The Essentialist Aspect of Naive Theories

Michael Strevens

Cognition 74:149–175, 2000

Abstract

Recent work on children's inferences concerning biological and chemical categories has suggested that children (and perhaps adults) are essentialists—a view known as psychological essentialism. I distinguish three varieties of psychological essentialism and investigate the ways in which essentialism explains the inferences for which it is supposed to account. Essentialism succeeds in explaining the inferences, I argue, because it attributes to the child belief in causal laws connecting category membership and the possession of certain characteristic appearances and behavior. This suggests that the data will be equally well explained by a non-essentialist hypothesis that attributes belief in the appropriate causal laws to the child, but makes no claim as to whether or not the child represents essences. I provide several reasons to think that this non-essentialist hypothesis is in fact superior to any version of the essentialist hypothesis.

Keywords: psychological essentialism, naive biology, concepts

Our theories shape the way we conceive of the world. Recent psychological work on concepts and conceptual development has suggested that several of these theories are, in some sense, essentialist, at least in their naive (prescientific) forms. This view is called the hypothesis of *psychological essentialism* or simply the *essentialist hypothesis*.[1]

The aims of this paper are

[1] Throughout this paper, an "essentialist" hypothesis is, of course, a hypothesis that attributes essentialism to children (and perhaps some adults), not a hypothesis that is itself committed to essentialism.

1. To distinguish several different varieties of psychological essentialism,
2. To argue that one of these is superior to the rest, and
3. To argue that there is a non-essentialist hypothesis better supported by the data than any essentialist hypothesis. Rather than attributing beliefs about essences to children, the non-essentialist hypothesis attributes beliefs about causal laws. As a result, I will show, it better explains the experimental results that have previously been thought to support essentialism.

What is the evidence for psychological essentialism? The hypothesis seems to be the best explanation of certain patterns of inference found in children and adults. The patterns in children include those described in S. Gelman and Markman (1986), Keil (1989), S. Gelman and Wellman (1991), and others cited below. The patterns in adults are described in Putnam (1970) and Rips (1989), among others. I will be most concerned with the patterns found in children's thinking about natural kinds, with emphasis on the kinds of naive biology, the subject of most of the relevant research.[2] Because these patterns of inference are associated with thinking about natural kinds, I will call them the *K-patterns* of inference. Most of this paper will be concerned with the ways in which different varieties of psychological essentialism explain, and sometimes fail to explain, the K-patterns.

1. Varieties of Essentialism

I will distinguish three varieties of essentialism that have been attributed to children, and I will add a fourth hypothesis of my own that is not essentialist at all.

[2] Relevant research on naive biology includes Atran (1990), Medin and Atran (1999), some of the studies just mentioned—S. Gelman and Markman (1986), Keil (1989), and Gelman and Wellman (1991)—as well as much other work cited below. Other naive theories that may be essentialist are our naive "chemistry" concerning kinds of substances (Gelman, 1988; Keil, 1989) and some theory or theories that constitute part of our understanding of social relations (Rothbart & Taylor, 1994; Hirschfeld, 1996). It is unclear whether any further naive theories, such as those of physics and psychology, are essentialist (see Gelman and Hirschfeld (1999) for a discussion). At least one writer has suggested that *all* of our concepts—including, for example, the concept of a wastebasket—are founded in essentialist theories (Medin, 1989, 1477). But there is strong evidence that many of the characteristics of essentialist thinking are not to be found in our reasoning concerning artifacts such as wastebaskets (Keil, 1989).

In what follows, I will use the notion of representation in a broad sense, so that it may be said that we represent something even if we know nothing about it (except, perhaps, that it exists). In particular, a child may be said to represent essences even if she does not have any beliefs about the nature of essences.

1.1 Pure Essentialism

A clear and concise version of the claim that naive theories are essentialist is made by S. Gelman, Coley, and Gottfried (1994):

> People seem to assume that categories of things in the world have a true, underlying nature that imparts category identity . . . The underlying nature, or category *essence*, is thought to be the causal mechanism that results in those properties we can see. For example, the essence of tigers causes them to grow as they do—to have stripes, large size, capacity to roar, and so forth. (p. 344)

In this passage, and in other papers by Gelman and her collaborators, a hypothesis is advanced that can be unbundled into the following three connected claims.

1. Some naive theories posit the existence of essences, though they may not represent what sorts of things essences are.

2. Essences are represented as what define (at least some of) the categories of a theory, in the sense that possession of the essence is represented as necessary and sufficient for category membership. In the example, it is believed that an organism is a tiger if and only if it has a certain essence.

3. Essences are represented as being causally responsible for certain observable properties; in the example, the essence causes the tiger's stripes, size, and so on.

I will call this hypothesis the *pure essentialist hypothesis*.[3]

[3] Note that the essentialist hypotheses considered in this paper do *not* ascribe to the thinker some of the standard posits of metaphysical essentialism, such as the belief that essences are fixed and immutable.

Pure essentialism may be represented in pictorial form as a diagram imputing a certain structure to a naive theory. A fragment of naive biology, according to the essentialist hypothesis, has the structure shown in figure 1.

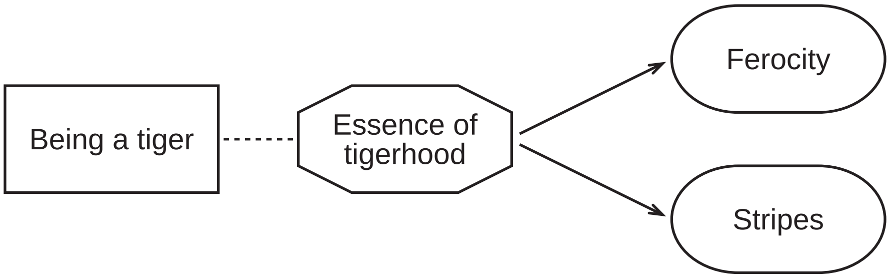

Figure 1: A pure essentialist theory

In this figure, the octagonal box represents a kind-defining essence, the rectangular box the kind it defines, and the rounded boxes characteristic observable properties of that kind. The arrows represent causation. The dashed line represents a semantic, or defining, link: an animal is considered to be a tiger—that is, to be a member of the natural kind "tiger"—if and only if it has the essence of tigerhood. I should emphasize that on the pure hypothesis, the naive essentialist may have no opinion as to what essences actually are. Their essentialism consists in their thinking that essences exist—whatever they may be—and that they play the category-defining and causal roles illustrated in figure 1.

1.2 Statistical Essentialism

Another version of the essentialist hypothesis appears in Medin and Ortony (1989). Medin and Ortony agree with the pure essentialist hypothesis that, although humans may not know anything about a category's essence, they believe that there is such an essence.[4] The authors say that in such a situation, the thinker's theory has an *essence placeholder*.[5] This placeholder is eventually replaced by a representation of the nature of the essence, if such knowledge is acquired.

[4] "...we are claiming...that people find it natural to assume, or act as though, concepts have essences" (Medin & Ortony, 1989, 184).

[5] "...we propose that the knowledge representations people have for concepts may contain what might be called an essence placeholder" (p. 184).

Medin and Ortony also propose that humans posit links between essences and observable properties. These links are often, but not always, represented as causal.[6] If they are not causal, they may be merely statistical. (Medin and Ortony do not enumerate all the possibilities.)

The major difference between pure essentialism and Medin and Ortony's essentialism concerns the link between essence and category membership. According to pure essentialism, having an essence is represented as being necessary and sufficient for membership of the corresponding category (tenet 2). Medin and Ortony disagree:

> . . . it may be part of the represented essence of bird that birds fly, even if it happens that not all birds do fly and that people know this. (Medin & Ortony, 1989, 184)

For Medin and Ortony, then, having the bird essence is not necessary for membership of the category of birds. They do not say whether it is sufficient, nor do they say what relationship an object must bear to an essence in order to belong to the category.

The three tenets of Medin and Ortony's psychological essentialism, then, are the following:

1. Some naive theories posit the existence of essences, though they may not represent what sorts of things essences are. (Same as tenet 1 of pure essentialism.)

2. Essences are not represented as necessary, and perhaps not represented as sufficient, for category membership. (The denial of tenet 2 of pure essentialism.)

3. Essences are represented as being causally responsible for *or* statistically correlated with certain observable properties. Other kinds of links between essences and observable properties may also be possible. (A weakened version of pure essentialism's tenet 3: the links may be, but do not have to be, causal.)

I will call this hypothesis the *statistical essentialist hypothesis*. Statistical essentialism attributes to children theories with the structure represented in figure 1, except that some or all of the links (both between category and essence and between essence and observable properties) may be statistical.

[6] "[The theories that are structured around essence placeholders] often provide or embody causal linkages to more superficial properties" (p. 186).

1.3 Internal Essentialism

The pure and statistical essentialist hypotheses allow that children may represent the existence of essences while having no beliefs at all about the sorts of things essences are (although they may well acquire such beliefs as they grow older).

What I will call *internal essentialism* is pure essentialism with one additional tenet: from the time they start making K-patterned inferences, children have a firm belief about the nature of essences, namely, that an essence is a property of some or all of an entity's insides. It might be the entire insides, or just something buried within (the heart or the DNA, say). The hypothesis allows that (and it is expected that) children do not know where the essential part of the insides is located or what property of that essential part is the essence. What it rules out, unlike pure essentialism, are possibilities such as the following: (a) children represent the essence of a tiger as being a property of the tiger's skin (or as any aspect of the tiger's external appearance), (b) children have no beliefs about the location or nature of essences, and (c) children represent the essence as some aspect of the causal history of the tiger, such as a fact about its parents or its evolutionary lineage.

It is often unclear whether advocates of psychological essentialism are pure essentialists or internal essentialists, but essences are often described as "deep" or "underlying" (see for example Atran (1995, 219–20)), and many experiments hint that insides play an important role in generating the K-patterns of inference (for example, S. Gelman and Wellman (1991)). Related research has suggested that children see insides as doing some of the work that pure essentialism attributes to essences (R. Gelman, 1990; S. Gelman & Kremer, 1991). Thus whether or not internal essentialism has previously been clearly distinguished from pure essentialism, it is, so to speak, "in the air".

The tenets of internal essentialism are the following:

1. Some naive theories posit a special role for a certain key property (or properties) of an entity's insides, the *essential property*, though they may not represent which property this is.

2. Essential properties are represented as what define the categories of a theory, in the sense that possession of the essential property is represented as necessary and sufficient for category membership.

3. The essential property of an entity's insides is represented as being causally responsible for certain observable properties.

Internal essentialism, then, attributes to children theories with the structure represented in figure 2. It may not seem that the differences between pure essentialism and internal essentialism are very great, but I will show that internal essentialism has certain advantages over pure essentialism and also certain serious disadvantages.

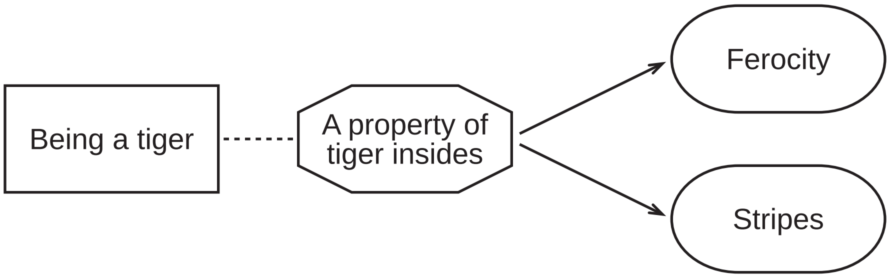

Figure 2: An internal essentialist theory

1.4 The Minimal Hypothesis

In the next section I will examine the way in which the essentialist hypotheses explain the K-patterns. I will show that the essentialist hypotheses have their explanatory power because they attribute to the child belief in certain causal laws, namely, causal laws connecting kind membership with observable properties. An example is the law that tigers have stripes. I emphasize that this is to be understood as a *causal* law. It is not just that, statistically, tigers tend to have stripes. Rather, *there is something about being a tiger that causes tigers to have stripes*. I take this formulation to be equivalent to *it is a causal law that tigers have stripes*. I will call these laws *K-laws*. (The 'K' is, once again, for 'kind'.)

It is clear that both pure and internal essentialism attribute belief in K-laws to the child. (Insofar as statistical essentialism fails to do so, I will later show, it proves unable to explain the K-patterns.) According to pure or internal essentialism, the child believes that whatever is a member of a kind has a certain essence, and that this essence causes the observable properties associated with the kind. It follows that the child believes that there is something about the kind that causes the observable properties, or in other words, that it is a causal law that members of the kind have the properties.

Now suppose, as I have asserted, that the essentialist hypotheses explain the K-patterns solely in virtue of the fact that they attribute belief in K-laws to the child. (I argue that this is so in the next section.) Then essences play no direct role in the explanation. To infer in accordance with the K-patterns, you must believe that there is something about tigers that causes them to have stripes, but you do not have to believe that this something is an essence. For example, you might have no opinion about what does the causing, or you might think that a mechanism that is not an essence does the causing (as in modern biology), or you might think that it is just a brute fact about the world that being a tiger causes an animal to grow stripes.

It follows that the K-patterns would be equally well explained by a number of hypotheses other than psychological essentialism, one hypothesis for each way in which a thinker can believe causal laws about tigers without attributing the operation of the laws to the causal powers of an essence. I will champion the simplest of these hypotheses: that children believe there are causal laws connecting natural kinds and their observable properties, but that they are not committed to any particular view about the implementation of these laws. In particular, children do not represent essences as underlying the laws. I call this the *minimal hypothesis*.

The minimal hypothesis, like the various essentialist hypotheses, may be represented by a diagram of naive theory structure (figure 3). The arrows are, as before, representations of causation. It will be seen that a child with the sort of essentialist theory shown in figure 1 and figure 2 will, in general, make exactly the same inferences as a child with the minimal causal theory shown in figure 3.[7] Thus if the pure and internal essentialist hypotheses explain the K-patterns, the minimal hypothesis does so too.

2. How the Essentialist Hypothesis Explains the Data

In what follows it is convenient to lump together pure, statistical, and internal essentialism, wherever they agree, as "the essentialist hypothesis". Where they do not agree they are distinguished and the merits of each are

[7] The exceptions are those inferences that are actually about essences or insides. In section 4 I will show that what we know of these inferences is better explained by the minimal hypothesis than the essentialist hypotheses.

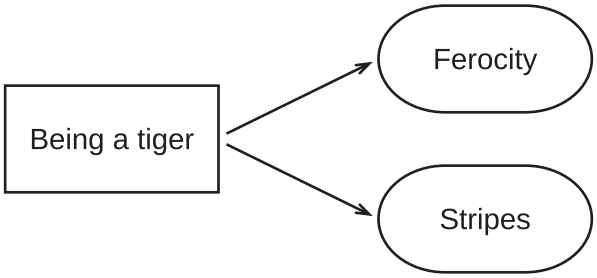

Figure 3: A minimal causal theory

discussed. Treatment of the minimal hypothesis is deferred until section 3.

I asserted in the last section that the essentialist hypothesis explains the K-patterns of naive inference because it attributes to the child belief in causal laws connecting kinds and their observable properties. What is more, this is the *only* way the hypothesis can explain the K-patterns. To argue for this proposition is my task in this section.

Before I begin, an important point: the bare fact that children represent essences cannot explain any of their inferences at all. Nor can any appeal to essence placeholders. A naive theory structures inference not in virtue of its elements (such as essences or essence placeholders), but in virtue of the connections between those elements. The essentialist explanation of the K-patterns is to be found in these connections.

In this section, then, I will show how the connections in an essentialist theory give rise to K-patterned inferences. It will emerge that the thesis that children represent causal laws connecting kinds and observable properties, together with some other posits, does all the explanatory work that the essentialist theory is capable of doing. None of these explanatory posits involves the representation of essences.

Let me stress that the following explanation is offered as much on behalf of the essentialist hypothesis as on behalf of my own minimal hypothesis. Insofar as any essentialist explanation of the K-patterns has appeared in the literature (and what has appeared has been very sketchy), it has been along the lines of the explanation offered below. Psychological essentialism cannot be defended from the minimal hypothesis by arguing that this kind of explanation is flawed, for if it is flawed, that is just as much a problem for psychological essentialism as it is for the minimal hypothesis.

2.1 Projection

K-patterned inferences are of two sorts: *projections* (sometimes referred to in the psychological literature as *inductions*) and *categorizations*. A projection is an inference concerning an as yet unobserved property of a given object. An example is our inferring that a sleeping tiger is ferocious. A categorization is an inference concerning the category of a given object. An example is our recognizing a ferocious striped creature as a tiger. I will deal with the explanation of K-patterned projections first.

Susan Gelman and her collaborators have created experimental situations in which the subject must choose between two projective inferences, one a projection according to kind, one a projection according to some other inferential cue, either similarity in appearance (S. Gelman & Markman, 1986; S. Gelman & Markman, 1987; S. Gelman & Coley, 1990), or similarity in environment and upbringing (S. Gelman & Wellman, 1991). The youngest children studied in the kind versus upbringing task were four years old; for the most part, they preferred to project by kind. The youngest children studied in the kind versus appearance task were two years old; even they preferred to project by kind.

S. Gelman and Markman (1986) showed children pictures of two squirrels and a rabbit. One of the squirrels was a gray squirrel, the other was a kaibab, a kind of squirrel that looks like a rabbit. Children were told that the gray squirrel eats bugs and that the rabbit eats grass. They were then told that the kaibab is a kind of squirrel, and asked what the kaibab eats. A majority of four-year-olds said that the kaibab eats bugs. Children's K-patterned projections, then, attribute new properties on the basis of natural kind rather than appearance, when the two conflict.

The essentialist hypothesis explains this projection as follows. In response to the kaibab scenario, the children (a) represent the kaibab as being a squirrel (since they are told that it is one), (b) infer that the kaibab has the squirrel essence, and (c) infer (from their belief that the squirrel essence causes bug-eating) that squirrels eat bugs. The inferences in (b) and (c) are based on, respectively, pure essentialism's tenet 2 and tenet 3.[8]

[8] K-patterned projections may be made even when the child is not told the kind (S. Gelman & Markman, 1987). It seems that in such cases, the child infers the kind from other information; Gelman and Markman obtained evidence that in their study, as expected, the information in question was the picture of the organism that they had shown to the children. (Thus a categorization from appearances plays a role in the inference.) The inference then

Appearances do not enter into the inference at all. Less obviously, essences are not important either. What is important in accounting for the K-patterned projections is that children believe that something about being a squirrel causes an animal to eat bugs. That the something is an essence adds nothing to the strength of the inference. Representations of essences merely provide one way, among others, of representing the facts that do warrant the inference, the K-laws. (Some ways of representing K-laws without representing essences were described in section 1.4.)

This explanation goes through smoothly for pure and internal essentialism, but not for statistical essentialism. Because statistical essentialism denies tenet 2 of pure essentialism, it allows that children may believe that some members of a category lack the category's characteristic essence. For example, children may believe that not all squirrels have the squirrel essence. But this essence is what causes bug-eating. Thus the squirrels that lack the essence may not eat bugs. If children believe this, it is unclear why they say that kaibabs do eat bugs, especially since the oddly shaped kaibabs are prime candidates for being essenceless. Statistical essentialism, I conclude, does not provide a very satisfying explanation of K-patterned projections.

The results of Gelman and Wellman's kind versus upbringing studies (Gelman & Wellman, 1991) may be explained in a similar way to the kind versus appearance results, at least by pure and internal essentialism. In one scenario, children were told of a baby cow that was raised by pigs. They were asked whether the cow, when it grew up, would have a straight tail or a curly tail, and whether it would say "moo" or "oink". On the whole, children predicted that the baby would grow up to exhibit cow, not pig, appearances and behavior. As in the kind versus appearance study, the children were provided with a piece of information—the animal's membership of a natural kind—which enabled them to make a prediction based on K-laws immediately, without attending to upbringing.

2.2 Categorization

Most of the evidence for K-patterns in categorization comes from the work of Frank Keil and his collaborators, reported in Keil (1989). The experimenters asked children to make categorizations in which "surface" information is pitted against "deep" information. (Biological categories predomi-

proceeds as described in the main text, based on the inferred knowledge of kind.

nate, with chemical categories making up the balance.) Children were presented with two kinds of scenarios:

Discovery: Children were told about some organism or substance that had the surface properties of one kind, but was related to another kind in a "deep" way. For example, a scenario was presented in which animals had the appearance and behavior of horses, but the insides and lineage of cows. (To have the lineage of a species is to have parents and children of that species.) The children were then asked whether these animals were horses or cows. Children of seven and older tended to categorize in accordance with the "deep" properties, saying that the animals were cows.

Transformation: Children were told about an organism or substance of one kind that had been cosmetically transformed (with paint, perfume and so on) so that it took on the appearance of another. They were shown appropriate "before" and "after" pictures. For example, a scenario was presented in which a raccoon had been made to take on the appearance of a skunk by restyling and dying its fur and adding the distinctive odor of "super smelly yucky stuff". Children were then asked if the animal was a skunk or a raccoon. Seven-year-olds usually claimed that, contrary to appearances, it was still a raccoon. In further experiments, the nature of the transformation was systematically altered. It was made either more superficial than a cosmetic change (a zebra was dressed up in a horse costume) or deeper (a young horse was given an injection that made it grow zebra stripes). The more superficial the change, it turned out, the younger the children who asserted that the change had not affected kind membership. Even preschoolers showed a tendency to ignore a costume change.

Knowing that a particular object exhibits appearances and behavior typical of a particular kind will normally cause both adults and children to infer that the object belongs to the kind. For one who believes the K-laws, however, such an inference is defeasible, that is, it may go from being reasonable to being unreasonable in the light of further information. This is because it does not follow from the fact that it is a causal law that tigers are catlike, ferocious and have stripes, that any ferocious, striped, catlike animal is a tiger. It does not even follow if one adds the information that tigers are the only animals for which such a causal law holds. This is because the appearance of a given animal may not have been caused by a biological law at all, as in the case where the animal's stripes are painted on. Thus information that the appearances were not caused in accordance with a biological law defeats the inference from appearance to kind (that is, makes the inference no longer reasonable). The essentialist hypothesis exploits this observation about the logic of causal inference to explain the K-patterns of categorization.

2.2.1 The Discovery Experiment

In the discovery experiment, everyone is agreed, children do not follow their inclination to make a categorization based on appearances because they consider the conflicting categorization based on lineage and insides to be more strongly warranted.

This fact about children cannot be explained by their belief in the K-laws alone. To explain the K-patterns, the proponents of the essentialist hypothesis must assume that the child represents certain facts about organisms' lineage and insides. The assumptions that must be made are:[9]

> *Lineage hypothesis:* Children believe that members of a kind invariably have children of the same kind, and
>
> *Insides hypothesis:* Children believe that members of a kind invariably have insides with certain kind-specific properties ("kind-specific" in the sense that things with the insides invariably belong to the relevant kind).

The word "invariably" here signals that inferences from insides or lineage to kind are not defeasible. In the case at hand, the child comes to believe that it is impossible that a tiger should fail to have tiger insides, that the child of a tiger should fail to be a tiger, and so on.

The beliefs attributed by the lineage and kind hypotheses license inferences from lineage or insides to kind membership. These inferences, I am supposing, are not treated by the child as defeasible. Thus in the discovery experiments, it is the defeasible inference—the inference from appearances—that will have to give way. (Note that there is no need to weight the two kinds of inference; they are qualitatively different in their logical status. Of course, one might alternatively hypothesize that children do not see lineage and insides as invariably correlated with kind, but merely regard them as a more reliable guide to kind than surface features.)

[9] For research on children's theories of inheritance, see Springer and Keil (1989) and Springer (1992). For work on children's theories of organisms' insides, see R. Gelman (1990) and S. Gelman and Kremer (1991). For the bearing of some of this research on essentialism, see section 4.2.

The three varieties of essentialism under discussion are all able to explain the discovery experiments in this way (and, I think, in no other way), but they do so with differing degrees of success. Prima facie, internal essentialism is in the strongest position, as it already incorporates the insides hypothesis. All and only tigers have the tiger essence, which is a property of tiger insides. Therefore all and only tigers have tiger insides. Pure and statistical essentialism allow that children may not represent essences as being in any particular part of the animal (and may not even represent essences as being a physical part of the animal at all), so they must posit the insides hypothesis as an additional assumption.

There is one major disadvantage to internal essentialism, however. Internal essentialism predicts that children will make the inferences they do in the discovery experiment influenced entirely (or at least largely) by what they are told about animals' insides. Transcripts of conversations that the experimenters had with their subjects after the questions were asked, however, do not bear out this prediction. Children seem to have been influenced roughly equally by lineage and insides. This point is developed into an argument against internal essentialism in section 4.2.

2.2.2 The Transformation Experiments

The transformation experiments differ from the discovery experiments in that the subjects are given no explicit information about the deep properties of the entity *after* the transformation. The children's reasoning in these experiments can be reconstructed in two rather different ways:

Inference from the superficiality of the transformation: On the first reconstruction, children's K-patterned inferences about transformations are similar to their inferences about discoveries, only a little more complex. It is assumed that older children know, or can work out, that cosmetic changes and the like, although they alter an animal's outsides, do not alter the insides. A raccoon made up as a skunk retains its raccoon insides. By the insides hypothesis (see above), what has raccoon insides must be a raccoon. The transformed animal retains its original kind. This inference is not defeasible, and so trumps the defeasible inference from appearances to membership of the new kind. Thus in cases such as this, older children consider that changes to an entity's outsides do not alter its kind.

Inference from the artificiality of the transformation: On the second reconstruction, the inference from transformed appearance to category is defeated because the appearance was brought about in an *unnatural* way, that is, because it was not brought about by the K-laws. Consider the biological cases. At some stage children must acquire a simple theory of growth, which specifies, rather broadly, the way in which an animal naturally acquires its characteristic properties. (Among perhaps other things, it narrates the story of a skunk's coming to have its adult appearance: mother skunks give birth to baby skunks, which start out small, lacking some adult features, and gradually become larger and develop those features, with no outside assistance.)[10] Clearly, Keil's transformations do not conform to this story. It follows that the organism's new appearance is not caused by the K-laws. In the case at hand, the skunk-like animal does not get its skunk likeness in the way that real skunks do. Rather, the appearance is caused by an unnatural or abnormal process. As shown above, artificially induced appearances are no evidence for category membership.

Of these two, the explanation from superficiality is the natural explanation for the internal essentialist, since all that is needed for the explanation to go through is the insides hypothesis, already part of the internal essentialist's theory. I will argue, however, that the explanation from superficiality cannot account for Keil's results concerning transformations of differing superficiality (Keil, 1989, chap. 11).

I will describe these experiments in more detail. The pictures used in the experiments (e.g., of the raccoon and the "skunk") were the same as in the original transformation studies. What differed was the story as to how the animal in the first picture came to have the appearance it did in the second picture. There were four kinds of story told about the transformation of appearances. They were, in order of decreasing superficiality:

Costume Change: The animal in the first picture is wearing a costume in the second picture. In one story, the tigers at a circus get sick, so the trainer puts the lions in tiger costumes.

Temporary Cosmetic Change: This is a cosmetic change (see next item) that is specified in the story to be temporary (e.g., the paint falls off, and must be reapplied each day).

[10] For pertinent work on the child's theory of growth, see Rosengren, Gelman, Kalish, and McCormick (1991).

> *Cosmetic Change:* This is the sort of change described to the children in the original transformation experiments. It involves makeup, paint, perfume, and so on.
>
> *Internal Change:* The animal is given pills or injections that cause it to take on the appearance in the second picture. The child is not usually told what sort of substance is being administered.

The experiment was conducted on four groups, a group of adults and three groups of children aged approximately five, seven, and nine. The more superficial the change, the younger the children who classified against appearances. Five-year-olds usually ignored the effects of costume changes, and sometimes ignored the effect of temporary cosmetic changes. Seven-year-olds almost always ignored changes in appearance brought about by costume changes and temporary cosmetic changes, but only sometimes ignored the effects of permanent cosmetic changes and internal changes. Nine-year-olds almost always ignored the effects of costume changes and both temporary and permanent cosmetic changes, but only sometimes ignored the effects of internal changes. Adults almost always ignored all changes.

The explanation from superficiality cannot easily account for these results. It is surely clear to children that costume changes and both temporary and permanent cosmetic changes leave insides unaffected. So the explanation from superficiality predicts, contrary to the facts, that children's performance will be the same whichever of these three causal mechanisms is described.

The explanation from artificiality can account for the data in the following way. As children's narratives concerning growth become more sophisticated, they are increasingly better able to distinguish between natural and unnatural changes in appearance and behavior. (Recall that a natural change is a change that occurs in virtue of the animal being the kind of animal it is, that is, in virtue of a K-law.) Even young children know that a costume change is unnatural; as they learn more about the world, they gradually come to see that other changes (cosmetic and then internal) are unnatural too, and so that appearances transformed by such operations are no basis for inferring a change in biological category.[11]

[11] A reader has objected to this explanation on the grounds that it is plausible that even young children will assume that *all* the changes Keil describes, including the changes induced by pills and injections, are unnatural, since children do not themselves experience

2.3 Comparison of Projection and Categorization

Given the vague claim that children are in some sense essentialists, there is a puzzling discrepancy between Keil's results concerning categorization and the results of Gelman, Markman and others concerning projection. K-patterns appear in children's projections as early as two years of age; in categorization, however, K-patterns show up only around seven years of age (although the costume change transformation elicits K-patterned inferences in five-year-olds).

S. Gelman, Collman, and Maccoby (1986) note that one sort of inference involves projection, the other categorization, and propose that projection is easier for children than categorization. In the light of the discussion above, this does not seem quite correct. Projection is not easier than categorization overall. Rather, Gelman's K-patterned projections are easier than Keil's K-patterned categorizations, for the reason that Keil's categorizations, as opposed to Gelman's projections, require the following of the child:

1. Knowledge of certain auxiliary theories (concerning growth, lineage, and insides), and the wherewithal to bring these theories to bear on the question, and

2. The ability to deal with a complex form of inference, namely, that in which a defeasible inference is in fact defeated.

One way to test this suggestion would be to experiment with projections that make use of auxiliary theories and in which inferences are defeated.[12] If the results were similar to Keil's results—if children had similar problems with these questions at similar ages—it would be reasonable to conclude that it is not the category of inference (projection or categorization) that makes

such changes in the process of growth. It seems to me to be highly unlikely that children base their beliefs about what is natural on their own experience. First, children are constantly given costume changes and are often given pills and injections. So these things *are* part of their experience. Their (eventual) opinion that such changes are unnatural must be based on something other than unfamiliarity. Second, children come to believe that, for example, a bird's hatching or a caterpillar's metamorphosis are natural, even though they never hatch or metamorphose themselves.

[12] An example: if we inject a baby tiger with a special substance taken from leopards, will it grow up to have spots or stripes? This question is not ideal, but it might create some age-related effects.

the difference, but the level of logical and theoretical sophistication required by the particular inference.

It is also worth noting that in S. Gelman and Markman (1987), some three and four-year-olds' inferences involved categorizations (see note 8). Since the projections in these inferences were mediated by K-laws, it seems reasonable to assume that the categorizations were too, although there is no direct evidence for this assumption. If the assumption is correct, three and four-year-olds are in many cases making projections and categorizations with equal facility.

2.4 Other Explanations of the K-patterns

The hypothesis that children believe K-laws is the *only* hypothesis that can account for all aspects of the K-patterns. This becomes evident as one moves step by step from the explanation of Gelman's experiments on projection through to the explanation of Keil's transformation experiments:

> *Gelman's Projection Experiments:* Any theory structure that posits a strong link between category and behavior but does not postulate an *unbreakable* link between category and appearances can account for Gelman's results, since on any such theory, (a) the child can infer that all squirrels behave the same way, and (b) the child can learn that rabbits eat grass without thereby committing to the belief that all rabbit-shaped creatures eat grass. It is not necessary to posit a representation of a *causal* law linking category and behavior. For example, Gelman's results can be explained by the hypothesis that children consider certain behaviors to be necessary and sufficient for category membership. Then squirrels *must* eat bugs, or they would not be squirrels, but there is nothing causal about it.
>
> *Keil's Discovery Experiment:* The discovery experiment shows that the child considers the laws linking kind and observable properties to have exceptions. They do not, however, show that these laws are causal. The laws might be non-causal statistical generalizations.
>
> *Keil's Transformation Experiments:* With the transformation experiments (assuming the explanation from artificiality) it becomes apparent that information about the abnormality of causal processes inclines older children to refrain from making inferences based on the laws linking kind

and observable properties. This distinctly causal pattern of inference strongly suggests that the laws are represented as causal laws.

3. The Minimal Hypothesis

Everything the essentialist hypothesis explains without the help of additional hypotheses, it explains in virtue of the child's belief in K-laws, causal laws connecting kinds and their observable properties.[13] Thus a more conservative hypothesis suggests itself: that the child represents K-laws but does not have any beliefs concerning the metaphysical foundation of these laws. In particular, the child does not represent an essence as doing the causing. This is the minimal hypothesis.

The minimal hypothesis denies all three tenets of the pure essentialist hypothesis: (1) that children represent essences, (2) that children define kinds in terms of essences, and (3) that children think that essences are causally responsible for natural kind members' observable properties. This is not to say that children think that there are no essences; rather, they have no opinion about what it is that makes the causal laws true.[14] All children are committed to is the existence of the K-laws.

It is important to note that one can believe the K-laws without implicitly committing oneself to essences. There are many ways that the K-laws could be true, but essentialism false; some are described in section 1.4. Indeed, many philosophers and biologists have argued that this metaphysical possibility is actual, that is, that the K-laws are true (of species and chemical substances) but metaphysical essentialism is false. Mayr (1970) makes the argument for biology; Mellor (1977) for chemistry.

One further point of clarification about the minimal hypothesis: a representation of a causal law (including a K-law) is something over and above a representation of a regularity. It is not just that most or all tigers are ferocious, it is that something about being a tiger *causes* ferocity. Exactly what additional psychological properties distinguish a representation of a lawful causal connection from a representation of a mere regularity is too big a question for this paper to answer, but it is clear that there is such a distinction. To believe that it has rained every prime-numbered day in April for

[13] The exception is internal essentialism's use of the insides hypothesis, which it incorporates, but this is problematic for reasons explained in sections 2.2 and 4.2.

[14] More exactly such opinions appear relatively late, perhaps largely as a result of formal education.

the last ten years is one thing; to believe that those days' prime numbering *caused* it to rain is, in its cognitive significance, quite another.[15]

Because the minimal hypothesis holds that children represent K-laws, it explains children's K-patterned inferences—but so does the essentialist hypothesis. The next and final section of this paper argues that the minimal hypothesis is superior to all variants of the essentialist hypothesis.

4. Against Psychological Essentialism

4.1 Against Statistical Essentialism

I present my case against the three varieties of psychological essentialism separately. The case against statistical essentialism is that, wherever it departs from pure or internal essentialism—that is, wherever it holds that members of a category do not necessarily have the category's essential properties (departing from tenet 2), or that the links between essence and observable properties are non-causal (departing from tenet 3)—it suffers a loss of explanatory power. Thus both pure and internal essentialism are explanatorily superior to statistical essentialism. I discuss the explanation of K-patterned projections and categorizations in turn.

When accounting for K-patterned projections, statistical essentialism stumbles wherever it departs from tenet 2 of pure essentialism. As I argued in section 2.1, to account for children's conviction that even a rabbit-shaped squirrel eats bugs, one must posit a link between being a squirrel and eating bugs. For the child essentialist, this link goes by way of the essence: squirrels all have the same essence; this essence causes them to eat bugs; therefore, all squirrels—even kaibabs—eat bugs. But insofar as it departs from pure essentialism's tenet 2, statistical essentialism denies that children believe all squirrels have the squirrel essence, and so undermines the reason for thinking that children will infer that any given squirrel will eat bugs. The statistical essentialist can hold that in this sort of case, children think all squirrels *do* have the essence, but that is just to say that in this case pure (or internal) essentialism is correct—and so on, for every K-patterned projection.

[15] There is, of course, an august philosophical tradition, initiated by David Hume, which denies that this psychological distinction corresponds to any metaphysical distinction. But these philosophers have not usually denied that there *is* a psychological distinction. To believe a connection is causal is to be in a different psychological state from believing that it is regular but non-causal. This is all that the minimal hypothesis requires.

When accounting for K-patterned categorizations, statistical essentialism stumbles wherever it departs from tenet 3 of pure essentialism. This is because, as argued in sections 2.2 and 2.4, to explain the K-patterned categorizations concerning transformations of differing superficiality, one must posit that the links between essence and observable properties are causal. Again, the statistical essentialist could maintain that the particular kinds mentioned in Keil's studies (a diverse set of birds and mammals) *are* represented as causally connected to the relevant observable properties—but that is just to admit that, with respect to these kinds, pure essentialism is right.

4.2 Against Internal Essentialism

In its favor, internal essentialism incorporates an explanatory hypothesis, the insides hypothesis stated in section 2.2, that other forms of essentialism and the minimal hypothesis must postulate separately. But there are three serious objections to internal essentialism that far outweigh this advantage.

4.2.1 Children's Explicit Views on the Causal Role of Insides

According to internal essentialism, children explicitly represent essences (a) as properties of organisms' insides and (b) as the causes of organisms' appearances and behavior. Thus children explicitly represent organisms' insides as causes of their appearance and behavior, and—if internal essentialism is to explain the K-patterns—they do so from an early age.

S. Gelman and Kremer (1991) asked children why objects had their characteristic behaviors and appearances. For example, the children were asked why rabbits hop and have long ears. Regardless of their answer, they were then asked whether the insides of the object might be responsible for the behaviors and appearances. Only half of seven-year-olds' responses were affirmative. Yet in Gelman's projection experiments, a greater proportion of significantly younger children (two-thirds of four-year-olds) made K-patterned inferences (Gelman & Markman, 1986). This strongly suggests that young children's K-patterned projections are not a result of their believing that insides cause observable properties. Understanding of the causal role of insides seems to be slow and incremental, and so cannot play a central role in all inference involving K-laws.

4.2.2 Insides Versus Lineage in the Discovery Experiments

Recall from section 2.2 that the explanation of Keil's discovery experiments requires two hypotheses:

> *The lineage hypothesis:* Children believe that members of a kind invariably have offspring of the same kind, and
>
> *The insides hypothesis:* Children believe that members of a kind invariably have kind-specific insides.

Either hypothesis could explain children's responses to the discovery experiments. But in fact, they seem to play roughly equal roles in the explanation: interviews with children during the experiments show that the children appeal to lineage about as often as they appeal to insides to explain their K-patterned inferences. I assume that this is because the two hypotheses carry equal weight, on average, for the children. (It would be interesting to conduct the experiment twice more, giving only information about lineage, then only information about insides.)

Internal essentialism incorporates the insides hypothesis, but not the lineage hypothesis. It follows that the internal essentialist child gets the relationship between insides and kind for free, but must learn the relationship between lineage and kind. According to internal essentialism, then, children's grasp of the relationship between insides and kind ought to come earlier than their grasp of the relationship between lineage and kind. Why, then, do the children in Keil's experiment appeal to lineage as often as they appeal to insides?

I consider three responses on behalf of the internal essentialist. The first concerns timing: it may be that children learn about lineage at the same time that they become essentialists. This suggestion squanders the advantage that internal essentialism has over the other explanations of the K-patterns: if children can learn about lineage at the same time that they are becoming essentialists, surely they can learn about insides, too. Then there is no motivation to incorporate the insides hypothesis (or any other like it) in the essentialist hypothesis.

The second internal essentialist response is to attempt to incorporate the lineage hypothesis in internal essentialism, that is, to posit that the beliefs attributed by the lineage hypothesis are entailed by the child's conception of essence. It would have to be argued that it is a consequence of the child's conception of essence that organisms with a certain essence always have children with the same essence. But it appears that this view is not consistent with internal essentialism. For this "same lineage same kind" principle to be entailed by the child's conception of essence, an organism's lineage would somehow have to be a part of its essence. But the property of having a certain parent is not an internal property of an organism. Thus the claim that children represent lineage as part of an organism's essence *contradicts* the internal essentialist's claim that children represent an organism's essence as being inside it. You can have either the insides hypothesis or the lineage hypothesis as a part of the notion of essence, but never both.

A third internal essentialist reply to this objection, which might also be given to the previous objection (children's explicit views about the causal roles of insides), is that children are bad at reporting the beliefs that go into their reasoning about natural kinds and categories. Resorting to this *ad hoc* hypothesis again squanders internal essentialism's explanatory advantage. The reason for preferring internal essentialism to other varieties of essentialism and to the minimal hypothesis was that it made one less assumption than its competitors—it did not assume the insides hypothesis. But it turns out that internal essentialism must make a far more sweeping assumption instead, that where the data contradict its predictions, children have simply become confused, falsely reporting their own reasoning. This is not, of course, impossible, but it does not reflect well on internal essentialism that it must make such a self-serving assumption. Some further comments on this kind of reasoning are made in section 4.3.

4.2.3 The Timing of the Discovery and Transformation Results

My third objection to internal essentialism is that it is unable to explain the fact that children's performance in the transformation study is, at every stage, better than their performance in the discovery study. I begin by showing that internal essentialism predicts the reverse, that is, that children ought to do better in the discovery than in the transformation study.

According to internal essentialism, a child's giving the correct answer (by our lights) to the questions in Keil's discovery experiment depends on her coming to reason in accordance with two precepts (ignoring for the duration of this section the problem of the influence of information about lineage discussed above):

1. That members of a kind have kind-specific insides (the belief attributed by the insides hypothesis). According to internal essentialism, children

believe this because they believe that essences are internal properties of an organism's insides.

2. That the relation between category and appearances/behavior is defeasible (a consequence of believing the K-laws).

Internal essentialism predicts, then, that a child will give the correct answers to the discovery questions as soon as she comes to believe (1) and (2). (1) is an integral part of naive theories from the start, according to internal essentialism, so belief in (1) will precede belief in (2). It follows that a child will give correct answers to the discovery questions as soon as she reasons in accordance with (2), that is, as soon as she appreciates that categorizations based on the K-laws can be mistaken.

Internal essentialism predicts that a child will give correct answers in the transformation experiments only if she has acquired knowledge of (2) and of a theory of growth. Since correct answers in the discovery experiments depend only on the child having acquired knowledge of (2), internal essentialism predicts that a child will give correct answers to the discovery questions either at the same time as, or before, she gives correct answers to the transformation questions (depending on whether (2) or the theory of growth comes first).[16] Pure essentialism and the minimal hypothesis, on the other hand, make no such prediction. They hold that knowledge of the relationship between insides and kind is acquired independently of the K-laws. Thus correct answers to the discovery questions will come first if acquisition of (1) precedes an adequate understanding of growth, and correct answers to the transformation questions will come first if acquisition of (1) lags an adequate understanding of growth.[17]

Keil's data very clearly go against internal essentialism on this matter. The responses of children to both the discovery and the transformation experiments were ranked on a scale from 1 to 3. A score of 1 corresponded to an incorrect answer, that is, a categorization in accordance with appearances rather than "deep" properties. A score of 3 corresponded to a correct answer. A score of 2 corresponded to, in effect, "undecided" (not a common

[16] I am speaking loosely, of course, since neither the acquisition of (2) nor that of a theory of growth is an all or nothing affair.

[17] If the argument from superficiality gives the correct explanation of the transformation experiments then the same conclusions hold, but with "an adequate understanding of the effects of external changes on the essential property" substituted for "an adequate understanding of growth".

response, Keil says). The data for nine-year-olds is the most striking (but the effect is visible at every age). In the discovery experiment, the nine-year-olds' average score for biological kinds was about 2.35. In the transformation experiments the average score for biological kinds was about 2.85—slightly *higher* than the *adults'* score for the discovery study. That is, not only do children do well in the transformation experiments before they do well in the discovery experiments; by the time they are nine they do better in the transformation experiments than they are *ever* going to do in the discovery experiments.[18] Because of the robustness of the evidence,[19] I consider this to be by far the strongest argument against internal essentialism.

This argument may be put together with the first argument against internal essentialism: children do well in the transformation experiments before either (a) they attribute causal powers to insides or (b) they do nearly as well in the discovery experiments, which on any account require the belief that members of a kind have kind-specific insides. This strongly suggests that children's understanding of the kind-specificity of insides *lags* their skill in making K-patterned inferences. This fact is completely at odds with the claims of internal essentialism. It is to the advantage of the pure essentialist hypothesis and the minimal hypothesis, then, that they hold that knowledge about the relationship between insides and kind is not at the core of naive theories.

4.3 Against Pure Essentialism

Of the essentialist hypotheses, pure essentialism is best able to account for the K-patterns of inference. The minimal hypothesis is just as successful. Is there any reason to prefer one over the other?

There is an argument from parsimony for favoring the minimal hypothesis. The pure essentialist hypothesis asserts the existence of certain mental entities—namely, representations of essences—that do no explanatory work. Nothing in the pure essentialist's explanation of the K-patterns depends on the representation of essences; everything depends on the repre-

[18] All claims are, of course, to be understood with "on average" affixed. Many children grow up to do equally well in the discovery and transformation experiments.

[19] Keil conducted two sets of discovery experiments and two sets of transformation experiments. The average score on both sets of discovery experiments was about the same, as was the average score on both sets of transformation experiments. There were about 50 children in each study.

sentation of causal laws about kinds. There is no reason, then, to posit the existence of representations of essences. We should prefer the minimal hypothesis.

Better than a parsimony argument, however, would be some piece of empirical data that can be explained by one but not the other hypothesis. The two hypotheses agree on many things; the question on which they clearly differ is whether a representation of a kind's essence mediates projections and categorizations involving the kind. This difference does not make much difference to children, who, as I have argued, have next to no opinions about essences. Where it is likely to show itself is in the reasoning of adults who do have such opinions, particularly about chemical kinds. In this final section, then, I turn to a study that involves adults' reasoning about chemical kinds.

The study is an examination of the relation between water and H_2O conducted by Barbara Malt (1994). Malt compiled a list of liquids, some water and some non-water, that is, some of which are normally counted by humans as instances of water, some of which are not normally counted as water. (Malt herself made these judgments; see my comments below, especially note 24.) She gave her subjects this list and asked them to estimate the percentage of each liquid that is H_2O. There turned out to be only a very loose correlation between percentage of H_2O and membership of the category "water". In particular, some liquids considered non-water by humans—tea, grapefruit juice, and lemonade—were estimated to have a far higher H_2O content (around 90%) than some liquids considered water, such as swamp water, radiator water, and sewer water (all less than 70%).[20] Malt concludes that having a high percentage of H_2O cannot be the sole criterion for membership of the category "water" (see also Chomsky, 1995, 22–3).

This raises a serious problem for pure essentialism. At some point in their lives, Malt's subjects presumably learned that water is H_2O. On the usual essentialist story (Medin & Ortony, 1989), learning this fact has the following effect. First, people come to believe that H_2O is the essence of water. Second, they update their theories accordingly, fleshing out their representation of water's essence or replacing an essence placeholder with a representation of H_2O. The result will be a theory similar to that pictured in figure 4.

In the essentialist's fleshed out theory of water, the property of being

[20] These estimates are, of course, quite inaccurate, but that is another matter.

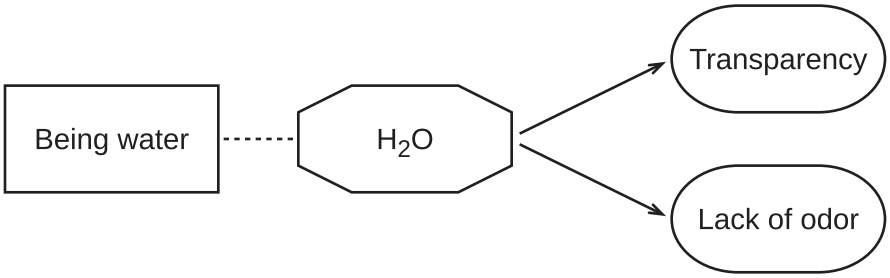

Figure 4: The theory of water updated to reflect the fact that water is H_2O

H_2O now mediates all inferences between the relevant observable properties and membership of the category "water". When a person categorizes swamp water as water, for example, according to essentialism her inference proceeds in two steps: (a) she first infers from the properties of the liquid that it is H_2O, then (b) she infers from its being H_2O that it is water.

What does it take to "be H_2O"? It is too much to demand that a substance must be 100% H_2O: humans count many liquids as water that are not that pure. "Being H_2O" must, in this context, mean being *mostly* H_2O. What, then, counts as being mostly H_2O? We know, first, that humans count swamp water as a kind of water, and second, that humans think swamp water is about 70% H_2O.[21] If follows that humans whose theories have the structure postulated by pure essentialism must consider being 70% H_2O adequate to be classified as water. If the threshold were higher, step (b) of their categorizing inference would not go through.

Pure essentialism attributes to humans the following inference: swamp water is (at least) 70% H_2O, so swamp water is water. Now, we have seen that humans think that tea, grapefruit juice, and lemonade are more than 70% H_2O. If humans were essentialists, and they were to infer consistently, they would conclude that these substances are kinds of water. But they do not. This suggests that essentialism is wrong.

The pure essentialist could deflect this problem if she could plausibly claim that the inference from being (at least) 70% H_2O to being water is an inductive inference, and is therefore defeasible. (The claim, then, is something like this: humans believe that substances that are 70% H_2O are usually water. But in the case of tea, blood, and so on, certain other information

[21] Of course, not all humans would agree on this figure, but Malt's study shows that enough agree on a figure close enough to this one to create a problem for pure essentialism.

about the substance defeats the inference.) In a recent paper, Susan Gelman seems to take this course of action in response to Malt's experiment (Gelman & Hirschfeld, 1999, 408). This is to abandon pure essentialism for a kind of statistical essentialism, the kind that denies essentialism's tenet 2 (and indeed, Gelman and Hirschfeld cite Medin as an inspiration). I showed in section 4.1, however, that an essentialism that abandons tenet 2 is unable to explain Gelman's own results concerning children's K-patterned projections. The pure essentialist, on the other hand, can explain the K-patterns, but is committed to tenet 2, that is, to the proposition that essences are represented as necessary and sufficient for category membership. In particular, pure essentialism is committed to the proposition that, if a person believes that H_2O is the essence of water, then she believes that water is *defined* in terms of H_2O (hence the dashed line between the two in figure 4). The person believes that the relation between H_2O content and being water has no exceptions. Thus the pure essentialist cannot explain why swamp water is counted as water when tea is not.

The minimal hypothesis has no such problem. According to the minimal hypothesis, when a person categorizes a sample of water, they infer directly from its observable properties to its kind. No essence mediates the inference, so the inferrer need make no decision about the relation between essence and kind—in the case at hand, no decision about how much H_2O water needs to have. Nor would any such decision, if it *were* made, affect everyday inferences about water (since these inferences are made independently of any information the thinker may have about the relationship between H_2O and water). Thus the minimal hypothesis explains how the average person can think so much about water, making constant categorizations and projections in everyday life, without being in any way practically affected by the eventual acquisition of the knowledge that water is H_2O. Knowledge about essences simply plays no role in our ordinary thinking about natural kinds.

I will consider five replies to this argument.[22] The first is that the subjects of Malt's experiment are confused in some way, and fail to draw the conclusions about water that follow from their own essentialist theories. That is, people somehow do not see that tea, chicken broth, grapefruit juice, and blood are, according to their own theories, kinds of water.

Pure essentialism is able to explain the K-patterns of inference by positing that children are essentialists, and then showing that a child or adult who

[22] Thanks to anonymous referees for suggesting several of these replies.

projects and categorizes in accordance with an essentialist theory will produce K-patterned inferences. The assumption made by pure essentialism (and by every other theory considered in this paper) is that children actually make the inferences that follow from their essentialist theories.

To excuse pure essentialism on the grounds that Malt's results are a product of confusion, then, is to endorse the following strategy:

1. When pure essentialism correctly predicts the inferences people make, these inferences are to be explained by the fact that people are following through the consequences of their essentialist theories.

2. When pure essentialism incorrectly predicts the inferences people make, these inferences are to be explained by the fact that people become confused and are unable to follow through the consequences of their essentialist theories.

Either this is a textbook example of unacceptably *ad hoc* reasoning, or there is some significant difference between the inferences that are made in (1) and those that are botched in (2). What could the difference be? An obvious answer is that the inferences in (2) are more difficult to make.

Unfortunately for the pure essentialist, it is the inferences in (1) that are more difficult to make; indeed, the inferences in (2) turn out to be steps in the more complex inferences in (1). The inferences in (2) are simply inferences from the fact that an organism or substance has a certain essence (e.g., being more than 70% H_2O) to the fact that it belongs to the relevant kind, and vice versa (from kind membership to essence possession). Call these kind/essence inferences. Kind/essence inferences are not only simple, they form an indispensable part of the more complicated chains of reasoning that the pure essentialist uses to explain the K-patterns. For example, in Gelman's projection experiments, preschoolers are supposed to infer (a) that a kaibab has squirrel essence from the fact that it is a squirrel, and (b) that it therefore eats bugs. But the inference in (a) is a kind/essence inference. If the kind/essence inferences in (2) are botched then the inferences in (1), which contain kind/essence inferences, ought also to be botched. At least, this is true unless there is something especially difficult about the kind/essence inferences in Malt's study. (Some possible difficulties will emerge in the next two objections.)

The second essentialist reply is made by Barbara Abbott (1997). Abbott proposes that tea, blood, and so on *are* kinds of water, but that they are not normally *called* water. That is, although humans classify, say, tea as a kind of water, this fact is obscured by the linguistic fact that humans invariably refer to tea as "tea".[23] This objection takes advantage of a weakness in Malt's experiment: Malt did not ask her subjects what they counted as water, but simply assumed that anything with "water" in its name is counted as water, and anything else is counted as non-water (Malt, 1994, 45).[24]

Abbott's claim that tea is categorized as water is open to an obvious objection: when asked, people say that tea is not water. (This is easily confirmed informally.) Could it be that, when they say this, they mean only that tea is not *called* water? This does not fit the facts: people say that snow is water, despite the fact that snow is never called water (outside of a chemistry class). Or to take an example from the biological realm, people asked whether a Bombay duck is a duck, and who know that it is a kind of fish, will say no. There is, I conclude, very good reason to think that these types of questions elicit information about categories, not names.

At this point the essentialist might suggest that, when people report that they do not count tea as a kind of water, they are mistaken. I have already said something above about the suggestion that people make this sort of error in kind/essence inferences. What is new here is a particular story about what causes the confusion: it is brought on by exposure to too much language. Because tea is always referred to as "tea", people tend to forget that they also count it as a kind of water.

There are at least two reasons to think that this line of thought is flawed. First, a simple confusion like this ought to be easy to overcome once it is pointed out that a thing's name is not always the last word about its nature. But even sophisticated adults aware of the issues resist the suggestion that tea, blood, and grapefruit juice are kinds of water.

Second, there are things that everyone counts as kinds of water, but that are usually referred to using other terms: rain, ice, and snow are good examples. If it were names that were causing the problem, it should be just

[23] Abbott is writing primarily about language, and the difference she emphasizes is not between categorization and naming but between the semantics and pragmatics of the meanings of names. However she assumes that semantics and categorization go together; see bottom of p. 312.

[24] For example, Malt counts tears as non-water. My own informal investigations show that many people are at least somewhat inclined to classify tears as water, suggesting that Malt's criterion is unreliable. The argument I make in the main text does not depend on any of these borderline cases.

as difficult to keep track of the fact that snow is water as it is to keep track of the fact that tea is water. Yet the same people who say that snow is water say that tea is not. Perhaps Abbott could reply that the difference results from the fact that we sometimes (although only in very special contexts) refer to snow as water but we *never* refer to tea as water. This is a dangerous fact for the essentialist to bring to our attention, however: surely what best explains the fact that we never refer to tea as water is that we do not think it is water.

A third essentialist defense holds that people make kind/essence errors in Malt's study because they are reluctant to classify a substance as belonging to two different basic level kinds.[25] On this view, (a) according to our theories of water and other liquids, a substance can be both tea and water, but (b) if a person classifies a liquid as tea, it is very difficult for them to then classify it also as water, thus people give the appearance of thinking that tea is not water.

This is very similar to the suggestion just considered, that people become confused by language, and is unconvincing for the same kinds of reasons. People have very little trouble with the idea that the stuff we classify as snow is also water, so why do even the most sophisticated consistently fail to classify tea as water?

A fourth defense is that "water" has a vague essence.[26] What I have in mind is the claim that the essence of water is (say) "85% ± 15% H_2O", with the exact percentage varying according to context. The problem with this suggestion is that our decisions about what counts as water do not seem to be at all sensitive to context. Tea is not water in any context. Swamp water is water in any context. Worse, in the context of this sentence, tea is not water yet swamp water is water. But this is impossible: whatever level of H_2O the context of the previous sentence sets for membership of the category of water, it cannot be that swamp water has it but tea does not.

Perhaps something other than context is responsible for determining the level of H_2O necessary for waterhood on a case by case basis. It is up to the pure essentialist to say what that something is, and to show that it does not interfere with the explanation of the K-patterns.[27]

[25] On the notion of the "basic" level of categorization and its role in human psychology, see Rosch, Mervis, Gray, Johnson, and Boyes-Braem (1976).

[26] Abbott (1997) brings up this possibility, but she uses it to explain our (allegedly) shifting criteria for applying the term "water", rather than claiming that the *semantics* of "water" is vague.

[27] See also LaPorte (1998), writing in response to Abbott (1997).

Finally, the fifth defense of pure essentialism is to suggest that water is represented as having an essence other than "at least 70% (or some other proportion) H_2O". This goes against the usual essentialist line (see especially Atran (1990)) that as children mature into adults, their theories become more and more like the currently accepted scientific theories, since if chemistry attributes any essence to water, it is H_2O. But suppose that, for the sake of saving the hypothesis, the pure essentialist retracts this claim.

Then pure essentialism owes us some suggestion as to what the represented essence of water is, such that it includes swamp water, the stuff in a radiator, and sewer water, but not tea, blood, or grapefruit juice. There is a very strong constraint: the K-patterned projections and especially the K-patterned categorizations rule out appearances and behavior as possible constituents of this essence. Thus a description of the essence must not mention the appearance or the chemical behavior of water. But what do people know of water, apart from its look, its behavior, and the fact that "water is H_2O"? Nothing. If there is something that is widely believed to be the essence of water, and that has nothing to do with looking like water or behaving like water, it must be something to do with H_2O. But what could that something be but the proportion of H_2O? There may be something that I have not thought of; it is up to the pure essentialist to say what it is. I might add: since we all represent water as having this essence, why doesn't it spring to mind when summoned?

Before concluding, I will make one more point in favor of the minimal hypothesis. The minimal hypothesis suggests a partial explanation of the fact that some things with high H_2O content are not counted as water. It will be noted that the non-waters—such as tea, juice, blood, and disinfectant—have special chemical (or biochemical) properties in addition to those possessed by water. It is our knowledge of these additional properties, I tentatively suggest, that prevents us from categorizing the fluids as water. For example, we know the following causal law: disinfectant kills germs. But we also know that this law is not true of water. So we infer from the germ-killing ability of disinfectant that it is not a kind of water.[28,29]

[28] The full story must be considerably more complicated than this. For example, chlorinated tap water is a kind of water, but chlorinated water also kills germs (if much less effectively). Or consider: there are causal laws about sewer water (it smells bad) that are not true of water in general.

[29] Essentialists can offer the same explanation. But although we may infer, defeasibly, from the special properties of coffee that it is not water, if we also believe (a) that coffee is

One final question: if essences are not represented in our naive theories, where does full blown metaphysical essentialism in, say, biology come from? To believe the biological K-laws is to be committed to the proposition that there is something about (for example) being a tiger that causes a tiger's observable biological properties. Other animals have a "something" that causes quite different properties. The philosopher of nature cannot help but ask: what is this something? And perhaps cannot help but answer that it is for each biological kind a hidden property uniquely characteristic of that kind. Thus essentialism is born.

In this way, from psychology, emerges philosophy. The journey back has taken quite some time.

90% H_2O and (b) that anything that is at least 70% H_2O is water, we must infer indefeasibly that coffee is water. Since (a) is apparently a common belief, and (b) is necessary if the essentialist is to explain the categorization of swamp water, the essentialist hypothesis predicts, contrary to the facts, that the indefeasible inference will be made—that coffee will be counted as a kind of water.

References

Abbott, B. (1997). A note on the nature of "water". *Mind*, *106*, 311–9.

Atran, S. (1990). *Cognitive foundations of natural history*. Cambridge: Cambridge University Press.

Atran, S. (1995). Causal constraints on categories and categorical constraints on biological reasoning across cultures. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition* (pp. 205–233). Oxford: Oxford University Press.

Chomsky, N. (1995). Language and nature. *Mind*, *104*, 1–61.

Gelman, R. (1990). First principles organize attention to and learning about relevant data. *Cognitive Science*, *14*, 79–106.

Gelman, S. A. (1988). The development of induction within natural kind and artifact categories. *Cognitive Psychology*, *20*, 65–95.

Gelman, S. A., & Coley, J. (1990). The importance of knowing that a dodo is a bird: Categories and inferences in 2-year-old children. *Developmental Psychology*, *26*, 796–804.

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

Gelman, S. A., Collman, P., & Maccoby, E. (1986). Inferring properties from categories versus inferring categories from properties: The case of gender. *Child Development*, *57*, 396–404.

Gelman, S. A., & Hirschfeld, L. A. (1999). How biological is essentialism? In D. L. Medin & S. Atran (Eds.), *Folkbiology* (pp. 403–446). Cambridge, MA: MIT Press.

Gelman, S. A., & Kremer, K. E. (1991). Understanding natural cause: Children's explanations of how objects and their properties originate. *Child Development*, *62*, 396–414.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, *23*, 183–209.

Gelman, S. A., & Markman, E. M. (1987). Young children's inductions from natural kinds: The role of categories and appearances. *Child Development*, *58*, 1532–1541.

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

Hirschfeld, L. (1996). *Race in the making: Cognition, culture, and the child's construction of human kinds.* Cambridge, MA: MIT Press.

Keil, F. (1989). *Concepts, kinds and conceptual development.* Cambridge, MA: MIT Press.

LaPorte, J. (1998). Living water. *Mind*, *107*, 451–5.

Locke, J. (1975). *An essay concerning human understanding.* Oxford: Oxford University Press.

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

Mayr, E. (1970). *Populations, species, and evolution.* Cambridge, MA: Harvard University Press.

Medin, D. (1989). Concepts and conceptual structure. *American Psychologist*, *44*, 1469–1481.

Medin, D., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–195). Cambridge: Cambridge University Press.

Medin, D. L., & Atran, S. (Eds.). (1999). *Folkbiology.* Cambridge, MA: MIT Press.

Mellor, D. H. (1977). Natural kinds. *British Journal for the Philosophy of Science*, *28*, 299–312.

Putnam, H. (1970). Is semantics possible? In H. Kiefer & M. Munitz (Eds.), *Contemporary philosophic thought: The international philosophy year conferences at Brockport.* Albany, NY: SUNY Press.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21–59). Cambridge: Cambridge University Press.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.

Rosengren, K. S., Gelman, S. A., Kalish, C. W., & McCormick, M. (1991). As time goes by: Children's early understanding of growth in animals. *Child Development*, *62*, 1302–1320.

Rothbart, M., & Taylor, M. (1994). Category labels and social reality: Do we view social categories as natural kinds? In K. Fiedler & J. Semin (Eds.), *Language and social cognition* (pp. 11–36). Thousand Oaks, CA: Sage Publications.

Springer, K. (1992). Children's beliefs about the implications of kinship. *Child Development*, *63*, 950–959.

Springer, K., & Keil, F. (1989). On the development of biologically specific beliefs: The case of inheritance. *Child Development*, *60*, 637–648.

Vosniadou, S., & Ortony, A. (Eds.). (1989). *Similarity and analogical reasoning.* Cambridge: Cambridge University Press.