

The Myth of the Final Criterion

Michael Strevens
Philosophy Department
Stanford University

Talk given at Psychological Essentialism Conference, University of Oregon,
February 2003

Abstract

Theories of psychological categories have all conformed to the thesis that for every psychological category, there is a *final criterion* that is the ultimate arbiter of category membership. This final criterion not only determines a category's boundaries but also key aspects of its cognitive significance. I propose that the thesis is false: there are no final criteria. I then explore what a theory of concepts might look like in the absence of final criteria. My chief aim is consciousness-raising: rather than presenting arguments against the thesis of the final criterion, I seek to show that some very interesting psychological possibilities lie outside its purview; we should not limit ourselves to theories of concepts that are built according to the criterion.

1. The Final Criterion

Psychological categories are those categories that we impose on the world in virtue of our concepts. They exist independently of what might be called natural categories: they may align with the natural categories well, badly, or not at all. (Then again, perhaps there are no natural categories to align with.)

Much of the literature on concepts, both psychological and philosophical, concerns the way in which our minds construct psychological categories. This is my topic here, and for the remainder of the paper, I will use the term *category* to refer to psychological categories.

Theories of category construction have, I assert, almost always posited the existence of something that I will call a *final criterion*. The final criterion for a psychological category is the final, indefeasible arbiter of category membership. As such, it plays two important roles in the construction of a category:

1. It determines the boundaries of the category, and
2. It determines the property or property complex shared by all members of a category in virtue of their category membership. I call this property complex the *basis* of the category. Some psychologists call it, or say it determines, the structure of the category.

In its first role, the final criterion explains much about what we count as members of a given category. In its second role, it explains important aspects of the category's cognitive significance, such as the kinds of inferences we are prepared to make about members of the category in virtue of their membership.

The *final criterion thesis* is the view that, behind every psychological category, there is a final criterion. I propose that this thesis is false. Most or all of our psychological categories have no final criterion. If I am right, then theories of concepts that posit a final criterion, such as the classical theory of concepts and psychological essentialism, are wrong.

It is far from obvious, however, that we could succeed in constructing psychological categories without final criteria. Most of this paper is an effort to show how it can be done, and thus, how the thesis of the final criterion could be false. The result is, I suggest, an approach to the psychology of concepts that has all the virtues of its predecessors and none of its vices. I urge its adoption.

2. The History of the Final Criterion Thesis

Rather than try to explain further what I mean by a final criterion in abstract terms, let me discuss briefly three influential theories of concepts—the classical view, the prototype view, and the essentialist view—so as to show how, despite their differences, they harbor a common commitment to the final criterion thesis. In each case I will pay special attention to the theories' treatment of "folk genus" concepts, such as TIGER and SWAN, and chemical substance concepts, such as GOLD and WATER.

2.1 The Classical Theory

A classical theory of concepts asserts that, for each of our psychological categories, we represent a definition. Our concept BACHELOR may have the definition *unmarried male*; the concept ISLAND the definition *body of land surrounded by water on all sides*. A specimen falls into a category, on this theory, just in case it satisfies the corresponding definition.

The classical view of psychological categories is, in many ways, the most direct realization of the final criterion thesis. The final criterion is, of course,

the definition. Its finality is indicated by its exceptionless as a criterion: nothing that fails to satisfy the definition can count as a category member, and nothing that satisfies the definition can fail to count as a category member. The classical view allows for human error, of course: a person may be mistaken in thinking that a certain specimen satisfies their definition. But the failure is always in the comparison between specimen and definition, not in the definition itself.

Typically, classical theories of concepts hold that concepts simply *are* definitions. Thus, the structure of the final criterion is the structure of the concept. This is what is meant, I think, when it is said that final criteria such as definitions or prototypes determine the structure of the corresponding psychological categories.

A paradigm of the classical approach is John Locke's theory of natural kind terms such as *gold* and *swan*. According to Locke, the concepts corresponding to these terms are literally built out of the concepts corresponding to the characteristic observable properties of the kinds. For example, the concept of gold is composed of the concepts YELLOW, METALLIC, MALLEABLE and so on. This conjunction of observable properties serves, in Locke's theory, as a definition of gold, since to think that something is gold just is to think that it is yellow, metallic, malleable, and so on. Locke's final criterion for natural kinds is, then, a definition couched in terms of the observable properties of the kinds.

Locke called this definition the *nominal essence* of the kind; by this he meant just what I mean by the categorical basis of a kind, that is, the property in virtue of which specimens do or do not belong to the psychological

category corresponding to the kind. He also posited that kinds have *real essences*: these determine the boundaries of objective categories in nature. Because, according to Locke, the nominal essences and real essences differ our psychological categories do not quite align with the natural categories.

2.2 The Prototype Theory

According to the prototype theory, psychological categories do not have definitions. Furthermore, they have vague boundaries. You might therefore suppose that they have no associated final criteria—wrongly, as it will turn out.

The prototype theory holds that every concept is associated with a *prototype*. Let me explain how this works in a relatively simple version of the prototype theory. A prototype is a cluster of characteristic observable properties which are used to classify specimens as follows. A specimen is considered a member of a category if it has sufficiently many of the features in the prototype. The prototype for a tiger, for example, may consist of the properties of ferocity, stripedness, having fearful symmetry, and so on. An animal would be counted as a tiger if it had enough of these properties; there is no one of them, however, that is absolutely essential for category membership.

The prototype theory differs from the classical theory in two obvious ways. First, there may be no particular property that all members of the category have in common. Second, though it is not obvious from the characterization given here, it may be unclear whether a particular specimen has sufficiently many features to belong to a category. Such a specimen is a *borderline case*.

Does a prototype constitute a final criterion? There is nothing about prototypes themselves that *demand*s that they be regarded as final criteria. This is why, for example, Rey (1983) can suggest that prototype-driven categorization is a heuristic procedure that serves an essentialist final criterion. It follows that the prototype theory is not inherently committed to the final criterion thesis, unlike the classical theory and, as we will see, the essentialist theory. Nevertheless, prototypes are almost always interpreted as specifying an ultimate basis for psychological categories.

I offer two pieces of evidence for this claim. First, advocates of the prototype theory are quite explicit that according to the theory, the prototype defines and determines the structure of the corresponding psychological category. Rosch (1978) writes, for example, that “categories tend to become defined in terms of prototypes” and frequently says that prototypes determine category structure. This is also the way that the theory is presented in standard reviews of the literature (Smith and Medin 1981; Margolis and Laurence 1999).

Second, the existence of borderline cases is always adduced to show that, if prototype theory is correct, then psychological categories are thereby vague. This inference follows only if prototypes are the last word on category membership, so that what is not definitely categorized by a prototype is not definitely a category member at all. (We see here that a test can be interpreted as a final criterion even if, in some cases, it does not deliver a definite verdict.)

If there is no single property that all members of a category, according to prototype theory, share, then what is the basis of the category? A basis is, after

all, supposed to be just the shared property that underlies category membership. A moment's thought shows that the shared property is just similarity to the prototype, or as some writers say, a certain family resemblance.

2.3 The Essentialist Theory

Recall from the discussion of the classical theory Locke's distinction between nominal essence and real essence: the nominal essence is the distinguishing mark of the members of a psychological category, while the real essence is the distinguishing mark of the members of the corresponding objective category. The essentialist theory of concepts claims that nominal essences are real essences, that is, that psychological categories are delineated by real essences.

On the face of it, this seems impossible. Our psychological categories ought to be given a basis that is represented somewhere in the mind, but the features that determine the nature of real categories are often beyond our grasp. The psychological essentialists concede that many people are largely ignorant of the nature of real essences,¹ but hold that it is enough for a person to have the *intention* of using nature's categories. What is in the mind is not a representation of the real essence itself, but rather a commitment to use the real essence to determine category membership as far as possible.

Our categorizing behavior, then, deploys whatever facts have come to light about real essences, and otherwise tries to allow for the existence of the other, as yet unknown, facts. As some writers say, humans for the most part

1. Or even that there are no real essences.

do not represent essences (meaning real essences) but rather *essence placeholders*, where an essence placeholder is some kind of (perhaps inchoate) theory about the nature of the relevant essences, along with the fact that

there are people, experts, who really know what makes the thing the thing it is, or scholars who are trying to figure out exactly what it is (Medin and Ortony 1989, 185).

This placeholder is eventually replaced by a representation of the nature of the essence, if such knowledge is acquired.

The essentialist picture becomes clearer in the light of an example, the essentialist theory of concepts of biological taxa and chemical substances. For a concept such as TIGER a paradigmatic essentialist theory makes the following claims:

1. Humans posit the existence of an essence corresponding to the category of tigers, though they may not be able to represent what property or properties constitute that essence.
2. Possession of the essence is represented as necessary and sufficient for category membership. It is believed that an organism is a tiger if and only if it has the tiger essence.
3. Essences are represented as causally responsible for certain observable properties: the tiger essence causes the tiger's stripes, ferocity, fearful symmetry, and so on.

The existence of a tiger essence is posited, then, but the essence is identified entirely by its causal connection to tigers' characteristic observable properties. It is quite possible that a possessor of the tiger concept has no beliefs

whatsoever about the nature of the tiger essence. Or perhaps they have a few, rather vague beliefs: that the essence is an internal property of the tiger, that it has something to do with genetic material, and so on. Experts, of course, have a much richer set of beliefs about the essence, but even non-experts have concepts structured around the posit of an essence.

Now the crucial question: is essentialism committed to a final criterion? Yes; the final criterion is the real essence. In this respect, essentialism offers an account of the final criterion that departs in an innovative and unexpected way from previous theories: the final criterion is no longer explicitly represented in the head. Rather, the nature of the criterion is determined by extra-psychological factors. The head does, however, commit itself to the *finality* of the criterion. It is our determination to use real essences as the final arbiters of category membership, then, that makes real essences final criteria for membership of psychological categories.

3. Remarks on the Final Criterion Thesis

The paradigmatic final criterion is perhaps a mental definition, just as the classical theory would have it. But there are other ways to have a final criterion. As the prototype theory shows, a final criterion need not take the form of an explicit set of conditions, nor need it make a clear judgment about every specimen. As the essentialist theory shows, a final criterion need not even be known by the categorizer, provided that the categorizer espouses a commitment to its finality.

For all these differences, theories of concepts that endorse the final cri-

terion thesis share two deep assumptions, one about categorization and one about inference.

First, all categorization is driven by the final criterion. This is, of course, in some sense true by the very definition of finality. But theories of concepts tend to posit a very active role for the final criterion: it is not only the criterion of last resort, but is deployed day to day. For the classical and prototype theories, this seems especially true: virtually all classification is carried out using the relevant definition or prototype. For the essentialist theory, it is perhaps less true. Some essentialists allow that a prototype, or something like it, may be used in a heuristic capacity for run-of-the-mill classification (Rey 1983). Still, the essentialist final criterion is kept close at hand, and is used even by fairly young children for specimens that are out of the ordinary (Keil 1989).

Second, other kinds of inference are also deeply affected by the nature of the final criterion. Because the final criterion determines not just category boundaries but also the category basis—the property or complex of properties that all members of a category must share—the form of the inferences that can be made on the strength of category membership will be shaped by the form of the final criterion itself. Such inferences in effect have the form: x is a member of the category K , therefore x has properties P (determined by the final criterion), therefore x has Q , where possession of Q follows from possession of P .

4. What's Wrong with the Final Criterion Thesis

It is not an aim of this paper to refute the final criterion thesis; however, it seems reasonable, before I go on, to point out its principal weakness. The weakness is simply that we do not categorize in accordance with any known final criterion.

The case against the classical Lockean criterion is well known, that against the prototype criterion almost as well known, and that against the essentialist criterion is quite controversial. Let me briefly recap each.

The final criterion posited by the Lockean account makes the wrong prediction about our classification of *uncharacteristic specimens*. An uncharacteristic specimen lacks one or more typical features of its category: green lemons and three-legged tigers are examples. Because, on the classical Lockean account, a specimen must possess all of a category's characteristic properties if it is to fall within the category, uncharacteristic specimens ought not, if the classical account is correct, to be classified as members of the corresponding categories. Green lemons ought not to be counted as lemons; three-legged tigers ought not to be counted as tigers. But of course, as the expressions *green lemon* and *three-legged tiger* show, we do allow for the possibility of uncharacteristic specimens. Thus classical Lockean definitions cannot be the final arbiters of category membership.²

The final criterion posited by the prototype account makes the wrong

2. This does not, of course, show that what we count as lemons or tigers do not all satisfy some other mental definition; the classical approach is therefore not definitively refuted, only a version of that approach, the Lockean account.

prediction about our classification of *radically uncharacteristic specimens*. A radically uncharacteristic specimen lacks most or all of the typical features of its category. Frank Keil has provided a nice paradigm for generating uncharacteristic specimens. Keil imagines a raccoon that has been cosmetically transformed to resemble a skunk. The made-over raccoon looks just like a skunk and smells like a skunk. If desired, we can imagine that it also behaves like a skunk, thanks to training or other behavioral manipulation. The transformed animal now has few if any of the features normally supposed to be in the raccoon prototype, and indeed strongly resembles the skunk prototype. Yet, as Keil demonstrates, we classify it as a raccoon, not a skunk. The prototype account's final criterion cannot be the last word on what it takes to be a raccoon.

The essentialist account is often touted, not without reason, as the solution to these problems. On the essentialist story, uncharacteristic specimens, even radically uncharacteristic specimens, retain the underlying essence of the relevant kind—a transformed raccoon, for example, still has the raccoon essence—and so count as members of the kind.

But the essentialist final criterion has its own problems, as Malt (1994) has shown. Essentialists normally assume that the essence of chemical kinds has something to do with their molecular structure. The essence of water, for example, is H_2O . A fully informed individual will count a substance as water, then, just in case it is composed of H_2O , or at least is largely composed of H_2O .

Malt's findings cast doubt on this view. She shows that there is only a weak correlation between people's judgments of a specimen's H_2O content

and their judgments of whether something is water, weaker, at least, than predicted by essentialism. Her conclusion is that possession of the “essential property”, being composed of H₂O, cannot be the final criterion for membership of the category of water, and she posits a final criterion that takes into account various properties of the specimen such as taste and appearance (though without explaining the Keil results).

Essentialists have occasionally tried to deflect this problem by positing two concepts of water, an expert’s concept and an everyday concept. Even if conceptual dualism is reasonable, however, this does not get essentialists what they want. It is the expert’s concept, on their view, that has as its final criterion H₂O content. But what was wanted was an account of the everyday concept; essentialists have so far not offered an essentialist final criterion for waterhood that explains Malt’s results.³

The reader may conclude that we have yet to discover the form of the final criterion that lies behind the psychological categories corresponding to biological taxa and chemical kinds. Maybe so. There is, however, another intriguing possibility, that there is no final criterion. I wish to show how the psychology of concepts and categories might look under such an assumption.

3. Meanwhile, I might add, philosophers of science have expressed serious doubts about the view that experts’ concepts have an essentialist final criterion (Mellor 1977; Weisberg 2003).

5. Categorization without a Final Criterion

The role of a final criterion is to determine what does and does not fall into a psychological category. How can psychological categories exist, without such a standard? And if there are no psychological categories, what could we be doing when we categorize?

Let me begin to answer this question with an observation: a categorization is an inference that has a conclusion of the form

x is a K ,

where x is some specimen and K is a category. It is possible to draw a conclusion of this form without using a final criterion; all that is needed is some appropriate knowledge about x and K . For example, if I know from a guidebook that I am passing through a grove of California-nutmegs, and I know that California-nutmegs have short, spiky needles, then upon observing a tree x with short spiky needles, I may well infer that x is a California-nutmeg, that is, I may well categorize x as a California-nutmeg. The facts about California-nutmegs used to make the inference are

1. The trees around here are California-nutmegs, and
2. California-nutmegs have short, spiky needles.

By no one's lights do these two facts constitute a final criterion for membership of the class of California-nutmegs; the example shows, then, that categorization can proceed without the use of a final criterion.

Readers will quite reasonably object that, although the inference does not make explicit use of a final criterion, it presupposes a final criterion's

existence. I trust the guidebook, for example, because I believe that the author is in possession of the final criterion, and has used it to verify that the trees in the grove are all California-nutmegs. Or if the author is not an expert, then I suppose that they have consulted an expert. My assumption that the guidebook is reliable, then, presumes that someone, somewhere is an expert, and so can tell for sure whether or not a tree is a California-nutmeg.

But this last train of reasoning is fallacious. My trust in the guide book presumes that someone, somewhere is an expert, but it is possible to be an expert—to categorize with great reliability—yet not to possess a final criterion.

To see this, suppose that the essentialist theory is true, thus, that the final criterion for being a California-nutmeg is possession of some deep property, let's say a configuration of DNA. For thousands of years people knew nothing about DNA. But there were experts about California-nutmegs all the same. These experts were able to identify California-nutmegs accurately because they possessed a large amount of information about the appearance and habits of California-nutmegs. We are assuming that essentialism is true, remember, and therefore that this information does not constitute a final criterion. It does not need to: what is important is that it constitutes a body of facts about California-nutmegs that is sufficient to distinguish the California-nutmegs from all the other trees. It is the informativeness of the facts—their inductive richness—that confers expertise on their possessors, then, not any special status the facts might have as final arbiters. This is not to say that experts will not become even more expert by possessing the final criterion, just that they need not know it to qualify as experts.

On the essentialist view, then, the experts of bygone times were not unlike me in my grove of California-nutmegs. I rely on two pieces of information, a guidebook's map of California-nutmeg groves, and a fact about California-nutmegs' leaves. The experts have no map (their job is to draw the map), but they know many more facts than I do about California-nutmegs' appearances. What we have in common is that our categorization—our conclusion that a certain tree is a California-nutmeg—is not based on a definitive test, but is a result of inductive reasoning that leads us to the conclusion that the tree is overwhelmingly likely to be a California-nutmeg.

In this respect, our conclusion that the tree has the property of being a California-nutmeg is rather like the conclusions we might draw about other properties of the tree: that its nuts would be good to eat, that it would not give much shade in summer, that it would look good in the shady corner of the back yard. In each case, we make an inductive inference from a set of pertinent facts to the tree's having a certain property. What makes the inference about the tree's species seem, unlike the others, a paradigmatic *categorization* is the property—specieshood—that appears in the conclusion, not the nature of the reasoning by which the conclusion was reached.

The viability of the essentialist theory shows, then, that categorization can be both widespread and successful in the absence of any known final criterion. Of course, the essentialist account asserts the existence of a final criterion, but denies that the criterion plays an active role in categorization, at least in pre-scientific days. The success of categorizing practices in the distant past can be put down to, at most, human faith in the existence of a final criterion.

I propose that we take one further step. Human categorization neither requires nor supposes the existence of a final criterion. Our sole tools for placing specimens into categories are, on this view, an abundance of beliefs about those specimens and categories, none of which constitute, or are regarded as constituting, a final criterion for category membership.

To illustrate how such a practice of categorization might work, let me sketch a view of psychological categories developed in Strevens (2000). The view, which applies to categories corresponding to “folk genus” and chemical substance concepts, is called *causal minimalism*.

Causal minimalism holds that typically associated with any low level biological or chemical category are a number of causal laws connecting membership of the genus with various observable properties. For example, concerning the category of tigers, we might believe the following laws:

1. All tigers are striped,
2. All tigers are ferocious,
3. All tigers have fearful symmetry.

Because these are causal laws, they should be interpreted as saying that there is something about tigers that *causes* them to be striped, something that *causes* them to be ferocious, and so on.

Note that the essentialist also posits causal laws of this sort, although the essentialist laws explicitly mention tiger essence. Otherwise, essentialism and causal minimalism treat the laws in the same way in two important respects. First, on both views, the characteristic properties mentioned in the

laws do not make up any kind of final criterion. Thus the laws may have exceptions, so that, for example, some tigers are not striped. (The tigers in Frank Keil's lab that are altered to look like lions are an example.)

Second, the laws themselves do not constitute a final criterion: it is in no sense considered a necessary truth that tigers obey just the three laws stated above. It is allowed that the laws may be false, or even that they may be true now but may change later. The point is that there is nothing about these laws that binds them more tightly to the tiger concept than any other belief about tigers is bound to the tiger concept. If we are more reluctant to abandon our belief in the laws than to abandon, say, the belief that there are tigers in the San Francisco zoo, it is because we have more evidence for the laws than for the presence of tigers in the zoo. In short, the laws are considered to be just a certain class of matters of fact about tigers, albeit an inductively very rich kind.

How does causal minimalism explain categorization? Suppose that we see an animal at the San Francisco zoo. It is striped, ferocious, and its symmetry is simply spine-tingling. What might it be? We know that an animal's being a tiger will cause it to have all these properties. We also know, or most of us do, that there is nothing else in the vicinity that will result in an animal's having the same complex of properties. (Frank Keil is taking the day off.) Thus we infer that the properties were very likely caused by the animal's being a tiger, and so that the animal is a tiger. The inference, then, has the form:

1. x has properties P ,

2. Membership of category K causes properties P ,
3. Nothing else in the vicinity is likely to cause properties P , therefore
4. What likely caused x to have P is membership of K , and so
5. x is a member of K .

No final criterion plays a role in this categorizing inference.

Where I differ from essentialism and all the other theories mentioned in this paper is in holding that final criteria play a role in *nobody's* inferences. All categorizations, even the experts' categorizations, take the above form, or the form of some other purely inductive argument. What marks out the experts is that they know more and more accurate causal laws.

Can this be right? Is it not the case that some of today's experts claim to know the final criteria for various categories, and to implement those criteria in determining category membership? Don't chemists claim that to be water just is to be H_2O , and don't they hold that their tests for water are based on this knowledge?

Some of them do make these claims. My counter-claim is that they make them falsely. Although they may define water as H_2O , they do not act as though water is so defined. To establish this point is a long and difficult undertaking, and I will say no more about it here.

Let me conclude with a few remarks about the features of categorization on the minimalist account. First, on the minimalist theory, there may be borderline cases or cases in which the facts about category membership are forever unclear. This is because our beliefs about a specimen or a category

are not sufficient to warrant the conclusion either that the specimen belongs to the category or that it does not. But this phenomenon should not upset us; it is a feature, after all, of categorization according to the prototype theory.

Second, on the minimalist theory, categorization goes on without any application of or even knowledge of a final criterion. This phenomenon should not upset us either; it is a feature, after all, of categorization according to the essentialist theory.

Third, on the minimalist theory, categorization goes on without the existence of a final criterion. This phenomenon, perhaps, should upset us. I have explained how there can be experts, that is, reliable classifiers of things, who make no use of a final criterion. But does it make sense to talk about experts if no final criterion exists at all? How can we talk about the reliability of an expert if there is no standard that determines whether or not an expert's classifications are correct? What makes the experts' categorizations right? What makes their judgments any better than the idiots'?

The answer is that the experts' categorizations, unlike the idiots', are based on true facts, for example, on true causal laws, about the categories in question. Very well then, but what makes the laws true? That is a difficult question; the best I can say here is that the philosophers are working on it.

6. Lessons for the Study of Concepts

In my final section, let me turn to a more practical topic, that of the ways in which an overriding concern for final criteria has distorted psychological work on concepts. I have in mind not only those aspects of theories of con-

cepts that explicitly presuppose a final criterion, but also aspects that I believe arise from an implicit commitment to the dominant explanatory role of final criteria.

All of the following errors can be seen, I think, as issuing from a failure to notice a certain possibility, that important facets of categorization and inference might not depend very much or at all on a final criterion.

The first error is, of course, to assume that a theory of concepts must be, in the first instance, a theory of category-determining criteria. As I have explained at length above, all major theories of concepts have made this mistake.

Future theories of concepts seem headed the same way. For example, Rips (2001) proposes what he calls an interaction-based account of categories, which emphasizes the role played by our beliefs about a natural kind's characteristic causal properties in determining the characteristic inferential patterns associated with the kind. His view of categories is in many ways congenial to my own causal minimalist project.

Yet Rips insists on moving from the overriding psychological importance of causal properties to the proposal that category structure is *determined by* causal properties. The resulting account is similar to the prototype theory: a specimen is a member of a category if it instantiates sufficiently many of the causal laws characterizing the kind, or is sufficiently similar in its causal properties to the prototypical members of the kind.⁴ Rips' "prototypes" embody

4. "An object's membership in a natural kind depends on whether the object instantiates the laws for that kind" (p. 846).

the deep causal properties, rather than the shallow observable properties, that jointly contribute to category membership, but they constitute a final criterion all the same.

As a result, Rips has some difficulty in explaining our classification of organisms that have undergone Keil-like transformations. He notes that, on his view, violating a “small subset” of the relevant laws will not invalidate category membership (p. 848). But the Keil victims violate more than a small subset; they violate all the causal laws that play a role in our everyday inferences. Rips would do better simply to let go of his final criterion altogether.

The second error is to suppose that inductively rich beliefs are indicative of conceptual structure. Ahn (1998), for example, argues that, because information about causal structure strongly influences our categorization decisions, causal facts must appear in the corresponding final criteria.

The influence of causal information does indeed show that causal facts play a central role in inferring category membership. But this is compatible with those same causal facts not appearing in the final criterion (as in the variant of essentialism that I call *pure essentialism* in Strevens (2000)) and, as causal minimalism shows, with there being no final criterion at all. Indeed, as I show in Strevens (2001), causal minimalism is not only consistent with, but predicts Ahn’s results.

Another perpetrator of this error is, I suspect, Locke himself. Noticing that a natural kind’s characteristic observable properties play a key role in categorization, Locke infers that the kind must have a final criterion formulated in terms of those same properties. (Of course, Locke’s view was also strongly motivated by his empiricist theory of concepts.)

Another example of this error is the prototype theorists' doctrine that psychological categories have fuzzy boundaries, a conclusion they draw from the existence of unclear cases, that is, specimens that are judged not to fall clearly into any category. Prototype theorists, taking these judgments as the dictates of a final criterion, suppose that where there is no clear judgment, the psychological category itself has no clear boundary. If all categorizations are regarded as mere inductive inferences, however, then this conclusion does not follow. It may be that our inductive techniques and our knowledge base simply lack sufficient resolving power to discern the truth about the unclear cases.

The third error—the converse of the second—is to think that characteristic patterns of inference concerning a class of categories, such as biological taxa or chemical substances, must be explained by the form of the categories' final criteria.

This error is pervasive, I believe, in the essentialist literature. Consider, for example, Keil's transformation experiments, in which subjects judge that a raccoon made up to look like a skunk is still a raccoon. Essentialists assume that it is the fact that Keil's subjects posit a raccoon essence that plays the pre-eminent role in the explanation of Keil's results. As I have shown elsewhere (Strevens 2000), this is not correct even by the essentialists' own lights: it is the causal beliefs that go along with essentialism—the same causal beliefs attributed to the reasoner by causal minimalism—that do most or all of the explaining.

I contend, then, that essentialists misunderstand the explanatory resources of their own theory, because of a commitment to the idea that it is the nature

of the final criterion that is responsible for all the most distinctive features of a psychological category.

The fourth error is to suppose that a kind of master metaphysical theory is required for concept possession. This supposition is in evidence in Susan Carey's view that children do not possess natural kind concepts until they possess the concept of a biological essence (Carey 1985). Why should this be? Because, Carey supposes, following the essentialist line, that the final criterion for a natural kind concept involves the concept of an essence. To acquire a natural kind concept, then, one must represent the final criterion and so one must acquire a metaphysical theory of essences.

The idea that final criteria not only exist, but must be fully theorized, leads Carey to the view that an explicitly metaphysical theory is required to disentangle categories from their characteristic observable properties. In a discussion of Keil's transformation experiments, Carey writes:

Five-year-olds have no notion of biological essence . . . Therefore, their concepts of fruits, plants, and animals must be exhausted by their knowledge of the characteristics by which they are recognized . . . (p. 180).

There are only two possibilities, on this view: an empiricist psychology, whether in the style of Locke or Rosch, or a psychology based on deep ontology. But this is a false dichotomy: causal minimalism shows that there is a third way.

The fifth and final error, closely related to the second, is to attribute properties of inductive reasoning to the final criterion itself. One property of inductive reasoning is sensitivity to context: I am more likely to judge that a

certain spiky-leafed tree is a California-nutmeg, for example, if I know that I am in a grove of California-nutmegs than if I am well outside their normal range.

Some writers, in particular those sympathetic to the prototype theory, have inferred from the effect of context on categorization that the criteria for membership of psychological categories are themselves context-sensitive—that what counts as a member of a given category depends on the context.

For example, Sloman et al. (2002) showed that whether an object was inferred to be a pen or a marker depended on the shape of other pens and markers that subjects had seen as part of the same study. This result can be explained as a “grove” effect: what the subjects infer from the various pens and markers they see in the course of the study is what pens and markers tend to look like “around here”. This information is then used to guess, in an educated way, whether a particular ambiguously shaped writing tool in the same locale is itself a pen or a marker.

Sloman and Malt (2003) propose a far more radical interpretation of this and similar results:

Theories of artifact categorization must contend with the fact that artifact categories are not stable, but rather depend on the categorization task at hand (article abstract).

The final criteria for artifact categories, Sloman and Malt seem to be saying, change with the context. Such a strong conclusion is hardly justified by the evidence. Once we realize that categorizations are regular inductive inferences, we see that context-sensitivity, that is, sensitivity to local information, may not tell us anything about final criteria—not even that they exist.

References

- Ahn, W. (1998). Why are different features central for natural kinds and artifacts?: The role of causal status in determining feature centrality. *Cognition* 69:135–178.
- Carey, S. (1985). *Conceptual Change In Childhood*. MIT Press, Cambridge, MA.
- Keil, F. (1989). *Concepts, Kinds and Conceptual Development*. MIT Press, Cambridge, MA.
- Malt, B. (1994). Water is not H₂O. *Cognitive Psychology* 27:41–70.
- Margolis, E. and S. Laurence. (1999). Introduction to *Concepts: Core Readings*. MIT Press, Cambridge, MA.
- Medin, D. and A. Ortony. (1989). Psychological essentialism. In S. Vosniadou and A. Ortony (eds.), *Similarity and Analogical Reasoning*, pp. 179–195. Cambridge University Press, Cambridge.
- Mellor, D. H. (1977). Natural kinds. *British Journal for the Philosophy of Science* 28:299–312.
- Rey, G. (1983). Concepts and stereotypes. *Cognition* 15:237–262.
- Rips, L. J. (2001). Necessity and natural categories. *Psychological Bulletin* 127:827–852.
- Rosch, E. (1978). Principles of categorization. In E. Rosch and B. Lloyd (eds.), *Cognition and Categorization*. Lawrence Erlbaum, Hillsdale, NJ.
- Sloman, S. A., M. C. Harrison, and B. C. Malt. (2002). Recent exposure affects artifact naming. *Memory and Cognition* 30:687–695.
- Sloman, S. A. and B. C. Malt. (2003). Artifacts are not ascribed essences, nor are they treated as belonging to kinds. Forthcoming, *Language and Cognitive Processes*.

Smith, E. and D. Medin. (1981). *Categories and Concepts*. Harvard University Press, Cambridge, MA.

Strevens, M. (2000). The essentialist aspect of naive theories. *Cognition* 74:149–175.

———. (2001). Further comments on Ahn et al. Electronic document available at <http://www.stanford.edu/~strevens/research/cogsci/ahnetal/ahnetal.pdf>.

Weisberg, M. (2003). Water is not H₂O! Forthcoming in *Philosophy of Chemistry: Synthesis of a New Discipline*, Kluwer.