The Myth of Depth:
Toward a Shallow Theory of Concepts

Michael Strevens

Draft of March 2010

Abstract

By a "deep" theory of concepts, I mean a theory according to which a concept's possessor represents the grounds of membership in the corresponding category, that is, the properties in virtue of which entities qualify for membership of the category. This paper has three aims. The first is to show that there is an interesting psychological distinction to be made between "deep" and "shallow" approaches to concepts. The second is to show that past theories of concepts have tended, on the whole, to take the deep approach. The third is to show that there is very little evidence that concepts are in fact deep, and so that it is worth exploring shallow alternatives to the existing deep theories.

We get by in the world by imposing upon its motley inventory of being a scheme of psychological categories—that is, by using our concepts to carve up the things in the world into classes that provide some basis both for further learning and for practical action. This sorting of things into man-made classes is my topic here, and for the remainder of the paper, I will use the term *category* to refer to psychological categories rather than to natural categories "out there", and *concept* as though all concepts were concepts of categories rather than concepts of properties, individuals, and so on—though of course, we are quite capable of thinking of things one-by-one as well as en masse, and we have singular concepts to suit (Rips et al. 2006).

Work on concepts and psychological categories is hobbled, I will contend, by its subscribing to a certain precept about the relation between concepts and categories:

> A concept carves out a psychological category by representing the grounds of membership in that category.

This is what I call the *depth thesis*, because it posits knowledge of the "deep structure" of a category, that is, knowledge of what it is at bottom that makes things members of the category. As my title implies, I regard the depth thesis with suspicion—it is, at the very least, a thesis whose popularity far outstrips the weight of evidence in its favor.

How could you possibly construct a new psychological category and go on to decide what is and is not a member of that category without representing grounds for category membership? How could the depth thesis be false? This question may be divided into two parts. How can a category be deployed in cognition—in categorization, induction, and so on—if the depth thesis is false? And how can a category be constructed in the first place if the depth thesis is false?

The answer to the first question, the question of category use, begins with the observation that you can categorize an object without invoking the ultimate grounds for category membership—by using a heuristic. Rejection

of the depth thesis means commitment to the view that in the case of the psychological categories, it is heuristics all the way down. It is this *shallow*—as opposed to deep—view of the nature of concepts that I will defend in this paper.

The idea that we could represent only heuristics for category membership is easy to accept if the categories in question are *natural* categories: if the grounds for category membership are facts out there in the world, it is easy to see that we might not have full knowledge of these facts, even if we know enough about the corresponding categories to make tentative classifications. It is less easy to see that we could represent only heuristics for membership of a *psychological* category, that is, a category that we have ourselves constructed—for how is a category built if not by designating some shared property or other criterion as the basis for category membership?

That is of course to pose the second question about depth, the question of category construction. Much of this paper will be devoted to investigating possible answers; I will not preview my conclusions here.

Before I continue, however, let me briefly advertise what I take to be the principal advantages of abandoning the depth thesis. They are four: first, accepting a shallow theory of concepts leads to new ideas about concept acquisition and category construction; second, it likewise leads to new ideas about the process of categorization; third, it explains how new knowledge about a category can be smoothly incorporated into the existing concept; fourth, and consequently, it explains how a prototype and a theory can be two aspects of the same concept, and so how the prototype theory and "theory theory" of concepts can be reconciled.

1. Deep Category Construction

So pervasive is the depth thesis that it is easily entirely overlooked—you may not realize that you make such an assumption, or that it constitutes a

substantial theoretical commitment. I will take some care, then, to introduce the thesis, to show how it figures in some of the major theories of concepts (in this section), and to show that it has considerable empirical significance (in the next section).

The depth thesis is firmly embedded in the psychology of concepts above all, I think, because it is entailed by an assumption about category construction that is shared by all the major approaches to thinking about concepts, at least in their canonical versions. I will introduce the thesis in its various guises, then, by discussing category construction in several important theories of concepts.

My focus is on "unsupervised learning" rather than supervised learning, or as Murphy (2002, chap. 5) puts it more revealingly, on concept construction rather than concept learning—on the process in which a person constructs a novel category to better organize newly acquired information about the world, rather than the process in which a person attempts to construct a category matching that of the users of some unfamiliar word.

As you will see, there are more similarities than differences between the accounts of construction assumed by the classical theory of concepts, prototype theories, and various forms of the "theory theory" such as psychological essentialism. In each case, a new concept is created by way of what I call a *stipulation*, a decision to class together under a new mental rubric a set of objects sharing some particular set of properties, a certain "family resemblance," a distinctive causal structure, or whatever. The stipulation lays out the grounds for membership of the psychological category so constructed.

A caveat: much valuable experimental work on concepts does not endorse the depth thesis, simply because it shies away from significant theoretical commitments altogether. The comments that follow apply only to large-scale theorizing about concepts, the sort of work that attempts to delineate the fundamental properties of the processes by which psychological categories are created and deployed.

And another caveat: the following discussion of theories of concepts will be, for reasons of space, rather selective. There is no mention of exemplar theories, and the discussion of the "theory theory" is focused on psychological essentialism and its variants.

1.1 The Classical Account According to the classical account, a concept is built around a definition that lists a set of properties possession of which is necessary and sufficient for category membership. The definition, then, sets out in the most forthright way the grounds for category membership; thus, the classical account is committed to the depth thesis—and it is in many ways the canonical form of the thesis, the archetype of the myth of depth.

Concept construction on the classical account is, clearly, a matter of formulating a definition, presumably in response to the observation of the regular co-occurrence a cluster of properties. On observing, for example, a number of white aquatic birds with red beaks that make trumpeting calls, and that do not fit into any existing category, a cognizer might as it were say to themselves: "Let me call a bird with these properties a *swan*," where *swan* is a coinage in their language in thought (or perhaps in their spoken language). Through this stipulative act, then, the cognizer introduces a new representation—a new mental predicate—and gives it cognitive significance by way of a definition. Through the same act they construct a new psychological category, the category of swans, containing all and only those things that satisfy the definition.[1]

Generalizing, the process of classical concept acquisition may be divided into three stages:

1. A certain pattern is observed in the world: the observation of a number of entities (say, birds) all sharing a certain aspect. If the entities do not fit easily into some pre-existing category, the conditions are ripe for concept acquisition; however, no concept is yet created. The first

1. Locke (1975), II:23, §14, p. 305

stage, then, is simply a matter of observation and (perhaps) inductive learning.

2. A new mental predicate is created. How this is achieved, we have very little idea, though we are of course quite familiar with the corresponding process by which a new word is introduced into a natural language.

3. The predicate is given cognitive significance by an act of stipulation: it is defined to apply to just those entities that fit the observed pattern. In the swan case, the new predicate *swan* is stipulated to apply to just those birds having the observed properties of whiteness, red-beakedness, and so on.

As I have told the story, all three stages proceed rather self-consciously. This is for expository convenience only; it may well be that the latter two stages, in particular, occur sub rosa—we observe a pond full of these fascinating new red-beaked, trumpeting, aquatic birds, and the next thing we know we have the swan concept and can think about them as members of the same category, rather than merely as identically endowed individuals.

Must classical category construction proceed by way of these three stages? As far as I can see, this is the only feasible story for the classicist. (Concept learning, as opposed to construction—for example, the process by which you learn the meaning of the preexisting English word *swan*—may proceed by a different process in which various hypotheses concerning the correct definition are tested against known examples of swanhood.)[2]

1.2 The Prototype Approach Consider a simple version of the prototype theory of concepts according to which a concept is associated with a list of features, and an entity is classified as a member of the corresponding category

2. For more on the distinction between category learning and category construction, see Murphy (2002, chap. 5).

if it has sufficiently many of these features. "Sufficiently many" will be given a precise meaning by a system of weights and a threshold, or something similar; the details do not matter in what follows, however, and so will be glossed over here.[3]

The differences between the prototype view and the classical view are well known: the prototype view explains "typicality" effects; the prototype view allows for vague boundaries between categories; the classical view predicts that there will be properties that are considered essential for category membership (Rosch and Mervis 1975; Rosch 1978; Smith and Medin 1981).

Yet the standard picture of concept acquisition on the prototype view is rather similar to the picture of classical acquisition laid out above, having as it does the following three stages:

1. A certain statistical pattern is observed in the world: the observation of a number of entities tending to share certain properties, or sharing an overall family resemblance, not characteristic of any pre-existing category.

2. A new mental predicate is created.

3. The predicate is given cognitive significance by an act of stipulation: it is defined to apply to just those entities that have sufficiently many of the correlated properties.

What is important for my purposes is that the third step involves, just as it does in classical category construction, an act of stipulation. Stipulation by its very nature establishes grounds for category membership—it says that to be a category member *just is* to satisfy such and such a criterion—and so

3. What I say about the prototype approach should also give you some idea of what I would say about the exemplar approach, but as noted above, for reasons of space, I will have to leave the treatment of exemplar theory as an exercise to the reader. For the possibilities inherent in the prototype approach, and the relation between the prototype and exemplar approaches, see Smith and Medin (1981) and Murphy (2002).

the prototype approach to concepts, if it endorses the present story about concept acquisition, is committed to the depth thesis.

Does it have to be this way? Does the third step in prototypical concept acquisition, by which the new concept is given cognitive significance, have to be a *stipulative* act? In the case of the classical account the answer appeared to be *yes*, because classical cognitive significance comes by way of a definition, and to frame a definition is to make a stipulation.

But in the case of the prototype account, it is not so clear. In a well-known paper, Osherson and Smith (1981) propose a view on which concepts have two parts, a classical core and a prototype that functions as an identification procedure. The core stipulates what it is for an entity to belong to the corresponding psychological category—it represents the ultimate grounds for category membership. The prototype functions as a heuristic, a procedure that is good enough, but not perfect, at determining category membership. On this view, then, when a concept is acquired a prototype is acquired, but the prototype is not attached to the concept by a stipulative act in which the concept is *defined* in terms of the prototype. It is rather attached to the concept by an empirical posit, a supposition that entities falling under the concept will tend to fit the prototype (and vice versa). Thus, there can be prototypes that do not represent the grounds of category membership.

In this particular case, the prototype nevertheless piggybacks on such a stipulation: a new category is constructed by the positing of a classical definition, and the prototype comes along as a good enough test for that definition's holding. (Thus, Osherson and Smith's account subscribes to the depth thesis.) Could a prototype concept be acquired without the stipulation of a classical core, yet also without the prototype itself being taken as the grounds of category membership?

I will later argue that the answer is *yes*: you can relinquish the depth thesis yet retain something that is recognizable as a prototype account of concepts. But at this early stage, my goal is to characterize rather than to

condemn the depth thesis, and so I will satisfy myself by pointing out that in its paradigmatic statement, the prototype theory subscribes to the thesis and takes the prototype to lay down the basis for category membership. In taking stipulation to be canonical, I have in mind the claim of Rosch and Mervis (1975) and others that psychological categories have a prototype or "correlational" structure, in which the basis for category membership is a kind of "family resemblance". Rosch and Mervis write, for example, that when psychological categories are constructed using prototype concepts, "the categorical relationship can be understood in terms of the principle of family resemblance" (p. 603), and they say that the categories "mirror the correlational structure of the environment" (pp. 586, 602). These proposals strongly suggest that when a prototype concept is acquired, the prototype criterion is not regarded as constituting a mere heuristic, but as determining the structure of the category itself, by spelling out the ultimate grounds for category membership.[4] That is the essence of the depth thesis. I will, therefore, call this canonical version of the prototype theory the *deep* prototype theory; a shallow prototype theory will be discussed in section 4.3.

1.3 Psychological Essentialism The psychological essentialist view (Gelman and Wellman 1991; Gelman et al. 1994; Gelman 2003) will for the most part in what follows stand in for the theory theory of concepts. This is something of a stretch, because essentialism is not usually thought even by its proponents to be true of all categorical concepts, but only of a subclass, including concepts of biological taxa (tiger, oak) and chemical substances (water, gold). Not all lessons drawn from psychological essentialism, then, will apply to the theory theory as a whole, but I do not think that this will matter too much for the purposes of the discussion of the depth thesis.

Like the classical view, the essentialist view has the cognizer representing

4. Rosch and others have since weakened their claims about the significance of their research; it is quite unclear that it is appropriate to call the watered-down view a prototype account of concepts at all.

from the moment of concept acquisition a condition that defines what it is for an entity to belong to the category, namely, possession of an "essence" that is causally responsible for many of the characteristic observable properties of category members. The essentialist theory of concepts, then, by its very nature involves a belief about the grounds of category membership, and so a commitment to the depth thesis.

Let me break down the process of essentialist category construction into three stages, as I did for the classical and prototype theories. A cognizer observes a number of animals, all with similar properties, which do not fit into any known category. Perhaps they are white-feathered, red-beaked, trumpeting aquatic birds. The cognizer hypothesizes that these shared features have an unobservable common cause. They then introduce a new mental predicate—say, *swan*—and they stipulate that it applies to all and only those entities that possess the causal property in question. The act of stipulation creates a new category, then, that has the putative common cause as its essence. Note that the stipulators may have very little or no information about the nature of the common cause; their representation of the cause is then in some sense a mere "placeholder".

More generally:

1. (a) A certain statistical pattern is observed in the world: the observation of a number of entities tending to share certain properties, or sharing an overall family resemblance, not characteristic of any pre-existing category.

   (b) A common cause is posited for the correlated properties.

2. A new mental predicate is created.

3. The predicate is given cognitive significance by an act of stipulation: it is defined to refer to just those entities that have the property hypothesized to be the common cause.

The result, in the case of the acquisition of the swan concept, is a commitment to a new theory with the starburst structure shown in figure 1.

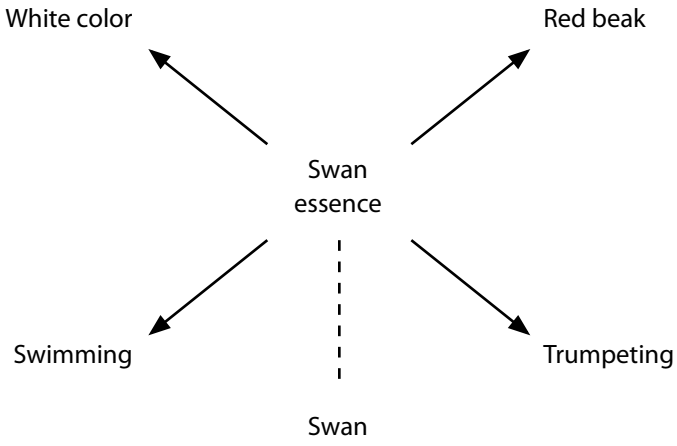

Figure 1: What the essentialist represents upon constructing the category of swans. Arrows represent causal relations; the dashed line represents a defining relation created by stipulation.

Psychological essentialism adheres to the depth thesis, because it assumes that the cognizer represents, for each of their psychological categories, the ultimate grounds of category membership. But it allows that very little goes into that representation; it may be no more than a placeholder, "a common cause, I know not what". The "deep" layer of an essentialist concept, then, is somewhat thin. In this respect, essentialism comes close to being a shallow theory of concepts, that is, a theory on which concepts have no deep layer at all.

There is a tendency in recent writing on psychological essentialism, however, to suspend the theory's flirtation with shallowness by enriching its deep layer, in two steps. First, it is suggested that the cognizer assumes more about essences than their causal role in producing characteristic properties strictly implies. Second, it is suggested that these assumptions are built into the

cognizer's conception of the very nature of essences, so that in stipulating that a mental predicate is to be defined in terms of an essential property, it is also stipulated that the defining property satisfies the assumptions articulated in the first step.[5]

Ahn et al. (2001), for example, hold that cognizers attribute to biological essences an inherent tendency to get passed from parent to child, and further, that merely to believe that biological kinds have essences is to attribute to them this "inheritability". (They conclude that essentialism therefore has an explanatory advantage over Strevens's (2000) causal minimalism (discussed in section 3) in explaining cognizers' evident commitment to inheritability, on the grounds that essentialism explains inheritability "for free" whereas the causal minimalist must make an additional empirical posit to the effect that cognizers believe that kind membership is inheritable.)

An essentialism with even "deeper" commitments can be obtained if essences are represented as having their causal powers essentially, so that (for example) the swan essence causes its bearers to have red beaks by definition. This is how Rips (2001), for example, understands essentialism.[6,7]

---

5. Note that it is only in the second step that the assumptions about essences are included in the category-creating stipulation and so are made to be a part of the grounds for category membership.

6. The distinction I make here may be seen as arising from the distinction between two possible stipulative acts that might be taken as the final stage of essentialist concept acquisition. One of these acts is that specified above: let *E* be the unobserved cause of the actual characteristic properties of swans; define swans as whatever entities have the property *E*. The other act widens the scope of the definition: define swans as whatever entities have a property that is (a) identical to *E* and (b) causes the actual characteristic properties of swans. The difference between the two definitions shows itself when counterfactual questions about swans are asked. Could it be that swans had yellow beaks? On the first definition, the answer looks to be *yes*: it might have been that *E* causes yellow rather than red beaks, if the biological laws had only been different. On the second definition, it is impossible: it is not enough to be a swan that you have *E*; in addition it must be the case that *E* causes red beaks. If *E* causes yellowness, then, there can be no swans, whether or not there are birds that have *E*. This latter view, and Rips' case against it, are further discussed below.

7. Note also that I am running two distinct views together here. On the first, the essentialist cognizer stipulates that the power to cause the characteristic properties, *whatever they are*, is constitutive of category membership. On the second view, the cognizer goes further and

In the same paper, Rips introduces a different kind of "theory theory" approach to biological kind concepts, according to which the grounds of category membership are possession of sufficiently many causal properties: "An object's membership in a natural kind depends on whether the object instantiates the [causal] laws for that kind" (p. 846). As Rips remarks, what results is a kind of prototype view on which the membership-determining features of a category are not surface appearances but the causes of those appearances (p. 848). Thus Rips subscribes to the depth thesis in much the same way as Rosch and Mervis (1975).

More generally, proponents of the theory theory tend to adhere to the depth thesis in the following way. First, as theory-theorists, they hold that theoretical beliefs about a category—its internal and external causal connections in particular—play an important role in people's reasoning concerning category membership and so on. Second, they further suppose that cognizers represent as the grounds of category membership an entity's conforming to the theory in certain ways—for the essentialist, in possessing a certain unobserved but causally potent property; for Rips, in conforming to sufficiently many of the theory's causal laws; and so on. This paper will propose that there is much to be gained in resisting the second step.

You might wonder: does the second step make any difference at all? Are there any empirical ramifications to the depth thesis? Let me answer that question next.

---

specifies the properties in question. Thus on the first view, the cognizer says that whatever the actual characteristic beak color of swans, it is essential for category membership that the essential property cause that very color, while on the second view, the color—say, red—is named up front, so that what is specified as being necessary for category membership is that the essential property causes *red* beaks. When I discuss the "deep essentialist" view below, I continue for simplicity's sake to conflate these two views, by assuming that the cognizer knows the identity of the characteristic properties for sure.

2. The Significance of Depth

So the mind represents resemblance to a prototype, or possession of a causally important property, not merely as another interesting characteristic of a category, but as the grounds of category membership—so what? Does such a distinction have any psychological implications?

It does; the form of the implications depends, however, on the epistemic status of the cognizer's representation of the grounds of category membership, and in particular, on whether the cognizer considers their belief about the grounds as a mere hypothesis or as the result of a category-creating stipulation. Because I take the principal motivation for the depth thesis to be a canonical story about acquisition that involves such a stipulation, I will simply assume in this section, and in the remainder of the paper, that the latter condition applies, that is, that the criterion that is considered to state the grounds for category membership is so regarded because it articulates the content of a stipulation made at the time of category construction. In effect, I am focusing on one particular variant of the depth thesis, albeit a variant that I take to be dominant in the psychological world.[8]

I will uncover the empirical significance of the depth thesis in two steps, first showing that category-creating stipulations result in the stipulator representing what I will call a *final criterion* for category membership, and second, discussing the psychological consequences of representing a final criterion.

A final criterion for a psychological category is a condition that is, and is regarded by the cognizer as, the final, indefeasible arbiter of category membership. It spells out the ultimate grounds of category membership, and thus determines

1. The boundaries of the category,

2. The property or property complex shared by all members of a category

8. The shallow approach to concepts defended later in this paper rejects the depth thesis in all its possible forms, not only the form discussed here.

in virtue of their category membership, and so

3. What many psychologists call the *structure* of the category.

Further, the cognizer knows all of this; finality subsists not only in the relation between criterion and category, but in the cognizer's attitude towards this relation, in their considering the criterion to be the last, incorrigible word.

It is quite clear and uncontroversial, I hope, that the sort of category-creating stipulation that constitutes the third and final step of the concept acquisition processes surveyed above results in the acquirer representing a final criterion. Indeed, this is hardly a side-effect: in the classical theory, the deep prototype theory, and the essentialist theory a new psychological category is created precisely by specifying a final criterion for that category. The stipulation is a stipulation of the grounds of category membership, thus the self-conscious provision of a final criterion.

The question, then, is whether there is a psychologically meaningful difference between representing a final criterion for a category and representing a mere matter of fact about a category. Consider, for example, the red-beakedness of swans. Is there a difference between the belief that red-beakedness is definitive of swans, and the belief that, as a matter of fact, swans are red-beaked?

To see that there is, consider first the classical account of concepts. On the classical account, the cognizer considers red-beakedness to be an indefeasible, necessary condition for swanhood. As a consequence, they will refuse to allow two possibilities: first, that some unusual swans might be yellow-beaked rather than red-beaked, and second, that they might have been wrong about the red-beakedness of swans all along, for example, that most swans might be yellow-beaked rather than red-beaked. This refusal is not obstinacy; rather, it is completely reasonable, indeed compulsory, given the way they have constructed their swan category. The classical cognizer is not failing to recognize the possibility that there might be birds that are like swans in every respect except for the color of their beaks, or that the birds that motivated

them to construct their swan category might be mostly yellow-beaked rather than red-beaked. They are simply recognizing the fact that, because their swan category is *defined* to contain only red-beaked creatures, the yellow-beaked specimens do not belong to the category. They may well construct a new category that contains both the red-beaked and the yellow-beaked birds; the old category, however, because of its structure cannot contain any animal with a yellow beak. By contrast, a mere belief that swans are yellow-beaked is defeasible every which way.

Other final criteria will have the same kind of psychological implications. Unlike ordinary beliefs, they will impose certain absolute limits, determinable in advance, on category membership. These limits cannot be lifted; they must be bypassed by the construction of a new category with fewer or different limits. The old category, by its very nature or structure, will not admit of change. Let me discuss some more examples of final criteria from the prototype and theory theories of concepts, then move on to consider briefly the advantages and disadvantages of a final criterion.

A concept with prototype structure does not, of course, stipulate individual necessary conditions for category membership. A bird that is a swan in every respect save its yellow beak is sufficiently similar to the swan prototype to count as a swan; even if red-beakedness is one of the swan prototype's features, then, category membership is not limited to red-beaked birds. But what I have called deep prototype theory still puts some absolute, inviolable limits on category membership, because the construction of a deep prototype category involves a stipulation about family resemblance and so creates a final criterion: it is definitive of category membership that a specimen resemble the prototype to a certain degree. Something utterly unlike the prototype in all important respects cannot be a swan.

It is for this reason that the prototype theory (along with the classical theory) finds it difficult to accommodate the results of Keil's (1989) transformation experiments, in which older children and adults said that a raccoon

that had been given skunk appearances and behaviors by various superficial transformations was still a raccoon, not a skunk: because raccoons are stipulated, on the deep prototype theory, to have an overall raccoon aspect (though they may lack individual raccoon characteristics), something with an utterly different aspect cannot be a raccoon. (This is not a refutation of the deep prototype approach, however; Keil's results are consistent with the view that the category-determining prototypical features are unobservable causal features rather than surface appearances and behaviors.)

Another example of a belief's status as a final criterion having empirical consequences can be found in Rips's (2001) ingenious argument against the sort of rich psychological essentialism, outlined above, on which not only the deep essential property but its specific causal powers are stipulated as constitutive of category membership. On this rich essentialist view it may be stipulated as part of what it is to be a member of the psychological category of swans, for example, that a bird's deep properties cause it to grow a red beak. It follows that it is possible for individual swans not to have red beaks, since something may go wrong with the causal mechanism that gives beaks their color without changing the fact that it is a red–beak-producing mechanism. However, because of the final criterion established by the stipulation, it is not possible that swans as a whole have yellow beaks; there might be birds just like swans whose beaks are naturally yellow rather than red, but they do not fall into the swan category with its stipulated red–beak-causing mechanism. Rips argues that, although people believe that swans have red beaks (or the equivalent), they find the possibility that swans have yellow beaks and the like quite coherent. For example, they are well able to reason intelligently about such questions as "If Pekinese could leap over cars, could Dobermans?", despite the fact that on the rich essentialist view, Pekinese are stipulated to have causal properties limiting their jumping abilities (or so Rips supposes).[9]

9. Rips seems not to notice that his argument does not militate against lighter-weight versions of essentialism, on which members of a category have their deep properties essentially,

Rips's argument turns, then, on his observation that representing the red–beak-causing power of swans' essential properties as a final criterion, rather than as a mere belief, has readily observable consequences.

What about the lightweight essentialism on which the one and only property required for category membership is an unobserved common cause of a complex of correlated observable properties?[10] Does the final criterion associated with this view have testable consequences? That is a subtle question—the lightweight essentialist constructs their categories using a stipulation that enshrines almost no empirical commitments as apodictic truths about category membership. One commitment, however, is clear: that the characteristic observable properties have a common cause. Consequently, essentialists should not be willing to provide substantive answers to Rips-style questions about counterfactual scenarios in which (say) there is no essential property unifying a biological species—since in these scenarios, nothing falls into the category in question.[11]

However, let me not pursue this line of thought here. It is enough, for now, that I have shown that a commitment to a final criterion has significant psychological consequences—as you ought to expect, given that the criterion determines the structure of the corresponding psychological category. It therefore makes sense to ask of any domain of concepts: what form for a final criterion fits best with the empirical data? But it also makes sense to ask whether there is evidence for any final criterion at all. I will suggest that there is not: there is much evidence against final criteria with various forms, and no evidence in favor of the existence of a final criterion of any other form. Adherence to the idea of a final criterion—to the idea that categories are constructed by stipulating grounds for category membership—is a result

but those deep properties do not have their causal powers essentially (see note 6).

10. In Strevens (2000) I called this "pure essentialism".

11. It is not clear, however, that Rips' experiments are decisive: people may be willing, on request, to reason about conceptual impossibilities—they might accept, for example, that "If 6 were prime, it would have no divisors other than 1 and itself".

not of empirical pressure but of a commitment to the depth thesis.

What is wrong with depth? Is it not obvious that psychological category construction involves the stipulation of grounds for category membership and, more generally, that representing a category involves representing the grounds for membership?

The deep view of concepts has two great advantages, I think. First, it has a clear and coherent story to tell about concept acquisition, centered on the stipulative act. Second, it offers a simple explanation of the groundedness and stability of psychological categories (see section 4): it explains what anchors a category, that is, what prevents it from "drifting off" over time, changing its structure and cognitive significance (see section 4).

And the shallow view? Perhaps the best reason to explore the shallow view is curiosity. "Big" theories of concepts have since the inception of the field labored under the assumption that concepts are built around a defining supposition about the grounds for category membership. Why not relax that (rather classical-sounding) assumption and see what happens? What happens if the supposition is not taken to be certain, but defeasible? What happens if the supposition is removed altogether, resulting in a concept's user's no longer representing any grounds for category membership, but only heuristics? Does the mind fall apart? Or do new possibilities swim into view?

3. Shallow Categories: Categorization

Could the depth thesis be wrong? Can there be a viable shallow theory of concepts, that is, a theory on which a concept's user does not represent criteria for category membership?

In the introduction to this paper, I divided the obstacles to a shallow theory into two classes:

1. The question of use: how can a shallow concept play the various cognitive roles that it must, in categorization, induction, and so on?

2. The question of acquisition: how can shallow concepts be acquired in the first place?

I will first tackle the question of use. In a single paper, it is not possible, but also not really necessary, to discuss all possible cognitive roles a concept may play. In what follows I will treat only that psychologically most canonical of roles, categorization (section 3). I will then discuss what I take to be the potentially most telling objection to the shallow theory as a theory of use, the complaint that shallow concepts are ungrounded or pathologically unstable (section 4). Acquisition is investigated in section 5, followed by an overview of the shallow approach (section 6).

* * *

On a shallow account of concepts, I wrote above, categorization is a matter of "heuristics all the way down". How does that work? 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 such a conclusion 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 ultimate grounds for membership of the class of California-nutmegs; the example shows, then, that

categorization can proceed without the use of a final criterion (or any other deep beliefs, that is, any other beliefs about the ultimate grounds of category membership).

You might quite reasonably point out 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 spiky-needled trees in the grove are 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 knows for sure what makes a tree a California-nutmeg.

But this last train of reasoning is fallacious. My trust in the guidebook 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, say, a configuration of DNA, that causes California-nutmegs' characteristic observable properties. 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. As Keil's transformation experiments show, this information is not regarded as giving the ultimate grounds for category membership, and so 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 have it to qualify as experts.

The experts of bygone times, then, were in a certain important sense like 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 a certain corner of my 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.

Let me make these observations more concrete by putting them in the context of a positive theory of concepts of biological and chemical taxa, namely, the "causal minimalist" theory proposed by Strevens (2000), which is—very roughly—psychological essentialism without its essences.

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

1. All swans are white,

2. All swans trumpet,

3. All swans have red beaks.

Because these are causal hypotheses, they should be interpreted as saying that there is something about swans that *causes* them to be white, something that *causes* them to trumpet, and so on.

Note that the essentialist also represents causal hypotheses of this sort; the essentialist hypotheses, however, explicitly mention and attribute a causal role to the swan essence. (Compare the essentialist theory of swans in figure 1 to the causal minimalist theory in figure 2.) Otherwise, essentialism and causal minimalism treat the hypotheses in the same way in two important respects. First, on both views, the characteristic observable properties mentioned in the hypotheses do not make up any kind of final criterion. Thus the hypotheses may have exceptions, so that, for example, some swans are not red-beaked.

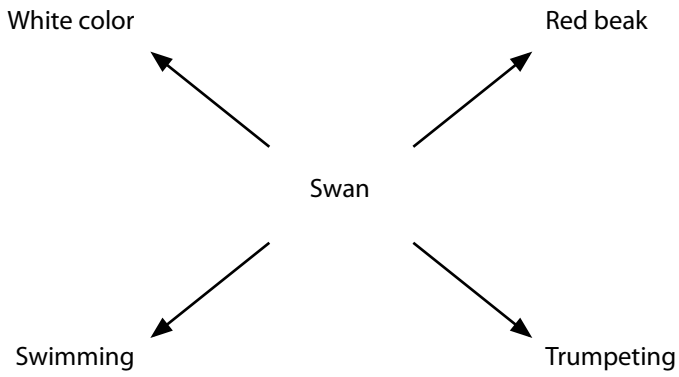

Figure 2: A causal minimalist theory of swans. The arrows represent causal relations. (More exactly, each arrow represents the putative fact that there is something about swans that causes the relevant property.)

Second, the causal hypotheses themselves do not constitute a final criterion: it is in no sense considered a necessary truth that swans conform to the hypotheses stated above (or at least, this is true on the "lighter-weight" varieties of essentialism). It is allowed that the hypotheses may be false, or even that they may be true now but may become false later (as a result of

evolution, say). The point is that there is nothing about these hypotheses that binds them more tightly to the swan concept than any other belief about swans is bound to the swan concept. If we are more reluctant to abandon our causal hypotheses about swans than to abandon, say, the belief that there are swans in Lake Rotorua, it is because we have more evidence for the hypotheses than for the presence of swans in the lake. In short, the hypotheses are considered to articulate just a certain class of matters of fact about swans, albeit an inductively very rich kind.

How does causal minimalism explain categorization? Suppose that we see a bird swimming in Lake Rotorua. It is white-feathered, red-beaked, and it is given to trumpeting. What might it be? We know that an animal's being a swan 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 swan, and so that the animal is a swan. 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 causal minimalism differs from essentialism and the other theories of concepts mentioned above 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 have more and more accurate causal hypotheses.[12]

Taking a step back, and thinking about the mind from an engineering perspective, is there any reason to add a final criterion to a causal-minimalist concept? Is there any reason to take some particular part of a cognizer's theory of swans—any particular belief (or set of beliefs) about swans—and make it a final criterion, by stipulating that satisfaction of the belief constitutes the ultimate grounds for category membership?

As far as categorization is concerned, the answer is surely *no*. Such a stipulation will distort your pre-existing theory of the category in question, by exaggerating the importance of one part of your theory over the others, while at the same time limiting the uses to which you can put your category, as the (deep) prototype theory limits the application of the raccoon concept in Keil's transformation studies, or the (rich) essentialist theory limits the application of the Pekinese concept in Rips' counterfactual studies. Why unbalance and limit a perfectly good mental theory of swans in this way?

There are two interesting reasons, to be considered in the next two sections—and rejected.

4. Shallow Categories: Groundedness

That categorization and other inferences about a psychological category should go on without necessarily making use of a final criterion should not seem so peculiar: after all, we use heuristics in reasoning all the time. What

12. This claim ought to be qualified: there are a few scientists or philosophers who have self-consciously constructed biological or chemical categories for themselves using explicit final criteria (cf. Mayr's (1970) "biological species concept"). My argument is not that such acts of classical category construction are impossible or unknown, but rather that they are rare, recent, and have little influence on everyday cognition. I might also point out that people typically do not act in accordance with their professed beliefs about the grounds for category membership (Malt 1994; Strevens 2000).

might seem peculiar is that inference should go on without a final criterion's being represented at all. Inference may not be actively guided at all times by a final criterion, the thought goes, but surely such a criterion must be held in reserve, to provide validity to whatever heuristics are in play? For if a psychological category's constructor does not know the grounds of category membership, how do they know good from bad heuristics?

This *groundedness* worry—that without representation of a final criterion, psychological categories will be ungrounded—comes, I think, in two varieties which might be called the *external* and the *internal*.

4.1 External Groundedness The external groundedness worry concerns reliability and truth. It takes as its master premise the thought that, if a psychological category's constructor fails to represent the grounds for category membership, then there can be no fact of the matter about those grounds. From this premise it appears to follow that there can be no fact of the matter as to whether the cognizer's beliefs about the category are true or whether their inferences about the category, including their categorizations, are reliable.

Is the master premise correct? You might think not, on the grounds that, although an individual cognizer may not represent the final criterion, there are experts who do.[13] But of course it is a part of the wholesale rejection of the depth thesis advocated in this paper that even the experts do not represent the ultimate grounds for category membership. They are experts about a category not because they have a final criterion for that category, but because they know many important facts about the category, for example, the causal laws connecting membership of biological categories to characteristic appearances. Very well then; what makes the experts' beliefs true? No one, it seems, has a final criterion on offer to provide the necessary grounding.

13. I am using the "final criterion" somewhat loosely here—strictly speaking, a criterion is final for a cognizer only if they both represent it explicitly and recognize it consciously as providing the grounds for category membership.

The external problem is solved by, and only solved by, there being some fact of the matter about the extension of a psychological category. What is needed, in other words, is a semantic or philosophical account of which things in the world a psychological category picks out. A natural and appealing account of this sort goes as follows:

> A psychological category has as its members just those things in the world that satisfy the associated final criterion.

This view is not available to rejecters of the depth thesis. But there are other options in the philosophical literature.

Fodor (1998), for example, suggests that a psychological category picks out just those things that the category's constructor would pick out as members under certain special circumstances, or in a certain special way. This privileged extension-determining categorization behavior need not (and according to Fodor, does not) make use of a final criterion. It will include many of the kinds of categorizations described in section 3; Fodor's theory of reference is therefore quite compatible with a shallow approach to categorization.

Another theory of reference, advocated by Boyd (1988), equates a category's extension with whatever the category's constructor would ultimately pick out as members, given enough time and information. Yet another theory, perhaps the most popular among philosophers for biological and chemical categories, is the causal-historical theory on which a category's members depend on the circumstances in which the name for the category was originally introduced (no stipulation of a final criterion is presupposed) along with certain facts about natural categories, the idea being that the psychological categories tend to hook on to natural categories. On this view, the final criterion for category membership is, in effect, out in the world—it is whatever serves as the grounds for membership of the corresponding natural category—not in the head of the category's constructor (Putnam 1975). To put it another way, the constructor achieves the construction by "borrow-

ing" a final criterion from nature, without themselves ever appreciating the content of the criterion.

It is up to philosophers to adjudicate among these options, and I will say no more about them here; the point is simply that it is possible to have facts about reference, and thus facts about truth and reliability, without the internal representation of a final criterion.

4.2 Internal Groundedness It is quite natural for psychologists to react to philosophical theories of reference with indifference—how could the facts about natural categories, or about the future history of a concept, possibly affect the way that the concept is deployed in reasoning *now*?

For such psychologists, the external problem of groundedness was never quite the right worry—the right worry is not whether there is a fact of the matter as to whether our categorization heuristics and other inferences are reliable, but whether we have some internal procedure for validating their reliability, in the absence of an internally represented final criterion. How can we check the effectiveness of a heuristic, without having some standard against which to assess it?

There is a simple answer to this question. The "heuristics" for categorization and so on that I have discussed are the result of the application of inductive logic to beliefs about the category in question. For example, the procedure used to classify something as a swan, based on its physical appearance and behavior, emerges from the logic of causal inference, as applied to our causal beliefs about swans. The validity of inductive logic itself is not in question, I take it (that is a topic for another time!), so the question of the validity of the heuristic is the question of the validity of our causal beliefs about swans.

Phrased in internal terms, then, the question we should ask ourselves if we want to assess our categorization techniques is: do I have good reason to believe that my swan beliefs are true? If there is a way to answer such

questions without invoking a final criterion, then there is a way to validate my heuristics without invoking a final criterion.

But of course, there is such a way: in two words, *inductive inference*—or more formally, the scientific method. I find some swans; I observe their properties and take note of the background conditions; I compare these facts to the predictions of my beliefs. Is a final criterion necessary? One especially secure way of taking the first step—finding some swans—is to use a final criterion. Then I can be *certain* that the animals at hand are indeed swans. But it is not *essential* to proceed in this way. All I need are good reasons for thinking that the animals at hand are swans, and a final criterion is not needed to have such reasons, as any scientist knows. (That is how science has been able to investigate electrons, or genes, or beliefs, without having a final criteria for any of these categories.)

To be sure, there is a subtle kind of bootstrapping at work: I may use the same beliefs that are under investigation to make my preliminary judgments that my test animals are swans. But such bootstrapping, done right, is a reasonable way of acquiring knowledge. How did we discover that some swans are black, for example? We applied our "swan theory", which among other things ascribed whiteness to swans, to certain Australasian birds. Our theory told us that they were, despite their not being white, likely swans. We concluded that one part of our theory, the causal hypothesis connecting swanhood to whiteness, was false. Similar cases are to be found everywhere in science—and in everyday life.

4.3 Conceptual Stability Related to worries about groundedness are worries about stability. A final criterion provides, for reasons laid out in section 2, an anchor—a fixed point from which the concept cannot stray. You might think that a concept is better off with such an anchor to keep it pointing at some particular part of the world.

Depending on the kind of "pointing" involved, this thought might be

developed as an internal or as an external worry about the shallow approach to concepts. Let me consider an internal version of the question, as follows. Without an anchor, is there any reason to think that there will be some kind of convergence in what is counted as a category member, as the cognizer learns more about the world? Or will the class of entities falling under an unanchored category drift like an unanchored ship, narrowing, widening, moving, never settling down? (Some concepts arguably do drift rather than converging, but many do not; it is the latter that are the focus of the stability concern.)

A final criterion will certainly serve the purpose of nailing down a category. But a sufficiently rich body of defeasible, non-definitive beliefs can also create the kind of tendency to convergence that will assuage worries about stability. The causal minimalist account of concepts of biological and chemical categories provides one example. The hypothesized causal connections linking category membership and observable properties are normally sufficient, even if not all of the connections are correct—as in the case of the belief that swans are white, discussed immediately above—to fix the concept to a particular group of animals. As further information about the animals comes in, the theory is updated and categorizations converge onto something like the actual species in question.

As another example, consider what might be called a shallow prototype concept, that is, a prototype-based concept without a final criterion. Let me take a few sentences to explain this notion. One way to understand a prototype-based concept is as a group of statistical beliefs linking category membership and possession of observable properties. A prototype-based swan concept, for example, might have at its core the following probabilistic beliefs: *Most swans have red beaks*, *Most swans trumpet*, and so on.[14] What

14. It would be truer to the prototype approach to give more precise values for probabilities, and to do so using a joint probability distribution, since the probabilities are not independent. Such sophistications will not matter here.

I call the "deep" prototype theory adds that these statistical generalizations are stipulated to constitute the ultimate grounds of category membership (thereby determining category structure). But suppose that this assumption is abandoned: there is no stipulation, thus no tendency on the cognizer's part to regard the generalizations as providing an ultimate grounding for the category of swans. In effect, then, the cognizer regards themselves as having some statistical knowledge about swans but no knowledge about the ultimate basis of swanhood. The resulting concept of swan is what I mean by a shallow prototype concept—it is based around prototypical knowledge, but that knowledge is not represented as constituting a final criterion.

A shallow prototype concept typically has, I suggest, more than enough built-in stability for everyday cognition, at least in those non-degenerate cases where the concept corresponds, even imperfectly, to a real correlation in the world. After all, the user of the concept will in the short term make precisely the same categorizations as the user of the corresponding deep prototype concept, since they apply precisely the same probabilistic criterion for category membership. The difference is simply that the deep user, but not the shallow user, regards the criterion as apodictic.

One important advantage that the shallow concept has over the deep concept is that, in the longer term, the shallow concept has a natural tendency to converge on real-world correlations, if it does not already represent those correlations entirely accurately. Consider again the case of the black swan. Suppose that a cognizer begins with a prototype according to which swans are (usually) white. They find a continent of black birds that are otherwise entirely like swans. Despite their color, these creatures resemble the prototype sufficiently well that they will be classified as swans. So far, the course of events is the same for the shallow prototype as for the deep prototype.

At the next stage, however, the stories diverge. The user of the shallow swan prototype regards *Most swans are white* as a defeasible empirical hypothesis about swans. They will regard the discovery of large numbers of

black swans (believed to be swans because of their overall resemblance to the prototype) as evidence against the hypothesis. Thus they will presumably abandon their belief in favor of the more liberal belief that *Most swans are either white or black*. This is a normal change in belief in the face of new evidence, yet it constitutes in effect an amendment of the swan prototype so that white color is no longer prototypical, with the new prototype better reflecting the correlational structure of the world.

The deep swan concept, by contrast, cannot undergo such a change. Its user has *stipulated* that whiteness is a part of their swan concept. In the face of substantial numbers of black swans, they must go on regarding these swans as atypical category members; the best they can do is to construct a new concept that applies equally well to black and white birds. (Note that this new concept does not in any meaningful sense *replace* the old swan concept; each concept picks out a psychological category in its own way, and is valid on its own terms. The most that can be said is that the new concept is more useful than the old concept, and may as a consequence be more frequently applied.) The deep swan cognizer reacts to new statistical evidence, then, by constructing new concepts. The shallow swan cognizer in effect revises their old concept to better correspond to the way things are in the world.

What lessons to draw? First, a shallow prototype concept has considerable stability. Second, insofar as it is subject to change, the change is in a desirable direction—the corresponding "instability" is much to be desired. Third, to deal with new empirical information the shallow cognizer needs only to have the usual equipment for bringing new evidence to bear on existing beliefs—the usual inductive logic. The deep cognizer needs a policy and procedure for constructing new concepts of groups of things that are already picked out by an existing concept. None of this fatally undermines the deep approach to concepts. But it does illustrate how natural and attractive a shallow approach can be.

5. Shallow Categories: Acquisition

All three theories of concepts surveyed in section 1 posited a three-step procedure for constructing new categories (that is, acquiring new concepts through unsupervised learning). Schematically, the steps are as follows:

1. Learn new facts about the world (typically, that certain observable properties come in clusters),

2. Create a new mental predicate,

3. Give the new predicate cognitive significance by stipulating that it stands for a category with such and such grounds for category membership.

The grounds stipulated in the third step may be: possession of each of the correlated properties (classical account), possession of sufficiently many of the correlated properties (prototype account), possession of an unobservable property hypothesized to be the common cause of the observable properties (essentialism), possession of an unobservable property stipulated to be the common cause of the observable properties ("rich" essentialism), possession of sufficiently many of the causes of the observed properties (Rips' "interactionist" account), and so on. (Please forgive the roughness of some of these formulations.)

Whatever the details of step three, a final criterion is created: because the new category is created by stipulating grounds for category membership, the creator—the acquirer of the concept—has certain knowledge of those grounds. The above story of category construction is therefore committed to the depth thesis.

Among the consequences of renouncing the depth thesis is the need for a new story about category construction. Can there be such a story? Is there any way to construct a new psychological category without a final–criterion-creating stipulation? You might think not—how, after all, can a new category

come into being without there being *some* fact of the matter about what makes an entity a member of that category, and how could there come to be such a fact of the matter if it is not specified by the category's creator?

To answer this question, what is needed is a new and different candidate for the third step in the process of category construction, a candidate that does not involve a stipulative act, yet that gives the category the grounding and stability—the connections to other categories in the mind, and the connection to the world outside—that it requires (cf. Margolis 1998).

I argued in the previous section (section 4) that grounding and stability do not require a final criterion; a sufficiently rich and coherent set of non-stipulative beliefs about a category, such as *Most swans have red beaks* or *Something about swans causes red beaks*, give it both grounding (internal and external) and stability. To construct a new category, then, it is enough to acquire the kinds of beliefs about that category posited by the shallow prototype theory or the causal minimalist theory. To put it another way, to give a new mental predicate cognitive significance it is enough to embed it in a causal theory of the sort shown in figure 2 (the minimalist theory), or in its statistical counterpart (that is, a theory with the same structure but with the arrows representing correlation—or more exactly, high conditional probability—rather than causation), or in some other theory of equally rich empirical import.

How, then, to acquire the empirical beliefs in question? How, if acquiring the swan concept is a matter of acquiring a set of beliefs of the form *Most swans have P* or *Something about swans causes P*, for several characteristic properties *P*, to come to have beliefs of this sort?

The standard route to the acquisition of empirical beliefs is hypothesis formation and testing: you formulate the hypothesis; derive some predictions from the hypothesis; then check the predictions against the world. For example, to acquire the belief *Most swans have red beaks* you formulate the hypothesis and then find some swans and check to see whether they have red

beaks. If most of them do—if the evidence supports the hypothesis—then you come to believe the hypothesis.

But category-creating empirical beliefs cannot be learned in this way. In order to form the hypothesis *Most swans have red beaks*, you must have the concepts that occur in the hypothesis, one of which is the swan concept. Thus, the acquisition of the swan concept must precede the testing of any hypotheses about swans, from which it follows that the concept cannot be acquired by learning such a hypothesis in the usual way. (Compare and contrast Fodor 1981.) It might appear, then, that when it comes to category construction, there is after all no alternative to a stipulative act.

But perhaps this is to give up too quickly. Perhaps there is a process by which belief in a hypothesis can be attained without first testing that hypothesis (or stipulating its truth). What would such a process look like?

Suppose that you observe a kind of animal unlike any other you have seen before—a white-feathered, red-beaked, trumpeting aquatic bird. You say to yourself: it looks like this is not a species of animal that I have seen before; it is, then, a new species. Let me call that species the *swans*. Now, what do I know about swans? That they are mostly white-feathered, red-beaked, and so on.

This chain of reasoning results, I suggest, in the formation of the empirical beliefs in question, yet it arguably involves no act of stipulation. How does it work? The first step involves acquisition of information about the world: the observation of a particular kind of animal with particular properties, and the inference that the animals belong to no known species. The second step involves the creation of a new predicate (linguistic and also, presumably, mental). The third step involves the creation of new beliefs in which the predicate is embedded. It is in effect a formal operation, in which I take my new predicate and create a representation in which that predicate is connected to predicates standing for the various relevant properties: having white feathers, having a red beak, trumpeting. I do not in any sense test this

new representation; I simply adopt the attitude of belief toward it. I will call the mental process that comprises the third step *enduction*.

Is enduction a kind of learning? It depends on what you mean by learning. If learning is acquiring a new belief, then enduction is learning. But if learning is acquiring a new belief based on new information, then enduction is not learning. I do not receive any new information after step one; the validity of my enduction depends entirely on my old information. Yet enduction is not reasoning, either: it is not the end point of some train of thought about swanhood, since I acquire the ability to think about swanhood only once the enduction has finished, since only once it has finished does my predicate *swan* come to have cognitive significance—it is only once enduction is complete that I acquire the swan concept. Enduction is, I suggest, a kind of mental process that is not unfamiliar, yet which does not have a place in our folk psychology.

You might wonder whether it is correct that no stipulation occurs in the above story about the acquisition of the swan concept. Do I not stipulate that the new predicate *swan* picks out a category? I see no reason to resist this point. So yes, I will accept for the sake of the argument that the new predicate is stipulated to pick out a category.[15] Such a stipulation does not create a final criterion, however, which is all that concerns me here.

A further objection: might it be that I not only stipulate that *swan* picks out a category, but also that it picks out a certain sort of category, a natural kind or (even more specifically) a biological species? My answer here is that there is no need for such a stipulation. (Recall that my topic is the possibility of a shallow theory of acquisition; the question whether the empirical psychological facts support the theory is another matter.)

15. It is worth noting, however, that some philosophers of biology have argued that species are not categories but individuals (Hull 1978), the conceptual coherence of which position suggests that I may not assume anything other than that *swan* has as its extension a set of entities.

But without such a stipulation, you might continue, how will the psychological category of swans get itself embedded in my network of biological beliefs? How will I come to draw inferences about swans from my knowledge about birds, or vice-versa? The answer: such an embedding requires beliefs linking swanhood to other biological categories, such as the belief that swans are a kind of bird. It does not follow that I must stipulate that the beliefs are true, or that I must regard the satisfaction of the beliefs as a part of the ultimate grounds of swanhood. It is enough that they are beliefs—that I regard them as, to the best of my knowledge, true. Such beliefs can be acquired by enduction (or by a chain of inductive reasoning taking place after enduction).

I have posited an addition to the inventory of mental processes: enduction. If not entirely unfamiliar, it is at the very least under-discussed. I can hardly give it a full-dress treatment here, but let me make some comments about the range of enductive processes that might be found in the mind.

Two varieties of enductive process might be posited: domain-specific, automatic enduction, and domain-general, conscious enduction. I tentatively propose that the first kind of enduction occurs in the biological domain when the mind encounters an organism or group of organisms that do not fit into any known taxon at a certain level of classification (e.g., the folk genus level). A new predicate is automatically manufactured and connected via enduction to the observed properties of the novel organisms, I suggest—just as in the case of swans above.

The nature of the connection—the content of the beliefs formed by enduction—will be dictated by the correct theory of folk genus concepts; according to the causal minimalist approach, for example, they will be causal beliefs. So on encountering organisms with properties *P* that cannot be classified as members of an existing folk genus, the following process takes place automatically (without the need for conscious control):

1. A new predicate *K* is manufactured,

2. New representations of the form *Something about Ks causes P*, for each

of the properties *P*, are created,

3. The cognizer adopts an attitude of belief towards the represented propositions, and perhaps

4. The cognizer adopts an attitude of belief towards other propositions relating *K* to known biological categories, objects, and properties.

This is one example of what I am calling enduction.

A domain-general and perhaps consciously controlled form of enduction takes place, I suggest, whenever we acquire evidence that two properties have a previously unknown cause. For example, if instantiation of two properties *P* and *Q* is correlated, and the right sort of spatiotemporal relations hold between them, then under certain conditions the following process might be initiated:

1. A new predicate *C* is manufactured,

2. New representations of the form *C causes P* and *C causes Q* are created,

3. The cognizer adopts an attitude of belief towards the represented propositions.

In this way, the cognizer comes to have a new concept picking out a new property that is hypothesized to be the common cause of *P* and *Q*. Note that no stipulation is involved: *C* is not defined to be the common cause of *P* and *Q*. Rather, the representation of *C* has the cognitive significance that it does (and presumably the extension that it does) solely in virtue of empirical, defeasible, non-definitive beliefs about its causal connection to *P* and *Q*. That is what makes the process an enduction.

As this latter example shows, enduction is rather like the procedure by which new theoretical terms are introduced in science. Indeed, philosophers of science have long known that many such terms are introduced simply by

embedding them in new hypotheses. Newton is regarded as having introduced the notion of Newtonian force by writing down his three laws and the gravitational force equation; Maxwell is regarded as having introduced the notion of the electromagnetic field by writing down his famous equations; and so on. Typically, such introductions have been regarded as implicit definitions (Lewis 1970): to write $F = ma$ is to stipulate that the force F is what you get when you multiply mass by acceleration, or something like that. They might better be seen as enductions.

The case of enduction in science (if that is what it is) shows, I think, that enduction need not involve the cognizer's coming to *believe* the newly constructed hypotheses in question. It is enough that they enter the hypothesis space as live candidates for belief, or to put it in Bayesian terms, that they acquire some subjective probability, however low. It is, after all, one thing to acquire the concept of *C*, and quite another thing to believe that the concept is instantiated—that the world actually contains some *C*s. We all have the concept of Newtonian force, but in the wake of modern physics, we are entitled to doubt that there really is any such thing.

6. Shallow Concepts

The purpose of this paper is to urge the viability of what I call shallow theories of concepts, that is, theories on which a concept's possessor does not represent (or pretend to represent) the ultimate grounds of category membership.

There are many shallow theories to choose from: the shallow prototype theory, causal minimalism (for concepts of biological and chemical kinds), and so on. In fact, for every deep theory of concepts there is a corresponding shallow theory, obtained by stripping away the stipulative or constitutive status of the "deep" beliefs while leaving their content otherwise intact. Thus the shallowness or depth of a theory of concepts is quite orthogonal to its other content.

Let me give two examples. First, take the classical Lockean account of the swan concept, on which swans are defined as white-feathered, red-beaked, trumpeting, aquatic birds (or something similar). What the classical account identifies in the head of the concept's possessor can be divided (though it is not usually so divided!) into two parts:

1. A set of universally quantified beliefs: *All swans have white feathers*, *All swans have red beaks*, and so on.

2. A stipulation that the swan category contains just those entities satisfying the beliefs. Thus an animal is defined to be swan just in case it is white-feathered, red-beaked, and so on.

It is the stipulation in (2) that turns the beliefs in (1) into a definition, and so into a final criterion for swanhood. Remove the latter belief, and you have a shallow concept of swanhood on which the beliefs about feather color, beak color and so on are in no way regarded as constitutive of swanhood. Note that a cognizer with the shallow concept will by and large classify the same things as swans as a cognizer with the deep concept. But they will not regard themselves as doing so on grounds specified by a final criterion for swanhood, and they will consequently keep open the possibility that their universal generalizations about swans are incorrect.

Second, take the "lightweight" psychological essentialist account of the swan concept illustrated in figure 1. What the psychological essentialist identifies in the head of the concept's possessor can also be divided into several parts:

1. A universally quantified belief : all swans have an unobservable property *C*.

2. A set of causal beliefs: *C causes white feathers*, *C causes red beaks*, and so on.

3. A stipulation that the swan category contains just those entities that have *C*.[16]

Remove the stipulation (3), and you have a shallow version of psychological essentialism (which for that reason, perhaps ought not to be called "essentialism"). A "shallow essentialist" will make more or less the same categorizations as a deep essentialist, but they will keep open the possibility that some swans may not have *C*, or that some creatures with *C* are not swans (without necessarily knowing how to take advantage of this possibility).

For every deep theory of concepts, then, there is—to repeat—a corresponding shallow theory. Whatever beliefs about a category a deep theory attributes to the cognizer, its shallow counterpart also attributes, with the sole exception of beliefs about the grounds of category membership. The shallow counterpart will therefore predict many of the same inferential behaviors as the deep theory—many of the same categorizations, inductions, analogies, and so on.

On what grounds, then, to choose between them? The deep theory's final criterion gives it a certain psychological fixedness, over and above the groundedness and stability that exists even on a shallow theory (section 4). Shallow concepts, unanchored, may drift more—not aimlessly, however, but in the direction of nature, in that the impact of new evidence on shallow psychological categories tends to shape them so that they better mirror the structure of the external world, as the case of the shallow prototype theory suggests (section 4.3). Why should this be? It is because all there is to a shallow category is an empirical theory, and the impact of new evidence on an empirical theory tends to push it toward veridicality.

There is something of a paradox here. If a psychological category that reflects the natural or objectively most inductively informative category is in

16. To get "rich" psychological essentialism, further stipulate that the beliefs in (2) are constitutive of swanhood, that is, that to be swan, you must not only have *C*, but also the beliefs in (2) must be true of *C*.

some sense the most philosophical or deep category you can have, then the route to deep, philosophical categories is not deep, philosophical concepts—that is, concepts that incorporate a final criterion— but rather shallow concepts. To put it another way, if you want your categories to be as "metaphysically valid" as possible, at all costs avoid putting metaphysics into your concepts. Or rather, avoid putting *your* metaphysics into the concepts; let nature, by way of your inductive inferences, fill in her own metaphysics instead.

For this reason, the great cognitive engineer in the sky might be expected to prefer to equip his creations with shallow concepts. Only empirical research—the sorts of evidence and arguments found in Strevens (2000) and Rips (2001)—can tell us whether humans represent final criteria for category membership, and so whether the depth thesis is as mythical as I have made out. But the possibility that cognition is largely or entirely shallow should not, I contend, be disregarded.

7. Some Lessons for the Study of Concepts

In this final section, let me turn to a more practical topic, describing some 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 concepts 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.

What I am criticizing in the work surveyed below is not the particular position that is taken, but a failure to notice other, equally viable rival views. All of the following errors, then, are errors of neglect—neglect, in particular, of the possibility that important facets of categorization and inference might not depend very much or at all on a final criterion.

The first and original 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 approaches to concepts have, in their canonical form, made this mistake.

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, our concepts must be "organized around causes" (139), which I take to mean that causal facts must appear in the concepts' final criteria. (Ahn cannot merely intend to say that we have many causal beliefs about our categories, as she takes her view to be inconsistent with psychological essentialism. The disagreement must concern the role of the causal facts in determining category structure.)

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 lightweight essentialism) 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.)

Or consider some 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.[17]

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 Keil's subjects' positing of a raccoon essence that plays the preeminent role in the explanation of the transformation 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 of our most distinctive reasoning about 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 Carey's (1985) view that children do not possess natural kind concepts until they possess the concept of a biological essence:[18]

> 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).

In this passage, Carey envisages only two possibilities: an empiricist psychology, whether in the style of Locke or Rosch, or a psychology based on deep

17. On this issue, including a defense of the prototype view, see Hampton (2007).
18. Carey (2009) appears to have relinquished this assumption.

ontology. But this is a false dichotomy: categorizers may have a scientifically sophisticated but philosophically innocent theory of biological taxa, as on the causal minimalist view. More generally, it is possible to have a rich set of beliefs about a category, generating subtle inferential behaviors, without any commitment to a thesis about the metaphysics of—the ultimate grounds of—category membership.

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, for instance, to judge that a certain spiky-leafed tree is a California-nutmeg 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 a writing instrument 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 grounds for membership in artifact categories, Sloman and Malt seem to

be saying, change with the context. Such a strong conclusion is hardly justified by the evidence. Once you see that category judgments are regular inductive inferences, you should expect to find context-sensitivity in categorization regardless of the nature of the ultimate grounds for category membership.

To conclude, let me say again that my principal aim is to expand the field of serious possibilities for theorizing about concepts. First, the characteristic behaviors of concepts are as likely to derive from empirical beliefs about the corresponding categories as from category structure determined by a final criterion. Second, there is a real possibility that final criteria are entirely absent in some or most concepts, in which case the very notion of the structure of a psychological category rests on a false presupposition.

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.

Ahn, W., C. Kalish, S. A. Gelman, D. L. Medin, C. Luhmann, S. Atran, J. D. Coley, and P. Shafto. (2001). Why essences are essential in the psychology of concepts: Commentary on Strevens (2000). *Cognition* 82:59–69.

Boyd, R. (1988). How to be a moral realist. In G. Sayre-McCord (ed.), *Essays on Moral Realism*. Cornell University Press, Ithaca, NY.

Carey, S. (1985). *Conceptual Change In Childhood*. MIT Press, Cambridge, MA.

———. (2009). *The Origin of Concepts*. Oxford University Press, Oxford.

Fodor, J. (1981). The present status of the innateness controversy. In *Representations*, pp. 257–316. MIT Press, Cambridge, MA.

Fodor, J. A. (1998). *Concepts: Where Cognitive Science Went Wrong*. Oxford University Press, Oxford.

Gelman, S. A. (2003). *The Essential Child: Origins of Essentialism in Everyday Thought*. Oxford University Press, Oxford.

Gelman, S. A., J. Coley, and G. Gottfried. (1994). Essentialist beliefs in children: The acquisition of concepts and theories. In L. Hirschfeld and S. A. Gelman (eds.), *Mapping the Mind*, pp. 341–365. Cambridge University Press, Cambridge.

Gelman, S. A. and H. M. Wellman. (1991). Insides and essences: Early understandings of the non-obvious. *Cognition* 38:213–244.

Hampton, J. A. (2007). Typicality, graded membership, and vagueness. *Cognitive Science* 31:355–384.

Hull, D. (1978). A matter of individuality. *Philosophy of Science* 45:335–360.

Keil, F. C. (1989). *Concepts, Kinds and Conceptual Development*. MIT Press, Cambridge, MA.

Lewis, D. (1970). How to define theoretical terms. *Journal of Philosophy* 67:427–446.

Locke, J. (1975). *An Essay Concerning Human Understanding*. Edited by P. Nidditch. Oxford University Press, Oxford.

Malt, B. (1994). Water is not H_2O. *Cognitive Psychology* 27:41–70.

Margolis, E. (1998). How to acquire a concept. *Mind and Language* 13:347–369.

Mayr, E. (1970). *Populations, Species, and Evolution*. Harvard University Press, Cambridge, MA.

Murphy, G. L. (2002). *The Big Book of Concepts*. MIT Press, Cambridge, MA.

Osherson, D. N. and E. E. Smith. (1981). On the adequacy of prototype theory as a theory of concepts. *Cogntion* 9:35–58.

Putnam, H. (1975). The meaning of 'meaning'. In K. Gunderson (ed.), *Language, Mind and Knowledge*, volume 7 of *Minnesota Studies in the Philosophy of Science*. University of Minnesota Press, Minneapolis.

Rips, L. J. (2001). Necessity and natural categories. *Psychological Bulletin* 127:827–852.

Rips, L. J., S. V. Blok, and G. Newman. (2006). Tracing the identity of objects. *Psychological Review* 113:1–30.

Rosch, E. (1978). Principles of categorization. In E. Rosch and B. Lloyd (eds.), *Cognition and Categorization*, pp. 27–48. Lawrence Erlbaum, Hillsdale, NJ.

Rosch, E. and C. B. Mervis. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* 7:573–605.

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. *Language and Cognitive Processes* 18:563–582.

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 commments on Ahn et al. Electronic document available at http://www.strevens.org/research/cogsci/ahnetal.pdf.