (2017). Why one model is never enough: A defense of explanatory holism. *Biology and Philosophy, 32*(6), 1105–1125.

Holm, L., Ritz, C., & Ladewig, J. (2007). Measuring animal preferences: Shape of double demand curves and the effect of procedure used for varying workloads on their cross-point. *Applied Animal Behaviour Science, 107*(1–2), 133–146.
Illari, P. M., & Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science, 2*(1), 119–135.
Kim, J. (1992). Multiple realization and the metaphysics of reduction. *Philosophy and Phenomenological Research, 52*, 1–26.
Kincaid, H. (1996). *Philosophical foundations of the social sciences*. Cambridge University Press.
Kincaid, H. (2004). Are there Laws in the social sciences? Yes. In C. Hitchcock (Ed.), *Contemporary debates in philosophy of science* (pp. 68–187). Blackwell.
Ladyman, J. (2008). Structural realism and the relationship between the special sciences and physics. *Philosophy of Science, 75*(5), 744–755.
Ladyman, J., Ross, D., Collier, J., Spurrett, D., Spurrett, D., & Collier, J. G. (2007). *Everything must go: Metaphysics naturalized*. Oxford University Press on Demand.
Lewis, D. K. (1969). *Convention: A philosophical study*. Blackwell.
List, C., & Menzies, P. (2009). Nonreductive physicalism and the limits of the exclusion principle. *The Journal of Philosophy, 106*(9), 475–502.
Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*(1), 1–25.
Mäki, U. (2009). MISSing the world. Models as isolations and credible surrogate systems. *Erkenntnis, 70*(1), 29–43.
Mas-Colell, A., Whinston, M. D., & Green, J. R. (1995). Microeconomic theory (Vol. 1). Oxford University Press.
Okasha, S. (2016). On the interpretation of decision theory. *Economics & Philosophy, 32*(3), 409–433.
Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese, 183*(3), 283–311.
Potochnik, A. (2017). *Idealization and the aims of science*. University of Chicago Press.
Prior, E. W., Pargetter, R., & Jackson, F. (1982). Three theses about dispositions. *American Philosophical Quarterly, 19*(3), 251–257.
Reiss, J. (2012). The explanation paradox. *Journal of Economic Methodology, 19*(1), 43–62.
Ross, D. (2005). *Economic theory and cognitive science: Microexplanation*. MIT Press.
Ross, D. (2006). The economic and evolutionary basis of selves. *Cognitive Systems Research, 7*(2–3), 246–258.
Ross, D. (2014). *Philosophy of economics*. Springer.
Ross, D., & Spurrett, D. (2004). What to say to a skeptical metaphysician: A defense manual for cognitive and behavioral scientists. *Behavioral and Brain Sciences, 27*(5), 603–627.
Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology, 1*(2), 143–186.
Skyrms, B. (2010). *Signals: Evolution, learning, and information*. Oxford University Press.
Sugden, R. (2004). *The economics of rights, co-operation and welfare*. Springer.
Thoma, J. (2020). In defence of revealed preference theory. *Economics and Philosophy.*
Vredenburgh, K. (2020). A unificationist defence of revealed preferences. *Economics and Philosophy, 36*(1), 149–169.
Weiskopf, D. A. (2011a). Models and mechanisms in psychological explanation. *Synthese, 183*(3), 313.
Weiskopf, D. A. (2011b). The functional unity of special science kinds. *The British Journal for the Philosophy of Science, 62*(2), 233–258.
Weiskopf, D. A. (2017). The explanatory autonomy of cognitive models. Integrating psychology and neuroscience: Prospects and problems. Oxford: Oxford University Press, forthcoming.
Zollman, K. J. (2011). Separating directives and assertions using simple signaling games. *The Journal of Philosophy, 108*(3), 158–169.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.