9-1-2018

Bridging the Gap between Similarity and Causality: An Integrated Approach to Concepts

Corinne L. Bloch-Mullins
Marquette University, corinne.bloch-mullins@marquette.edu

Accepted version. *The British Journal for the Philosophy of Science,* Vol. 69, No. 3 (September 1, 2018): 605-632. DOI.

Bridging the Gap between Similarity and Causality: An Integrated Approach to Concepts

Corinne L. Bloch-Mullins
Department of Philosophy, Marquette University, Milwaukee, WI

Abstract

A growing consensus in the philosophy and psychology of concepts is that while theories such as the prototype, exemplar, and theory theories successfully account for some instances of concept formation and application, none of them successfully accounts for all such instances. I argue against this 'new consensus' and show that the problem is, in fact, more severe: the explanatory force of each of these theories is limited even with respect to the phenomena often cited to support it, as each fails to satisfy an important explanatory desideratum with respect to these phenomena. I argue that these explanatory shortcomings arise from a shared assumption on the part of these theories, namely, they take similarity judgements and application of causal knowledge to be discrete elements in a theory of concepts. I further propose that the same assumption carries over into alternative theories offered by proponents of the new consensus: pluralism, eliminativism, and hybrid theories. I put forth a sketch of an integrated model of concept formation and application, which rejects this shared assumption and satisfies the explanatory desiderata I discuss. I suggest that this model undermines the motivation for hybrid, pluralist, and eliminativist accounts of concepts.

1 Introduction

Over the last few decades, three theories have dominated the philosophical and psychological literature on concepts: the prototype theory, the exemplar theory, and the theory theory.[1] The prototype theory maintains that a concept is a body of statistical knowledge about the properties of the members of a class (Posner and Keele [1968]; Rosch and Mervis [1975]). The exemplar theory argues that it is a body of knowledge about the properties of individual members of a class (Medin and Schaffer [1978]; Nosofsky [1986]). I refer to these two views as the similarity-based approaches (see discussion in Medin [1989]). Lastly, the theory theory asserts that the concept of a class stores knowledge that is explanatory of the properties of the members of the class (Carey [1985]; Murphy and Medin [1985]; Gelman and Markman [1986]; Keil [1989]). While proponents of the three theories differ in their views on the structure and function of concepts, they accept the traditional assumption that all concepts share a general common structure and that a single model could, in principle, be developed, which would account for the formation and application of all concepts. However, despite an accumulation of a large amount of data, none of these approaches has succeeded in formulating such a model. As a result, a growing number of authors have argued that while each of these theories of concepts is able to account for a range of empirical phenomena related to the formation and use of concepts, no single theory has been successful in accounting for all, or even most, of these phenomena (Piccinini and Scott [2006]; Machery [2009], [2014]; Weiskopf [2009]).

In accord with this view, which I refer to as the 'new consensus',[2] several solutions have been offered: hybrid, pluralist, and eliminativist theories of concepts. Hybrid theories of concepts maintain that each category is represented by a single concept, and every concept is comprised of several distinct parts (for instance, a theory part and an exemplar part). These parts store different types of information about the members of the category (Rips *et al.* [1973]; Smith *et al.* [1974]; Osherson and Smith [1981]; Anderson and Betz [2001]). Pluralist theories argue that concepts are not of a single kind. Rather, there are multiple kinds of concepts (such as prototype concepts, exemplar concepts, and theory concepts). Consequently, no single theory—not even the hybrid theory—can be formulated to explain the formation and application of concepts (Piccinini and Scott [2006]; Machery [2009], [2014]; Weiskopf [2009]). Finally, eliminativism accepts the above pluralist position, and further maintains that we should dispose of the very concept of concept. Instead of studying concepts, psychologists should study, for example, prototypes, exemplars, and theories (Machery [2009], [2014]).[3]

I argue against the new consensus, and maintain that philosophers and psychologists have mischaracterized the problem. The problem is not merely that each of the homogeneous theories of concepts discussed above accounts for some, but not all, instances of concept formation and application.[4] Rather, the issue is more severe: the explanatory force of each of these theories is limited, even with respect to the phenomena often used to support it. This is because, while these theories provide useful insights about certain aspects of these phenomena, they each fail to accommodate an important explanatory desideratum with respect to these phenomena. In other words, the difficulty is not just with the scope of the empirical phenomena that each theory explains, but also with the depth of the explanations each theory provides.[5] I show that an understanding of this problem gives rise to an integrated model of concepts, which resolves it.[6]

In Section 2, I discuss the similarity-based theories of concepts: prototype and exemplar theories. I argue that without incorporating causal knowledge into the model, these theories do not accommodate what I term the 'selection desideratum'. In Section 3, I discuss the theory-based theories, and argue that without incorporating a notion of similarity, these theories do not accommodate what I term the 'range desideratum'. In Section 4, I propose a solution in the form of a highly integrated model of categorization, which resolves the explanatory and conceptual difficulties I discuss. I suggest that the model undercuts the motivation, framed by the new consensus, for hybridist, pluralist, and eliminativist models.

First, however, something needs to be said about scope of this article, and the purpose of my discussion of empirical data and experimental protocols. I do not attempt to discuss all theories of concepts, and all the empirical phenomena that are relevant to a model of concepts (not even all the phenomena that are relevant to the model I sketch in Section 4). Rather, the empirical data and experimental protocols I discuss specifically target the central assumption of the new consensus. This assumption, which provides motivation for a specific kind of hybridism, pluralism, or eliminativism, is that each of the homogeneous theories discussed above explains some phenomena very well, but that none explains all. In response, my strategy is to take a 'flagship phenomenon'—an empirical phenomenon that the theory is taken as particularly successful in explaining—for each of the homogeneous theories and show that there are explanatory gaps even in the theory's account of such a phenomenon. For similarity-based approaches, I discuss studies that show, for example, that categorization tracks judgements about typicality. And for theory-based approaches, I discuss studies that show that categorization is influenced by judgements about causal structure. Thus, the data I discuss are not meant to provide empirical proof for the model I advocate (even though I hold that the data are compatible with it). Instead, they are used as illustrative devices to point out conceptual difficulties in each of the homogeneous theories discussed here.

2 The Similarity-Based Approach and the Importance of Theory

2.1 The similarity-based approach

Within the similarity-based approach, the two dominant theories are prototype theories and exemplar theories. According to prototype theories, the concept of a class is a prototype—a body of statistical knowledge about the properties of the members of this class (Posner and

Keele [1968]; Rosch and Mervis [1975]; Smith *et al.* [1988]; Smith and Minda [2002]). For example, one's concept cat would include properties such as 'has fur', 'purrs', and so forth. In order to determine whether a target, say Sylvester, is a cat, one calculates the similarity between Sylvester and the cat prototype. According to exemplar theories, concepts are sets of exemplars, which are bodies of knowledge about the properties of individual members of a class (Medin and Schaffer [1978]; Nosofsky [1986]). For example, one's concept cat may include representations of one's own cat, of the neighbours' cat, and so forth. In order to determine whether Sylvester is a cat, one calculates the similarity between Sylvester and the cat exemplar(s) one had previously encountered and stored in memory. The prototype and the exemplar theories hold that the categorization of a target as a member of a class depends on the computation of its similarity to the prototype or the exemplar, respectively. Thus, while there are interesting differences between prototype and exemplar theories (and the various versions of each), I refer to them collectively as the similarity-based theories, and contrast them with the theory theory. (For further discussion, see Medin [1989]; Rosch [1999b].)

The various versions of the similarity-based approach have been taken as successfully explanatory of a range of empirical phenomena (for review, see Laurence and Margolis [1999]; Murphy [2002]; Margolis and Laurence [2004]; Machery [2009]). For example, a prototype of a category is experimentally reconstructed by asking subjects to list features of particulars belonging to members of that category. Subjects rate instances that share a lot of properties with other members of a category as better examples of that category than they do instances with fewer such properties (Rosch and Mervis [1975]). Such instances are more likely to be classified as members of a category and are categorized faster than non-typical ones (Rosch and Mervis [1975]; Hampton [1979]). Among the evidence supporting the exemplar theory is the finding that previously experienced items are more easily classified than new items, even when they have the same degree of typicality (for a review, see Nosofsky [1992]). Additionally, new items that are similar to previously encountered category members are categorized faster than items that are less similar to old stimuli, even when they are farther from the category's prototype (Medin and Schaffer [1978]).

2.2 The selection desideratum

The similarity-based approach, however, faces an explanatory challenge. Similarity is not absolute—it must always be considered with respect to a particular set of properties (see Goodman [1972]; Medin [1989]). As Medin ([1989], pp. 1473–4) pointed out:

> […] a zebra and a barberpole would be more similar than a zebra and a horse if the feature 'striped' had sufficient weight […] attempts to describe category structure in terms of similarity will prove useful only to the extent that one specifies which principles determine what is to count as a relevant property and which principles determine the importance of particular properties.

Without such principles, a similarity-based theory is faced with what Machery ([2009]) calls the 'selection problem'. Accordingly, one desideratum of a similarity-based approach is that it be able to account for the way in which one chooses the individual features that serve as the basis for categorization. In line with Machery, I refer to this as the 'selection desideratum'.

Similarity-based approaches have attempted to accommodate the selection desideratum by taking the effects of the entire stimuli set into account. When the relevant features for similarity judgements are not specified, they are often inferred from the stimulus context—the set of objects under consideration (Torgerson [1965]; Tversky [1977]; Goldstone *et al.* [1997]). Different versions of similarity-based theories have incorporated the effect of stimulus context into their models in different ways. Some have suggested that basic categories are formed in a manner that maximizes both category resemblance and cue-validity of the properties of the particulars in the category (Rosch [1999a]) (I refer to this as the 'maximization principle'). Cue-validity of property y is the probability that an object belongs to a category X if it has the property y—for example, the probability that Sylvester is an instance of cat if he has fur. A property will have high cue-validity if particulars within the category have it and particulars in the contrast class (the class of items that the category is distinguished from) do not. For example, the cue-validity of 'has fur' for cats would be high if contrasted only with birds, but low if contrasted only with dogs. Thus cue-validity is one way of taking into account the effects of the set of particulars under consideration on the selection of features for similarity judgements.[7] In what follows, I use the term 'set of particulars under consideration' (hereafter SOP) instead of the narrower term 'stimulus context', to reflect the idea that such effects are not necessarily limited to the set of particulars one is currently exposed to.[8]

The incorporation of SOP effects into similarity-based models allows them to partially account for the selection of the relevant 'respects' for similarity judgements. Importantly, these and other adjustments—which take into account the typicality of properties in the target class and in contrast classes, their salience, their 'perceptibility', and so on—work to the extent that they rely on an assumption about the nature of the perceived world, namely, that it is carved at the joints in a manner that gives rise to the formation of a unique division into categories. Rosch's working assumption, for example, is that 'in the perceived world, information-rich bundles of perceptual and functional attributes occur that form natural discontinuities, and that basic cuts in categorization are made at these discontinuities' (Rosch [1999a], p. 192; see also Rosch [1973]).[9]

The assumption that the perceived world is 'carved at the joints' in a way that gives rise to a unique system of classification is problematic. Studies have shown, for example, that the fundamental taxonomic rank chosen by classifiers depends upon factors such as expertise (Tanaka and Taylor [1991]) and cultural factors (Dougherty [1978]). Moreover, classification based only on the maximization principle does not result in stable categories. The early numerical taxonomists attempted to classify organisms in a theory-free, inductive method. In order to calculate similarity, they used—in equal weights—all of the organisms' legitimate biological features (Hull [1988]; this example is also discussed in Griffiths [1997], p. 178).[10] There seemed to be an indefinite number of ways to subdivide organisms into features, and without any selection criteria, all subdivisions were equally plausible. Adding the maximization principle into a similarity-based model is insufficient to accommodate the selection desideratum.[11]

Of course, it is not necessary for a similarity-based approach to assume that there is only one way to carve up the perceived world (or the perceived SOP) that results in maximized in-category similarity and inter-category difference. In fact, Rosch herself concedes that with

respect to some categories, the classification system is not completely fixed by our biology, but is also influenced by an individual's culture and language (Rosch [1975b]). The present point is that if there are two or more ways to carve up the same perceived world, then a theory of concepts based solely on similarity judgements would be unable to explain why we prefer one classification to another.[12]

2.3 Causal knowledge as satisfying the selection desideratum

In order to satisfy the selection desideratum, a model should incorporate additional elements, such as the goal of classification and background knowledge (see discussion in Medin *et al.* [1993]). For example, people form categories for the purpose of goal-relevant inferences (Holland *et al.* [1989]). Therefore, it may be expected that goals would affect the selection of relevant respects of similarity, thus directly influencing similarity judgements. Barsalou ([1982]) found that, for *ad hoc* categories (such as 'things that can float', 'can be a pet'), exposure to the name of the category, which conveys information about the use or function of the items in the category, increases the similarity rating for pairs of items belonging to the category. Discussing the various factors that may be packed under the term 'background knowledge' and that may affect the selection of the relevant features for similarity judgements is beyond the scope of this article. Here, my focus is specifically on causal knowledge.

Causal knowledge affects the selection of features for similarity ratings and for categorization. Ahn *et al.* ([2000], [2002]) found, for example, that causally fundamental features were more important than causally superficial ones for various category-related tasks, including goodness-of-exemplar ratings, free-sorting, similarity judgements, categorization, and sensitivity to feature correlations. One might point out that people often categorize without having much causal knowledge. Thus it might seem that while causal knowledge affects the selection of the relevant properties in some cases of categorization, it may not affect the selection in all cases. It is not required, however, that one has all the relevant background knowledge to select the relevant properties for similarity judgements and categorization. Missing elements—including causally fundamental features—are often inferred from other features, from the subject's general background knowledge, or from analogy with other domains, and these affect categorization (Medin and Ortony [1989]; Rehder and Kim [2009]). Additionally, one's theory about the general domain may lead one to selectively focus on some properties. This suggestion is supported by the findings that functional features are important in classification of artefacts (Wisniewski [1995]) and biological kinds (Lombrozo and Rehder [2012]). Therefore, even minimal knowledge about the targets would enable causal considerations to factor into similarity judgements and classification.

I have argued that in similarity-based theories, the constraints imposed by the maximization principle do not accommodate the selection desideratum, and that additional factors, such as causal knowledge, should be added to accommodate it. It might be objected that even if, for some sets of objects (such as species categories), there are two (or more) alternative classifications that satisfy the maximization principle, there are at least some simpler cases for which only one classification achieves such maximization when all properties are given equal weights. In such cases, selection of specific properties is not required, and classification can

be explained by appeal to overall similarity. Therefore, it may be argued, the selection desideratum is not central to a general theory of concepts.

In response, it should be noted that even in cases we might take as much simpler than biological classification, one is unlikely to weigh all properties equally (consider, for example, the importance of shape versus colour in classification of kitchen tools). Given that factors such as background knowledge and goals affect the selection of relevant properties for similarity judgements and classification in various cases, there is no reason to assume that giving equal weights to the various relevant properties is one's default position in classification. Attributing equal weights to all properties is, thus, one possible (although unlikely) outcome of a selection process. Even if there are cases in which all properties are considered in equal weights, the selection process still requires explanation.

Moreover, if 'simple' means fewer properties, then it is not clear that there are any simple cases for categorization, as the list of possible properties for a given object is infinite. The issue is illustrated by Murphy and Medin ([1985], p. 292; see also Medin [1989]):

> Suppose that one is to list the attributes that plums and lawnmowers have in common in order to judge their similarity. It is easy to see that the list could be infinite: Both weigh less than 10,000 kg (and less than 10,001 kg …), both did not exist 10,000,000 years ago (and 10,000,001 years ago …), both cannot hear well, both can be dropped, both take up space, and so on. Likewise, the list of differences could be infinite.

If the list of possible properties for each object is infinite, then there are no simple cases that give rise to overall similarity judgements. What counts as a relevant property, therefore, must be somehow constrained. The similarity-based approach may provide some structural constraints, based on principles of cognitive economy (for example, a property should apply to many but not all particulars) in addition to the maximization principle.[13] However, as Smith and Medin ([1981], pp. 15–18) point out, these constraints are not sufficient, and it seems that the most important criterion for what counts as a relevant property for similarity judgements is that the property is used as input for categorization processes. If this is the case, explaining categorization purely in terms of similarity is circular. Similarity alone cannot satisfy the selection desideratum, which is warranted even in so-called simple cases.[14]

The above discussion is not meant to provide empirical support for the idea that similarity alone could not satisfy the selection desideratum. Rather, the discussion is meant to demonstrate the conceptual difficulty with the traditional notion of similarity, if such a notion is to be explanatory of the structure, formation, and use of concepts. The explanatory force of the similarity-based approach is limited, even with respect to the phenomena that are taken to support it.[15] I suggest, accordingly, that the criticism that models that are based on similarity alone can only explain some of the phenomena related to concepts is not strong enough. Rather, even with respect to the empirical phenomena that are often used to support them (such as categorization judgements that track typicality ratings), the explanations provided by these models fall short of satisfying the selection desideratum. In order to satisfy it, models of concept formation and application should include the effects of factors such as goals and

background knowledge on the selection of relevant properties (and correlations of properties) for similarity judgements.

In the following section, I examine whether classification (and other related tasks) can be explained solely within the framework of a competing approach, the theory theory, which relies heavily on the use of causal knowledge. I conclude that while the notion of similarity (as traditionally conceived) is not sufficient to accommodate the selection desideratum for a theory of concepts, it should not be discarded and, in fact, by excluding similarity from its account, the theory theory fails to accommodate a different explanatory desideratum. This shortcoming is rooted in the same assumption made by the similarity-based theories, namely, that similarity judgements and causal knowledge should be taken as discrete elements in a theory of concepts.

3 The Theory-Based Approach and the Importance of Similarity

3.1 The theory-based approach

According to the theory-based approach, the concept of a class stores knowledge that is explanatory of the properties of the members of the class, knowledge of laws, as well as causal, functional, and generic propositions. While various versions of the theory-based approach differ on the specifics of the categorization process, they generally agree that a target is taken to belong to a class if its properties are judged to be generated (or at least constrained) by the causal structure that characterizes the members of that class (Medin and Ortony [1989]; Sloman *et al.* [1998]; Rehder and Hastie [2001]; Rehder [2003b]).For example, one's concept cat may include the theory that the genetic structure of cats is causally responsible for cats' various properties. If one judges that Sylvester likely has the same causal structure as members of the category of cats, one will classify Sylvester as a cat.[16]

The theory theory is often taken by proponents of the new consensus as successful in explaining the effect of causal knowledge on subjects' classification decisions in tasks involving a small number of properties (see discussion in Machery [2009]). I argue, however, that taken solely on its own terms, the theory theory is limited even with respect to the phenomena that are often used to support it, as it does not satisfy what I call the range desideratum. I first consider a representative example from experiments used in support of the theory theory in order to illuminate the theory's explanatory limitation. I argue that while the theory may explain how the properties relevant for categorization decisions are chosen, it does not explain how one determines whether a particular target shares these relevant properties to a sufficient degree (or what such a determination amounts to). In Section 3.3, I suggest that the missing element involves the application of similarity judgements.

3.2 The range desideratum

The experiments often invoked in support of the theory theory are usually composed of two stages: the first is a learning stage, in which subjects learn the causal relations among a few properties of members of a category. For example, Rehder taught subjects about a novel category, lake victoria shrimp. For members of this category, the property 'high amounts of

ACh neurotransmitter' is causally related to other properties, such as 'long lasting flight response' (Rehder [2003a], [2003b]). After the learning stage comes the testing stage, in which subjects are presented with new targets, and are asked to perform various tasks, such as categorizations, goodness-of-exemplar rating, free-sorting tasks, or similarity ratings (for instance, Ahn *et al.* [2000]; Rehder [2003a], [2003b]). In the experiment described here, for example, subjects rated targets as good category members to the extent that their features manifested the expectations induced by the causal story provided to them (Rehder [2003a]).

In these experiments, the properties provided to subjects are almost always binary ones—a target either has, or doesn't have, property y. Property y can, of course, belong to a dimension that is potentially continuous in character (such as amounts of ACh), and even stand for a range of values (such as the property 'high amounts of ACh neurotransmitter', which presumably stands for a range of possible ACh levels).[17] However, both in the learning stage and in the testing stage, participants are presented only with binary values for these dimensions: they are informed that members of the category (or a certain percentage of them) and, later, the target to be classified either possess a property or do not possess it. For example, particulars either have 'high amounts of ACh neurotransmitter' or they do not. As a result, during the testing stage, subjects effectively judge whether there is a sameness relation between the target and the representation of the category, with respect to each relevant property.[18]

The problem is that there is no reason to suppose that this is how the comparison of features is performed in real-life categorizations and, in fact, there are reasons to suppose that, at least in many cases, it is not. As in the case of ACh levels, dimensions that are used for classification may receive multiple values. In such cases, there is no demand that the values exhibited by members of a category be identical. For example, it is not the case that a target must have an exact value of ACh levels, say 7.23 units, in order to have 'high levels of ACh neurotransmitter'. Rather, there is a range of ACh level values that, when exhibited by the target, would be taken as sufficiently high.

A theory that is able to account for how one determines the appropriate range of values for a causally relevant dimension would have the advantage of telling a fuller story about how we make classification decisions with respect to categories with causal structure. This is the range desideratum. Since the majority of empirical phenomena used in support of the theory theory comes from experiments that do not present subjects with ranges of possible values, such phenomena are silent with respect to the range desideratum.

One might object that decisions about ranges of values for categories like lake victoria shrimp are outside the scope of categorization decisions made by a non-expert. True. But so is the classification of a particular kind of shrimp, and the consideration of ACh levels for that purpose. The above experiment, while using scientific categories and parameters, was meant to shed light on 'everyday' processes of classification. My point here is that for whatever dimensions one does take into account in classification, a theory of concepts should, ideally, be able to say something about how the appropriate ranges of values for these dimensions are chosen. To illustrate this, let us consider a toy example. Suppose one has the following causal theory about teenagers:

A theory theory about the concept teenager could satisfy the selection desideratum (that is, explaining how one determines the relevant dimensions for classification). If one has no special access to the target's age (or hormonal levels), one will use the dimensions 'eye rolling' and 'door slamming' for the purpose of categorization, because they are taken to be indicative of the deeper cause (if we apply Ahn *et al.* [2000]) or because they are taken to be indicative of a specific causal structure (if we apply Rehder [2003a], [2003b]).[19] However, since there is no requirement that individuals belonging to teenager actually be identical with respect to any of the relevant dimensions, the range desideratum needs to be satisfied. The causal theory alone cannot satisfy it, since it does not tell us how one determines the relevant ranges of values for these dimensions (for instance, how forceful should the shutting of the door be, in order to be considered a slam? What is the required frequency of eye rolling one has to engage in, to be considered an eye roller?). A theory that is able to satisfy the range desideratum would better account for why one classifies Jack (three eye rolls per hour) as a teenager, but not Jill (one eye roll per hour).[20]

3.3 Similarity as satisfying the range desideratum

I argued that even in the case of causal categories, one has to know the range of acceptable values for the relevant dimension in order to perform classification. I further argued that the theory theory does not tell us how one determines this range, that is, it does not satisfy the range desideratum. Note, however, that to determine such a range of acceptable values is precisely to answer questions like: how similar do ACh levels of particulars have to be to each other in order for the particulars to be members of the category lake victoria shrimp?[21] In other words, forming a range of acceptable values is a way by which one forms a similarity threshold for categorization decisions. Thus the comparison between the individual dimensions (and correlation among their values) of members (or potential members) of a category involves similarity judgements, even if causal knowledge affects the selection of the relevant dimensions.[22]

The need to move beyond matching of binary properties was recognized by Waldmann *et al.* ([1995], Experiment 2). In their study, subjects received causal explanations and then learned a category (such as a disease with various symptoms) through a categorization task with feedback. The dimensions used in the description of the targets were 'body weight' and 'pallor'. This experiment is of interest because unlike most experiments with causally structured categories, dimensions were not given merely binary values such as yes/no or high/low. Instead, they were given one of four possible value brackets. For example, the four values for body weight were 'slightly underweight', 'underweight', 'seriously underweight', and 'anorexic'. Subjects' performance (as measured by the number of errors) was better when classifying targets whose causal structure were closer to that predicted by the causal story they had learned. Their performance in the task was graded, exhibiting sensitivity both to the values of the two dimensions and to the correlation between them. The significant point here is that once

we accept that a relevant dimension is continuous, the notion of similarity becomes necessary for explaining how a causal representation is applied to specific targets that have different values along that dimension. It is essential to my view that similarity judgements are not restricted to perceptual features (Hahn and Chater [1997]; Hampton [1998]; Medin [1989]; Gärdenfors [2000]; Mazzone and Lalumera [2010]; Rheins [2011]). Similarity judgements can therefore be applied to various theoretical characteristics often featured in causal theories, including the causal relations themselves.

Importantly, the present point relates to a view about what categorization entails. Namely, that even if we accept the view that category members share a causal structure, it does not entail the stronger commitment that people group things in categories because they take them as sharing some identical property (or an identical relation between properties). Even if we concede, for example, that people believe that all raccoons share the internal, causally fundamental property 'having raccoon DNA', and that they believe that raccoons are members of racoon in virtue of sharing this property, we are still not entitled to the strong claim that people classify raccoons in this way because they think that there is something identical about the DNA of all raccoons. We are only entitled to the weaker claim, that they classify raccoons in this way because they think there is something similar about the DNA of all raccoons. Because most experiments that are invoked in support of the causal theory use binary values, they are silent with respect to the stronger claim. Given that the sort of causal properties subjects are often asked to consider are potentially continuous (such as levels of ACh), and given what we may reasonably assume about non-experts' general knowledge about such properties (for instance, that exact levels of neurotransmitter vary among particulars of the same species), the stronger claim, it seems to me, is not warranted. The same goes for non-scientific concepts. One may very well have a causal theory for teenager. But this does not pose the requirement that to classify Jack and Jill as teenagers, one must believe that they share an identical age, or an identical hormonal level, or that the causal chain between their hormone levels and their door slamming is identical. What Jack and Jill share, and what various instances of lake victoria shrimp share, is a similar causal structure.[23]

Four objections can be made here. The first is that one may receive the range of values from experts (for instance, scientists can tell us what the proper range for ACh levels is for lake victoria shrimp). In this case, one can apply the concept to a target without engaging in any similarity judgements. Furthermore, experts themselves often use their causal knowledge to determine the appropriate range of values for a given dimension.[24] I agree with both points. However, given my view, just discussed, on what categorization entails (and what similarity is), I would argue that for both the expert and the non-expert, grasping that there is such a range of values among members of a category amounts to grasping that the relation between members of a category with respect to the relevant dimension is that of similarity. Therefore, even if the determination of the range of values for the relevant dimension is performed by experts, and even if this determination is guided by causal considerations, grasping a concept requires at least the implicit grasp that members of a category stand in a relation of similarity to one another.

The second objection is that with my emphasis on 'acceptable range', my characterization may be taken to imply that one must have an explicit knowledge of the range of values in order to

make the classification, when in fact people rarely do. My argument, however, does not necessitate such knowledge (for example, it is not necessary that one has an exact cut-off value for how many eye rolls per hour are required for an individual to be taken as an eye roller). Knowledge of the proper range is often implicit, and fuzzy borders for the range of acceptable values may lead to fuzzy categories. Note, moreover, that any model of categorization that involves similarity judgements will have to incorporate some mechanism for determining how similar is 'similar enough', and there is nothing in my suggestion that demands a higher degree of explicit knowledge or accuracy about where the cut-off point should be.

A third, related objection is that my view may be taken to imply that the decision about whether a given target falls within the proper range of values, with respect to the relevant dimension is an all-or-nothing process. Similarity-based category judgements, however, often exhibit typicality effects (typical items are classified faster than non-typical items) (Rosch [1975a], [1999a]; Rosch and Mervis [1975]). My reply is that forming a range of acceptable values merely provides one with a mechanism of translating similarity judgements into 'yes/no' categorization decisions. Nothing in the account, however, precludes the various effects associated with similarity judgements themselves (for instance, it is perfectly consistent with my view that it may take us longer to classify targets that are closer to the range's extreme than those that are placed in the middle of the range, with respect to the relevant dimensions).

A fourth objection might go as follows: I suggested that the process of categorization based on similarity with respect to each casually relevant dimension can be explained as a two-step process. First, during the initial formation of the category, one forms a range of acceptable values. Then, in order to make a decision about the classification of target T, one judges whether T's properties fall within that range. This, it may be argued, disregards the flexible nature of similarity judgements, which may change as one is exposed to additional targets. It has been shown that the similarity rating for two targets, even when confined to a single dimension, is influenced by the presentation of a third target that serves as a foil. For example, two small black disks, presented along with a large black disk, were judged to be more similar as the size of large disk increased (King and Atef-Vahid [1986]). This means that what one takes as the range of acceptable values for a dimension might change as one encounters new targets. In response, I'd like to point out that my suggestion allows for such modifications of similarity judgements, and entails that these will be translated into categorization decisions through a change in the range of allowable values within the relevant dimension.[25] As I discuss in Section 4.1, there is no requirement that the borders of the range be rigid.

I have argued that the experiments often appealed to by proponents of the theory theory require participants to make sameness judgements with respect to dimensions with binary values. Such experiments do not enable researchers to test for similarity judgements with respect to individual features or relations used in categorization. It is not surprising, therefore, that proponents of the theory theory often consider similarity judgements to be irrelevant for categorization or for category-based inferences performed in the presence of causal knowledge (Rehder and Hastie [2001]; Rehder [2006]). Since the dimensions that are used in actual categorization and category-based inferences often receive non-binary values, I argue that a theory that explains how the proper range of values is determined would be

advantageous. I further suggest that the notion of similarity may facilitate such an explanation, even in the case of categories whose members share a causal structure.

Again, this discussion of classification decisions is not meant to be exhaustive of the tasks that involve the use of concepts or the phenomena that a theory of concepts aims to explain. Rather, the discussion is meant to highlight a conceptual difficulty with the idea that non-identical particulars can be classified as members of a category without the use of similarity judgements.[26] This difficulty was demonstrated by showing the explanatory limitation of the theory, even with respect to phenomena that are taken to support it. If I am right, then the diagnosis of the new consensus—that the theory theory is unsatisfactory because it only accounts for some of the category-based phenomena—is not strong enough. Rather, the theory theory, when it does not incorporate similarity judgements, falls short of accommodating the range desideratum, and therefore does not provide a satisfactory account of categories whose members share a causal structure.

4 An Integrated Approach to Concepts

In Section 2, I argued that similarity-based approaches do not satisfy the selection desideratum unless they admit theoretical considerations into their models. In Section 3, I argued that theory-based approaches do not satisfy the range desideratum unless they admit similarity judgements into their models. I now outline an integrated model of concept formation and application that arises from this characterization of the problem and satisfies both desiderata. I suggest that the model undercuts the motivation for the hybridism, pluralism, or eliminativism formulated by the new consensus.[27] I should emphasize that the suggested model should not be taken to be exact or comprehensive. Rather, it provides a framework within which the above discussion can be understood and integrated. In Section 4.2, I contrast my theory with hybrid theories of concepts, in order to explicate the strong sense of integration to which my view is committed.

4.1 An integrated model

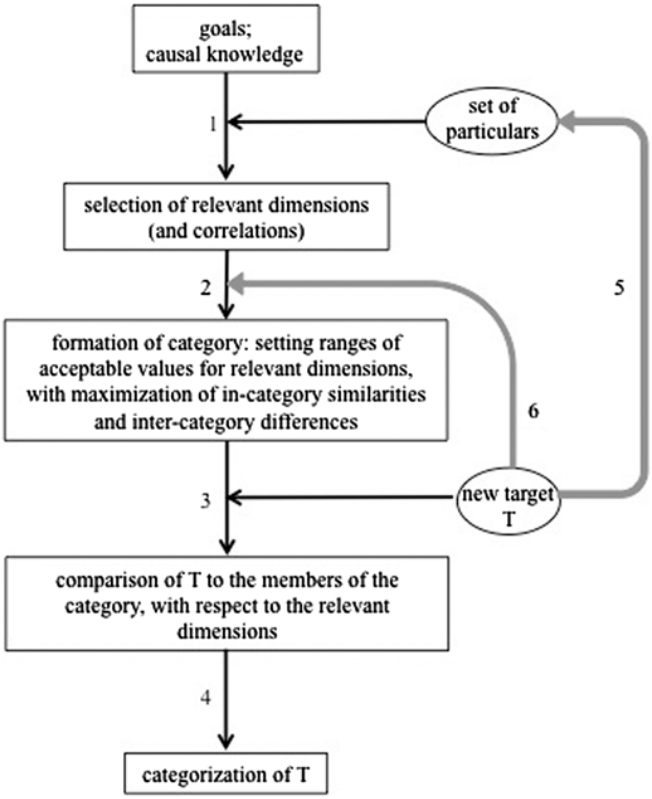
The suggested model is illustrated in Figure 1, which depicts the initial formation of a category and the subsequent categorization of a new target. Since the model integrates data discussed in some detail in Sections 2 and 3, in what follows, I only briefly relate the elements of the model to some of the previously discussed data, and then discuss a number of implications.

As discussed in Section 2, several factors affect the selection of dimensions for similarity judgements (Arrow 1 in Figure 1). First, the distribution of features among the SOP (contrast among particulars, perceptual salience of certain properties) affects the selection of features for similarity judgements (Rosch [1999a]). Goals and causal knowledge also affect the features used in similarity judgements (Barsalou [1982]; Medin *et al.* [1993]; Ahn *et al.* [2000]). As discussed in Section 3, the dimensions selected for similarity judgements may include causal relations (accordingly, what is judged to be similar between Jack and Jill is, I suggest, not only their age and their hormone levels, but also the causal chain between their age and their hormone levels). The involvement of causal knowledge in categorization is therefore not limited to the selection of dimensions for similarity judgements (in which case one might argue

that causal knowledge is, in some sense, external to such judgements); rather, causal relations are, in part, what similarity judgements are about.

As discussed in Section 3, the ranges of acceptable values for dimensions and correlations that are deemed causally relevant are determined (Arrow 2 in Figure 1). The ranges should maximize in-category similarity and inter-category differences, as suggested by Rosch ([1999a]), with the additional stipulation that this maximization is not absolute, but is relative to the selected dimensions. This range is dynamic, and depends on the various properties of the SOP. The model sketched here is silent with respect to the psychological mechanism that determines this range (it may involve, for instance, the use of prototypes, exemplars, or both).

Figure 1.



An integrated model for classification and categorization. Initial formation of a category, and the following categorization of new targets are described. Arrows indicate causal connections; grey arrows indicate feedback loops.

After the ranges of the relevant dimensions are formed, the relevant properties of a new target are compared to the ranges of acceptable values (Arrow 3 in Figure 1), and a classification decision is made (Arrow 4 in Figure 1). The model, however, emphasizes the dynamic structure of concepts (and similarity judgements). Along the lines of Gärdenfors's ([2000], Section 4.5) account of concept learning, the process described here is not one-directional, nor is it strictly divided into a category-formation stage and a category-application stage (see

feedback loops Arrows 5 and 6 in Figure 1). Given the effect of context on the selection of relevant respects for similarity (see discussion in Section 2.2), a presentation of a new target may, for example, affect the selection of relevant dimensions for similarity calculation. Additionally, in accordance with the anchor-range effect (King and Atef-Vahid [1986]), a new target may alter the ranges of the appropriate values with respect to the dimensions deemed relevant for similarity calculations, thus changing the resolution of the similarity judgements.

It is likely that additional feedback loops should be added to the model. For example, the formation of a mental category (or the labelling of that category) may shift one's attention towards new dimensions and affect similarity judgements (for discussion of some relevant phenomena, see Collins and Olson [2014]). Additionally, the formation of a new category may contribute to the development of one's theoretical knowledge about members of the category by, for instance, facilitating inductive inference (see, for example, discussion in Smith [1989]). This theoretical knowledge may, in turn, affect the selection of relevant dimensions, leading to a new classification. Thus, the suggested model highlights a view of concepts as dynamic constructs. I should point out, however, that despite the emphasis on these dynamic features, the model does not take concepts to be constructs in working memory, created on the fly.[28] While the debate on the stability of concepts is outside the scope of this article, it should be noted that one may consistently hold that concepts are fairly stable constructs, stored in long-term memory, yet that they may change as new knowledge comes to light. The current model may accommodate both approaches.[29]

The model proposed here takes a broad view of similarity. As I point out, similarity need not be merely perceptual, and can be calculated across abstract features and relations (Hahn and Chater [1997]; Hampton [1998]; Gärdenfors [2000]; Medin [1989]; Mazzone and Lalumera [2010]; Rheins [2011]). It may appear that in painting such a broad picture of similarity, I have turned the idea of similarity into a vacuous one—similarity just is anything that would account for putting things in the same category, whether it's perceptual or more abstract (see discussion in Medin *et al.* [1993]; Goldstone [1994]; Sloman and Rips [1998]). As discussed above, however, making a similarity judgement is a process in which two (or more) things are compared against a (provided or postulated) background that serves as a foil, and a judgement of similarity requires the specification (either implicitly or explicitly) of relevant dimensions for comparison. The term 'similarity', therefore, imposes specific demands on what a similarity-based account of concepts must explain: it must account for the effect of both SOP and background knowledge on the selection of the relevant dimensions for similarity judgements, and the determination of the proper ranges for these dimensions. Moreover, the view of similarity incorporated into the current model leads to specific predictions that may be tested empirically. The model predicts, for example, that the objects that serve as the foil in categorization decisions would affect both the selection of relevant dimensions for categorization and the ranges of values along these dimensions, and that these effects would be observed even in categories in which causal knowledge plays a central role.

4.2 The integrated theory versus hybrid theories of concepts

Some authors who have accepted the new consensus have tried to salvage the concept of the concept by appealing to hybrid theories. Various hybrid models have combined elements from

several homogeneous theories. In what follows, I briefly discuss some of these, and argue that since traditional hybrid approaches do not identify some of the explanatory issues at stake, they also fall short of resolving them. I suggest that the integrated approach has explanatory advantages over traditional hybrid theories. I end by comparing my view to a couple of highly integrated hybrid models, spelling out the strong sense of integration that is implied by my view.

Versions of hybrid theories differ from each other in what they take as the elements that compose concepts, and in the ways they take these elements to be connected and coordinated. Anderson and Betz ([2001]), for example, suggested a cognitive architecture in which elements from exemplar-based models and from rule-based models are mixed to produce classification behaviour. In their hybrid model of categorization, the system selects which of these strategies to employ depending on the task at hand. Other approaches have suggested that a concept is a single representation containing a core component, such as a definition, and stereotypical features (Rips *et al.* [1973]; Smith *et al.* [1974]; Osherson and Smith [1981]). According to such models, the stereotypical features are used for quick categorization decisions, and the core is utilized when a decision cannot be made based on stereotypical features alone, as well as for additional purposes, such as making inferences.

My contention is that by misconstruing the problem, such hybrid theories—like the pluralist and the eliminativist approaches—have missed the solution. If we assume that each homogeneous theory successfully accounts for some, but not all, phenomena, a model that treats these theories as discrete elements that are differentially activated in various cognitive tasks may seem promising. But if, as I argue here, each of these homogeneous theories does not satisfy an important explanatory desideratum even with respect to the empirical phenomena that serve to support it, then hybrid theories that incorporate elements from these theories may inherit the same explanatory challenges faced by the homogeneous approaches. The core/exemplar model, for example, faces two explanatory difficulties. For quick categorization decisions performed using the stereotypical features, the difficulty in accounting for the selection of features remains. For slower decisions based on ‘core properties’, the question remains as to how a set of dimensions—however limited it may be—is compared among particulars. In contrast to the hybrid approaches, the integrated model does not face such difficulties, as it accommodates both the selection and the range desiderata.

Of course, not all hybrid approaches argue that a single element is activated at a time. Rice ([2014]), for example, argues that (different combinations of) the various elements of concepts, which store different types of knowledge, are co-activated and integrated before the information is processed by any higher cognitive processes. Whether the specific hybrid concepts described by Rice are immune to the above criticism would depend on which elements are activated and the specific ways in which they are connected (or merged).

Still, there is an important difference between my approach and Rice’s, which pertains to what we take concepts to be. To see this difference, it would help to consider the different ways in which our approaches respond to a challenge put forth by Machery ([2009]).[30] Machery points out that a traditional hybrid theory, in which the system selects a single element to be used in a given task, is not distinguished from pluralist approaches to concepts. The system in the

core/exemplar hybrid model, for example, utilizes, at any given moment, elements either from the exemplar-based model or the rule-based model, but it does not use the two together. So why should we say that we have one concept with two modes of deployment, rather than two concepts? The problem does not seem to be alleviated by models in which a concept is taken to be a single representation comprising several distinct components that are differentially activated. The difficulty simply shifts, and we may then ask in what way the different elements comprise a single representation. In response to Machery, Rice ([2014]) points out that his own account provides a clear individuation criterion for concepts: the various elements should be taken as comprising a single concept because they are linked together and integrated before being used as input for the same cognitive process.

My own view differs from Rice's in that I do not take concepts to be constituted by several 'parts', which store distinct types of knowledge and which functionally interact (thus, my view is not vulnerable to Machery's particular criticism concerning concept individuation).[31] In other words, I do not suggest that knowledge about similarity and causal knowledge 'come together' in the process of concept formation or activation. It is crucial to my account that similarity judgements are very broad, and apply not only to simple dimensions but also to complex dimensions including causal relations among properties. Accordingly, causal knowledge not only serves in the selection of the relevant dimensions for similarity judgements, but it also figures into what is being compared in these very judgements. To go back to a previous example, Jack and Jill are similar not only with respect to 'hormone level', 'eye rolling', and 'door slamming', but also with respect to the causal mechanisms that link these dimensions. There is no sense, therefore, in which one's knowledge about Jack and Jill, *qua* teenagers, can be understood as comprising discrete, similarity-based knowledge items and causally based knowledge items. To have the concept teenager is to take Jack and Jill (and others) as members of a category by virtue of their similarities with respect to specific dimensions and with respect to the causal links between those dimensions.

The present view may be compatible with Vicente and Martinez Manrique's ([2016]) approach to hybrids. They argue for a hybrid framework that requires that the various elements of the concept are activated concurrently and have some functional significance for the task at hand. Unlike traditional hybrid models, they are not committed to the idea that the elements are 'semi-separable in the quasi-modular sense' (Manrique and Manrique [2016], p. 75). To the extent that a hybrid theory does not require such compartmentalization, the present account may be taken as a version of a hybrid account in which causal and statistical knowledge is organized according to complex similarity relations.[32]

5 Conclusion

I argue here against the premise, shared by proponents of hybridism, pluralism, and eliminativism, that each of the homogeneous theories of concepts successfully explains some, but not all, of the phenomena involving concept formation and application. I argue, instead, that each of the homogeneous theories, when taken on its own terms, fails to accommodate an important explanatory desideratum, even with respect to the very phenomena often used to support it. I further suggested that these explanatory shortcomings arise from an assumption, shared by the homogeneous theories, that our knowledge about the similarities of category

members and our knowledge about the causal relations among their properties are discrete knowledge items. I proposed a model of categorization, in which knowledge about causal relations and similarity are highly integrated. The model satisfies the two explanatory desiderata I discussed, and undercuts the motivation, framed by the new consensus, for pluralism, eliminativism, and traditional hybridism.

[1] For the purpose of this article, I am glossing over interesting differences between several versions of each of these approaches.

[2] Importantly, the new consensus, as I use the term here, does not refer to just any view that holds that the three theories discussed above should be rejected (indeed, the present article also rejects these theories, at least in their traditional construal). Rather, I use the term to refer to the particular motivation, described above, for holding that these theories should be rejected.

[3] This version of eliminativism, which develops out of pluralism about concepts, argues that prototypes, exemplars, and theories exist, but they do not belong to a single kind. It should be distinguished from the claim that there are no such stable mental constructs (for example, Casasanto and Lupyan [2015]).

[4] Throughout the article, I describe prototype, exemplar, and theory theories as homogeneous, to be contrasted with hybrid, pluralist, and eliminativist theories of concepts.

[5] As one referee remarked, we cannot determine *a priori* which aspects of the phenomena will turn out to be relevant for a theory of concepts. It is possible that certain aspects will be explained by a mechanism that is distinct from concepts. I suggest, however, that if (as I propose below) a theory of concepts can be constructed, which accounts for these aspects of the phenomena, this should be taken as an advantage of that theory.

[6] I limit my discussion to prototype, exemplar, and theory theories, as they serve as the main targets of the pluralists and eliminativists approaches I discuss here. Therefore, I do not discuss other theories of concepts, such as embodied approaches, which hold that concepts are encoded perceptually. What I term the new consensus, then, should not be understood as a broad statement about the state of play in the field of concepts research, but rather as representing a view that has been gaining support among proponents of the amodal framework.

I should point out, however, that some pluralists would be quite happy to add embodied approaches to their arsenal (see, for example, Rice [2014]). One criticism made against embodied approaches is that, while they may work well for concrete concepts, they do not properly explain how highly abstract concepts, such as justice, are grounded in perception (Mahon and Caramazza [2008]; for some proposed solutions, see Barsalou [1999]; Gibbs [2006]; Prinz [2005]). This does not pose a problem for pluralists who are willing to admit both embodied and amodal concepts into their accounts (see discussion in Dove [2009]). I do not address this type of pluralism here. Importantly, I distinguish between questions about the content of concepts and questions about their vehicle (Mahon [2015]; Weiskopf [2010]; see also Machery [2007]; Bloch-Mullins [2015]). The explanatory challenges I discuss here arise from issues about the content of concepts, and would therefore apply to embodied theories as well. For related reasons, I hold that the sketch I propose in Section 4 could accommodate embodied views of concepts, as

well as amodal ones. While I hold that concepts contain both perceptual and non-perceptual information, my view is silent with respect to whether this information is coded perceptually or non-perceptually. Due to limitation of space, I do not specifically address embodied approaches in this article.

[7] Tversky and Gati's notion of the 'diagnosticity of features' works along the same lines (Tversky [1977]; Tversky and Gati [1978]).

[8] As recognized by the models discussed above, when it comes to concepts (rather than artificial categories produced for the purpose of an experiment), the set of particulars under consideration is not limited to the objects that one might be facing at a given moment. It also includes other particulars that a given category naturally contrasts with within one's taxonomy (Rosch *et al.* [1976]). For example, since cats are usually placed in one's taxonomy as contrasting with other animals, the cue validity for various properties of cats will be determined by contrasting them with the properties of dogs, birds, and so on, even if, at the moment the concept is used, one is not exposed to dogs. Similarly, since chairs are usually placed in one's taxonomy as contrasting with other furniture, the cue validity of various properties of chairs will be determined by contrasting them with properties of sofas, tables, and so on.

[9] Note Rosch's use of the term 'perceived world'. She is not committed to the strong metaphysical assumption that such discontinuities exist 'out there', independently of us as perceivers.

[10] Of course, 'legitimate biological features' already implies some constraints.

[11] One might point out that similarity-based approaches need not make such a strong assumption about the perceived world. While the similarity-based approaches discussed here assume that concepts are stable across occasions and across individuals, not all similarity-based approaches must be committed to this assumption. Instead, they may take classification as strongly influenced by the specific 'stimulus context' at any given moment; that is, the SOP, in effect, may be much more localized. Such a model, then, only needs to assume that the localized SOP is 'carved at the joints', an assumption that may be easier to defend. I suggest, however, that even this more limited requirement is problematic, for reasons similar to those discussed in Section 2.3 for 'simple classification' cases.

[12] One might suggest that various contextual factors (specific goals, background knowledge, and so on) can explain such decisions. As will become clear, I am extremely sympathetic to this solution. My present point, however, is that a theory that is based solely on similarity, narrowly construed, cannot help itself to these resources. It must either appeal to separate mechanisms, distinct from concepts, or confine itself to SOP-effects. Constructs such as 'cue validity' or 'diagnosticity' do the latter.

[13] The term 'cognitive economy' is borrowed from (Rosch [1999a], p. 190).

[14] I do not mean to suggest here that one cannot calculate similarity or form categories without causal knowledge. One can learn, of course, artificial perceptual categories for which causal knowledge is irrelevant (for example, Posner and Keele [1968]). However, while these works provide valuable insight into some aspects of similarity judgements and their role in categorization, they bypass, at least in part, the problem of selection, and

therefore cannot provide us with the full explanation of how selection takes place in 'real world' classification.

[15] As will become clear in Section 4.2, the difficulty is inherited by other theories that take up the same narrow notion of similarity.

[16] One's theory does not need to be elaborate, however. It is not even necessary that one know that genes exist. One only has to have a general sense that there is something causally fundamental in cats, which brings about the other common properties of cats (Medin and Ortony [1989]; Gelman [2005]; Carey [2009]).

[17] Hereafter, I use 'dimensions' to refer to variables (for example, colour), and 'properties' (or 'features' or 'attributes') to refer to specific values (or ranges of values) along these dimensions (for example, red).

[18] I found Rheins's ([2011]) article on the species concept helpful in thinking about this point, and more broadly, about the role of similarity in categories whose members share a causal structure.

[19] One might point out that in this toy example, knowing the target's age would enable one to make a clear-cut classification decision without the need to rely on other, more superficial dimensions. However, one of the appealing features of the theory theory is that it purports to provide us with the mechanism by which we make classifications based on observable features, even when we are unable to observe the most causally fundamental feature in virtue of which particulars belong to the category. In these cases, the theory theory maintains, we use the causally superficial features to infer the presence of the underlying feature (see, for example, Rehder and Kim [2009]). Indeed, without the possibility for such causal inference, the theory theory would apply to a very limited range of categorization decisions. Therefore, in this example, I do not limit the discussion to the most causally fundamental feature.

[20] I do not mean to imply that one has to be able to attribute an exact range of numerical values to this dimension; I say more about this in Section 3.3.

[21] Rheins's definition of similarity is helpful here. He takes similarity to be 'the resemblance relationship that holds between two or more things (the "similars") when their differences in some specific respect(s) are dwarfed by their differences to one or more dissimilar things (the "foil")' (Rheins [2011], p. 255). Accordingly, a concept based on similarity is the 'grasp of a range of possible "measurements" along the relevant axes of variation, as contrasted against all other values that lie outside that range' (p. 256). Beyond this notion of similarity, the present article makes no claims about how similarity is assessed (see discussion in Hahn and Chater [1997]).

[22] This is not to imply that causal considerations are irrelevant for determining the appropriate ranges of values within these dimensions—I address this point below.

[23] One might argue that since lake victoria shrimp is a natural kind while teenager is not, instances of the former but not the latter share an identical essence. But the present point is precisely that while there may be important differences between the two concepts (for example, with respect to the number of generalizations they support), the empirical findings that support the theory theory do not warrant the strong claim that

classifiers believe that there is something identical about all members of lake victoria shrimp.

[24] I thank an anonymous referee for this point.

[25] See, in this context, the importance of the foil for the notion of similarity, discussed in Footnote 21.

[26] As will become clear in Section 4.2, the difficulty is inherited by other theories that take up a similarly flat view of the application of causal knowledge in categorization.

[27] It does not, however, serve as a direct argument against hybridism, pluralism, or eliminativism.

[28] For recent discussions see (Casasanto and Lupyan [2015]; Machery [2015]).

[29] For those who view concepts as constructs created on the fly, localized factors may carry more weight in the model than general ones. For example, SOP may become localized and much closer to the specific 'stimulus context', local goals will be emphasized over long-term cognitive goals, and so on.

[30] See also (Weiskopf [2009]). My aim here is not to evaluate the effectiveness of Rice's approach in responding to Machery, but merely to use Machery's objection to flesh out important differences between Rice's approach and my own.

[31] My view is further distinguished from Rice's in that (i) I do not argue that concepts are constructed on the fly, and (ii) I do not argue for pluralism. A discussion of the first point is outside the scope of this article (but see brief comments at the end of the previous section). With respect to the second point, my own view does not directly argue against pluralism. However, I argue against a specific motivation for pluralism, by suggesting that the new consensus has misconstrued the problem, and I propose that a proper construal of the problem gives rise to an alternative solution.

[32] As an anonymous referee pointed out, concept pluralists may admit my suggested model as one possible construct used in higher cognitive processes. Importantly, however, to the extent that the other constructs they allow require that we take similarity judgements and application of causal knowledge as discrete elements, their view is still vulnerable to the difficulties discussed above.

Acknowledgements

I thank Edouard Machery, Yoon Choi, James Lennox, Peter Machamer, Anthony Peressini, David Danks, and Eva Jablonka for written comments on previous versions of this manuscript. I also thank Patrick Mullins, Jason Rheins, Gregory Salmieri, Audrey Yap, Lina Jansson, Eileen Nutting, Kathryn Lindeman, and participants at the 2015 Meeting of the European Society for Philosophy and Psychology for their feedback on some of the ideas presented here. Last, I thank two anonymous referees who provided extensive and insightful comments that greatly improved the quality of this manuscript.

References

1 Ahn W. K., Kim N. S., Lassaline M. E., Dennis M. J. [2000]: 'Causal Status as a Determinant of Feature Centrality', *Cognitive Psychology*, 41, pp. 361–416.

2 Ahn W. K., Marsh J. K., Luhmann C. C., Lee K. [2002]: ‘Effect of Theory-Based Feature Correlations on Typicality Judgments’, *Memory and Cognition*, 30, pp. 107–18.
3 Anderson J. R., Betz J. [2001]: ‘A Hybrid Model of Categorization’, *Psychonomic Bulletin and Review*, 8, pp. 629–47.
4 Barsalou L. W. [1982]: ‘Context-Independent and Context-Dependent Information in Concepts’, *Memory and Cognition*, 10, pp. 82–93.
5 Barsalou L. W. [1999]: ‘Perceptual Symbol Systems’, *Behavioral and Brain Sciences*, 22, pp. 577–660.
6 Bloch-Mullins C. L. [2015]: ‘Foundational Questions about Concepts: Context-Sensitivity and Embodiment’, *Philosophy Compass*, 10, pp. 940–52.
7 Carey S. E. [1985]: *Conceptual Change in Childhood*, Cambridge, MA: MIT Press.
8 Carey S. E. [2009]: *The Origin of Concepts*, New York: Oxford University Press.
9 Casasanto D., Lupyan G. [2015]: ‘All Concepts Are *Ad Hoc* Concepts’, in Margolis E., Laurence S. (*eds*), *The Conceptual Mind: New Directions in the Study of Concepts*, Cambridge, MA: MIT Press, pp. 543–66.
10 Collins J. A., Olson I. R. [2014]: ‘Knowledge Is Power: How Conceptual Knowledge Transforms Visual Cognition’, *Psychonomic Bulletin and Review*, 21, pp. 843–60.
11 Dougherty J. W. D. [1978]: ‘Salience and Relativity in Classification’, *American Ethnologist*, 5, pp. 66–80.
12 Dove G. [2009]: ‘Beyond Perceptual Symbols: A Call for Representational Pluralism’, *Cognition*, 110, pp. 412–31.
13 Gärdenfors P.[2000]: *Conceptual Spaces: The Geometry of Thought*, Cambridge, MA: MIT Press.
14 Gelman S. A. [2005]: *The Essential Child: Origins of Essentialism in Everyday Thought* , New York: Oxford University Press.
15 Gelman S. A., Markman E. [1986]: ‘Categories and Induction in Young Children’, *Cognition*, 23, pp. 183–209.
16 Gibbs R. W. [2006]: ‘Metaphor Interpretation as Embodied Simulation’, *Mind and Language*, 21, pp. 434–58.
17 Goldstone R. L. [1994]: ‘The Role of Similarity in Categorization: Providing a Groundwork’, *Cognition*, 52, pp. 125–57.
18 Goldstone R. L., Medin D. L., Halberstadt J. [1997]: ‘Similarity in Context’, *Memory and Cognitio*n, 25, pp. 237–55.
19 Goodman N. [1972]: ‘*Seven Strictures on Similarity’, in his Problems and Projects*, Indianapolis: Bobbs-Merrill, pp. 437–47.
20 Griffiths P. E. [1997]: *What Emotions Really Are: The Problem of Psychological Categories*, Chicago, IL: University of Chicago Press.
21 Hahn U., Chater N. [1997]: ‘*Concepts and Similarity’, in Lamberts K., Shanks D. (*eds*), Knowledge, Concepts and Categories*, New York: Psychology Press, pp. 43–92.
22 Hampton J. A. [1979]: ‘Polymorphous Concepts in Semantic Memory’, *Journal of Verbal Learning and Verbal Behavior*, 18, pp. 441–61.
23 Hampton J. A. [1998]: ‘Similarity-Based Categorization and Fuzziness of Natural Categories’, *Cognition* , 65, pp. 137–65.

24 Holland J. H., Holyoak K. J., Nisbett R. E., Thagard P. R. [1989]: *Induction: Processes of Inference, Learning, and Discovery*, Cambridge, MA: MIT Press.
25 Hull D. L. [1988]: *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*, Chicago, IL: University of Chicago Press.
26 Keil F. C. [1989]: *Concepts, Kinds, and Cognitive Development, Cambridge*, MA: MIT Press.
27 King D. L., Atef-Vahid M. K. [1986]: 'Two Extensions of the Anchor-Range Effect', *Perception and Psychophysics*, 39, pp. 96–104.
28 Laurence S., Margolis E. [1999]: 'Concepts and Cognitive Science', in Margolis E., Laurence S. (*eds*), *Concepts: Core Readings, Cambridge*, MA: MIT Press, pp. 3–81.
29 Lombrozo T., Rehder B. [2012]: 'Functions in Biological Kind Classification', *Cognitive Psychology*, 65, pp. 457–85.
30 Machery E. [2007]: 'Concept Empiricism: A Methodological Critique', *Cognition*, 104, pp. 19–46.
31 Machery E. [2009]: *Doing without Concepts*, New York: Oxford University Press.
32 Machery E. [2014]: 'Concepts: Investigating the Heterogeneity Hypothesis', in Sytsma J. (*ed*.), *Advances in Experimental Philosophy of Mind*, London: Bloomsbury Academic, pp. 203–21.
33 Machery E. [2015]: 'By Default: Concepts Are Accessed in a Context-Independent Manner', in Margolis E., Laurence S. (*eds*), *The Conceptual Mind: New Directions in the Study of Concepts*, Cambridge, MA: MIT Press, pp. 567–88.
34 Mahon B. Z. [2015]: 'What Is Embodied about Cognition?', *Language, Cognition, and Neuroscience,* 30, pp. 420–9.
35 Mahon B. Z., Caramazza A. [2008]: 'A Critical Look at the Embodied Cognition Hypothesis and a New Proposal for Grounding Conceptual Content', *Journal of Physiology*, *Paris*, 102, pp. 59–70.
36 Margolis E., Laurence S. [2004]: 'Concepts', in Stitch S. P., Warfield T. A. (*eds*), *The Blackwell Guide to the Philosophy of Mind*, Oxford: Blackwell, pp. 190–213.
37 Mazzone M., Lalumera E. [2010]: 'Concepts: Stored or Created?', *Minds and Machines*, 20, pp. 47–68.
38 Medin D. L. [1989]: 'Concepts and Conceptual Structure', *American Psychologist*, 44, pp. 1469–81.
39 Medin D. L., Goldstone R. L., Gentner D.[1993]: 'Respects for Similarity', *Psychological Review*, 100, pp. 254–78.
40 Medin D. L., Ortony A. [1989]: 'Psychological Essentialism', in Vosniadou S., Ortony A. (*eds*), *Similarity and Analogical Reasoning*, New York: Cambridge University Press, pp. 179–95.
41 Medin D. L., Schaffer M., M. [1978]: 'Context Theory of Classification Learning', *Psychological Review* , 85, pp. 207–38.
42 Murphy G. L. [2002]: *The Big Book of Concepts*, Cambridge, MA: MIT Press.
43 Murphy G. L., Medin D. L. [1985]: 'The Role of Theories in Conceptual Coherence', *Psychological Review*, 92, pp. 289–316.

44 Nosofsky R. M. [1986]: ‘Attention, Similarity, and the Identification-Categorization Relationship’, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 115, pp. 39–57.
45 Nosofsky R. M. [1992]: ‘Exemplar-Based Approach to Relating Categorization, Identification, and Recognition’, in Ashby F. G. (*ed.*), *Multidimensional Models of Perception and Cognition*, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 363–93.
46 Osherson D. N., Smith E. E. [1981]: ‘On the Adequacy of Prototype Theory as a Theory of Concepts’, *Cognition*, 9, pp. 35–58.
47 Piccinini G., Scott S. [2006]: ‘Splitting Concepts’, Philosophy of Science, 73, pp. 390–409.
48 Posner M. I., Keele S. W. [1968]: ‘On the Genesis of Abstract Ideas’, *Journal of Experimental Psychology*, 77, pp. 353–63.
49 Prinz J. J. [2005]: ‘The Emotional Embodiment of Moral Concepts’, in Pecher D., Zwaan R. A. (*eds*), *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking* , New York: Cambridge University Press, pp. 93–114.
50 Rehder B. [2003a]: ‘Categorization as Causal Reasoning’, *Cognitive Science* , 27, pp. 709–48.
51 Rehder B. [2003b]: ‘A Causal-Model Theory of Conceptual Representation and Categorization’, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, pp. 1141–59.
52 Rehder B. [2006]: ‘When Similarity and Causality Compete in Category-Based Property Generalization’, *Memory and Cognition*, 34, pp. 3–16.
53 Rehder B., Hastie R. [2001]: ‘Causal Knowledge and Categories: The Effects of Causal Beliefs on Categorization, Induction, and Similarity’, *Journal of Experimental Psychology: General,* 130, pp. 323–60.
54 Rehder B., Kim S. [2009]: ‘Classification as Diagnostic Reasoning’, *Memory and Cognition*, 37, pp. 715–29.
55 Rheins J. G. [2011]: ‘Similarity and Species Concepts’, in Campbell J. K., O'Rourke M., Slater M. H. (*eds*), *Carving Nature at Its Joints: Natural Kinds in Metaphysics and Science,* Cambridge, MA: MIT Press, pp. 253–88.
56 Rice C. [2014]: ‘Concepts as Pluralistic Hybrids’, *Philosophy and Phenomenological Research*, 92, pp. 597–619.
57 Rips L. J., Shoben E. J., Smith E. E. [1973]: ‘Semantic Distance and the Verification of Semantic Relations’, *Journal of Verbal Learning and Verbal Behavior*, 12, pp. 1–20.
58 Rosch E. [1973]: ‘On the Internal Structure of Perceptual and Semantic Categories’, in Moore T. E. (*ed*.), *Cognitive Development and the Acquisition of Language*, New York: Academic Press, pp. 111–44.
59 Rosch E. [1975a]: ‘Cognitive Representations of Semantic Categories’, *Journal of Experimental Psychology: General*, 104, pp. 192–233.
60 Rosch E. [1975b]: ‘Universals and Cultural Specifics in Human Categorization’ in Brislin R. W., Bochner S., Lonner W. J. (*eds*), *Cross-cultural Perspectives on Learning*, New York: Wiley, pp. 177–206.
61 Rosch E. [1999a]: ‘Principles of Categorization’, in Margolis E., Laurence S. (*eds*), *Concepts: Core Reading*, Cambridge, MA: MIT Press, pp. 189–206.

62 Rosch E. [1999b]: ‘Reclaiming Concepts’, Journal of Consciousness Studies , 6, pp. 61–77.
63 RoschE., Mervis C. B.[1975]: ‘Family Resemblances: Studies in the Internal Structure of Categories’, *Cognitive Psychology*, 7, pp. 573–605.
64 Rosch E., Mervis C. B., Gray W. D., Johnson D. M., Boyes-Braem P. [1976]: ‘Basic Objects in Natural Categories’, *Cognitive Psychology*, 8, pp. 382–439.
65 Sloman S. A., Love B. C., Ahn W. K. [1998]: ‘Feature Centrality and Conceptual Coherence’, *Cognitive Science*, 22, pp. 189–228.
66 Sloman S. A., Rips L. J. [1998]: ‘Similarity as an Explanatory Construct’, *Cognition*, 65, pp. 87–101.
67 Smith E. E. [1989]: ‘Concepts and Induction’, in Posner M. I. (*ed*.), *Foundations of Cognitive Science,* Cambridge, MA: The MIT Press, pp. 501–26.
68 Smith E. E., Medin D. L. [1981]: *Categories and Concepts*, Cambridge, MA: Harvard University Press.
69 Smith E. E., Osherson D. N., Rips L. J., Keane M. [1988]: ‘Combining Prototypes: A Selective Modification Model’, *Cognitive Science*, 12, pp. 485–527.
70 Smith E. E., Shoben E. J., Rips L. J. [1974]: ‘Structure and Process in Semantic Memory: A Featural Model for Semantic Decisions’, *Psychological Review*, 81, pp. 214–41.
71 Smith J. D., Minda J. P. [2002]: ‘Distinguishing Prototype-Based and Exemplar-Based Processes in Category Learning’, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, pp. 800–11.
72 Tanaka J. W., Taylor M. [1991]: ‘Object Categories and Expertise: Is the Basic Level in the Eye of the Beholder?’, *Cognitive Psychology*, 23, pp. 457–82.
73 Torgerson W. S. [1965]: ‘Multidimensional Scaling of Similarity’, *Psychometrika*, 30, pp. 379–93.
74 Tversky A. [1977]: ‘Features of Similarity’, *Psychological Reviews*, 84, pp. 327–52.
75 Tversky A., Gati I. [1978]: ‘Studies of Similarity’, in Rosch E., Lloyd B. L. (*eds*), *Cognition and Categorization*, Hillsdale, NJ: Lawrence Erlbaum, pp. 79–98.
76 Vicente A., Martínez Manrique F. [2016]: ‘The Big Concepts Paper: A Defence of Hybridism’, *British Journal for the Philosophy of Science*, 67, pp. 59–88.
77 Waldmann M. R., Holyoak K. J., Fratianne A. [1995]: ‘Causal Models and the Acquisition of Category Structure’, *Journal of Experimental Psychology: General*, 124, pp. 181–206.
78 Weiskopf D. A. [2009]: ‘The Plurality of Concepts’, Synthese, 169, pp. 145–73.
79 Weiskopf D. A. [2010]: ‘Embodied Cognition and Linguistic Comprehension’, *Studies in History and Philosophy of Science Part A* , 41, pp. 294–304.
80 Wisniewski E. J. [1995]: ‘Prior Knowledge and Functionally Relevant Features in Concept Learning’, *Journal of Experimental Psychology: Learning Memory and Cognition* , 21, pp. 449–68.