PHILOSOPHICAL PERSPECTIVES

Philosophical Perspectives, 26, Philosophy of Mind, 2012

CONCEPTS, ANALYSIS, GENERICS AND THE CANBERRA PLAN[1]

Mark Johnston
Princeton University

Sarah-Jane Leslie[2]
Princeton University

> My objection to meanings in the theory of meaning is not that they are abstract or that their identity conditions are obscure, but that they have no demonstrated use.[3]
>
> —Donald Davidson "Truth and Meaning"

From time to time it is said that defenders of conceptual analysis would do well to peruse the best empirically supported psychological theories of concepts, and then tailor their notions of conceptual analysis to those theories of what concepts are.[4] As against this, there is an observation — traceable at least as far back to Gottlob Frege's attack on psychologism in "The Thought" — that might well discourage philosophers from spending a week or two with the empirical psychological literature. The psychological literature is fundamentally concerned with mental representations, with the mental processes of using these in classification, characterization and inference, and with the sub-personal bases of these processes. The problem is that for many philosophers, concepts could not be mental items. (Jerry Fodor is a notable exception, we discuss him below.)

We would like to set out this difference of focus in some detail and then propose a sort of translation manual, or at least a crucial translational hint, one which helps in moving between philosophical and psychological treatments of concepts. Then we will consider just how, given the translation, the relevant psychology, particularly including recent work on the generic character of many of our 'platitudes' or early developing central beliefs (Gelman, 2010; Hollander, Gelman, & Star, 2002; Leslie, 2007, 2008, in press a; Leslie & Gelman, 2012; Leslie, Khemlani, & Glucksberg, 2011; Mannheim, Gelman, Escalante, Huayhua, & Puma, 2011; Tardif, Gelman, Fu, & Zhu, 2011), bears upon

philosophical theories of concepts, and especially on the new style of conceptual analysis that now goes under the heading of "the Canberra plan".

A Philosophical Theory of Concepts

What follows in this section is a short account of the general form of a substantial philosophical theory of concepts. We do not endorse it. We simply take it to be a widely held view, at least in its significant sub-parts; we also take it to be a view that animates the common practice of conceptual analysis within philosophy, along with the so-called conceptual theory of the a priori — the idea that the source of the a priori lies in the application conditions of concepts.

On the substantial theory, *concepts are abstract objects individuated by their conditions of application to entities, and it is possession or grasp of these concepts which guides the use of terms*. Thinkers who grasp or possess a given concept C will often come to associate it with a term or representation "T" and then use "T" to classify or characterize entities in accord with the specific conditions of application of the concept C. For such thinkers "T" will then express the concept C. Other thinkers, who do not even implicitly know the application conditions associated with C might nonetheless defer to those who use "T" to express C, and intend to do likewise even though they themselves do not use "T" to classify or characterize entities in accord with the specific conditions of application of C. They can thus make conceptual mistakes, such as supposing that arthritis is a more general infirmity than inflammation of the joints. They can then be said to have an imperfect or incomplete grasp of the concept expressed by "arthritis".

Thus far, concepts are abstract objects encoding conditions of application of words, phrases and the like; in effect, they are rules of reference determination across possible situations, rules that we grasp and employ as guides in our use of terms. They are what give spoken and written words 'life'; that is, it is by being associated with this or that concept that a word or phrase comes to have conditions of application. They would also explain how it is that words and phrases — like "si" — that appear in distinct languages can have double lives. The speakers of the distinct languages associate different concepts with the same words. Likewise, within a language, the association of different concepts with the same words and phrases is what accounts for ambiguous words and phrases such as "bank" and "the bank he built up".

The postulation of concepts, and the hypothesis that it is the grasp of these concepts — if not by us then by the experts and lexicographers to whom we defer — that properly guides our use of words, makes for *concept publicity*; that is, the possibility of two speakers with very different beliefs about a subject matter sharing the same concept of that subject matter. Someone who knows that arthritis is a disease of the joints can genuinely disagree with someone who believes it can spread to the bones. The second is in error about the application conditions of a concept that both speakers share. As a result

he can be meaningfully corrected by the experts; for example they are not changing the subject on him when they tell him that one cannot have arthritis in the thigh.

Since words can combine compositionally to produce phrases of any arbitrary length, the life-givers that are concepts *had also better combine compositionally* to give life to phrases of arbitrary length (see, e.g., Fodor, 1998, Fodor & Lepore, 2002). So the concept female lion had better be some computable construction out of the concept female and the concept lion. Very likely, it had better amount to an intersection of these concepts; that is, a concept whose conditions of application are given by the conjunction of the application conditions of the constituent concepts.

A natural extension of this whole way of thinking involves associating a certain sort of concept with a whole sentence. The concept associated with a whole sentence is a computable construction out of the concepts associated with the sentence's constituent words and phrases. The special name for such a concept is "a thought", where a thought is not a sentential or representational vehicle but rather what is expressed by such vehicles. On this view, it is precisely because sub-sentential concepts have application conditions that sentences express thoughts, and so can be true or false. The special name for the application conditions of these structured concepts, the ones that are expressed by whole sentences, is "truth conditions".

On the substantial theory, if there are concepts, and if our use of terms is explained by our grasp of the associated concepts, then *there will be certain immediate inferences that will look like good candidates to be underwritten simply by our grasp of the relevant concepts*, since those inferences seem utterly reliable and do not seem to depend on any intermediate empirical premise. So it is with the inference from

Bonnie is a mare

to

Bonnie is a horse.

Furthermore, there is a route to recognizing the truth of the corresponding conditionals of such inferences by appealing simply to our grasp of the relevant concepts. So it is said that it is a priori that

If x is a mare then x is a horse

where this means that there is a route to recognizing its truth that does not depend for its justifying force on any empirical information about how the world works.

This also explains why the conditional could not be false. The conditional gives no hostages to fortune; it involves no bet on any contingent way the world

is. Accordingly, the conditional has the defining feature of a so-called conceptual truth: it is not only a priori but necessary.

Finally, if our use of terms is explained by our grasp of the associated concepts, and if concepts have necessary and sufficient conditions of application, then we might well hope to make our implicit knowledge of the application conditions of our terms *explicit*, by reflecting on what seems to be guiding us when we apply them in the vast range of imaginary cases that we can conjure up in the philosophical armchair. We can then form various hypotheses about just which sets of application conditions lie behind which terms, and we can test these hypotheses further by means of the method of consulting real and imaginary cases. The aim is, for each significant philosophical concept that is not simple — that is, for each concept that is compounded out of other concepts — to arrive at an a priori and necessary biconditional of the form

x is T if and only if x is . . .

where "T" is a term expressing the concept in question.

Not just any such biconditional will do. It has to be non-trivial. And it has to be rich enough to enable the derivation of all the a priori and necessary truths that arise from the concept associated with "T". *Finding a rich enough a priori and necessary biconditional is not easy; but if we succeeded we would have an 'analysis' of the relevant concept.* Generalizing, if we were to provide such analyses for all of the concepts of philosophical importance — *person*, *cause*, *essence*, *knowledge*, *mental state*, *matter, good* and so on and so forth — we would have articulated the bases for all of our distinctively philosophical a priori knowledge.

Something like this substantial picture has a good claim to be at the motivating core of what was once called 'analytical philosophy' (from Gottlob Frege, Kurt Gödel, A.J. Ayer, H.P. Grice and Roderick Chisholm through to George Bealer and Frank Jackson). Arguably, with the simple addition that the abstract objects that encode the application conditions of terms are *the rules we grasp* and which then determine our sense of the rightness of our applications of the term in particular cases, the picture is the real target of the skepticism about meaning that Saul Kripke (1982) saw in certain passages in Wittgenstein. Whether or not one accepts the picture, it goes some way toward illuminating just why a philosophical theory of concepts might take little notice of mental entities or mental processes.

Of course, the picture raises many questions: What are the specific application conditions of this or that particular concept? What it is to grasp a concept? What is it for a concept to be associated with a word or phrase? Are concepts really 'life-givers' or is talk of them simply a way of encoding, as it were after the fact, the details concerning the extension of this or that word- or phrase-in-use? Are concepts then idle wheels? Is a priori knowledge

really knowledge deriving from knowledge of the application conditions of our concepts? If so, is there any that goes beyond knowledge of merely stipulative definitions? Is the method of imagining a case and calling on our intuitions about whether a given concept applies in such a case a good way of making explicit our implicit grasp of the application conditions of the concept? Are some concepts, for example the familiar concept *plus,* more natural than others, in the sense of being privileged attractors of the terms that we use, attractors which we can only avoid by explicitly intending to use our terms in certain specific ways? Is this why our term "+" expresses the concept *plus* rather than the concept of some finite function that is like plus for a huge finite stretch but gives no answer just where our capacity to add actually gives out?

It may well turn out that these questions have disappointing answers. (For some of best arguments that this is so, see Willard van Orman Quine (1951), Hilary Putnam (1962, 1975), Gilbert Harman (1974, 1975), Jerry Fodor (1998), Timothy Williamson (2003) and Michael Devitt (2005).) Concepts may not be guides in any psychologically real sense. There may be no non-trivial biconditionals that neatly capture the application conditions of our concepts. There may be no a priori knowledge, period, or, at least, almost none beyond that delivered by logic and mathematics. Furthermore, the now popular idea of certain concepts, or the properties they demarcate, being privileged attractors may just be hopeful mysticism. At the end of the day, the frame provided by philosophical talk of concepts, analysis, and the distinctively philosophical a priori may have nothing very interesting left in it. (Obviously that pessimistic view also has its distinguished opponents; among them David Lewis (1970, 1994, 1997), Christopher Peacocke (1992) George Bealer (1987), Frank Jackson (1994, 1998), David Chalmers (Chalmers & Jackson, 2001) and Paul Boghossian (2003).)

What has so far been missing from the debate over these central philosophical issues is any systematic account of just how the psychological literature might bear upon them. By way of beginning to fill that gap, we will revisit these issues at the end of the paper, once the interrelations between philosophical and psychological theories of concepts have been set out in appropriate detail. Our ultimate view is that the psychology does resolve the philosophical standoff in a fairly convincing way.

In any case, it should be clear that on what we have been calling the substantial theory, concepts are abstracta that encode conditions of application; they are not mental processes or mental representations. One obvious route to the conclusion that concepts are not mental is to note that on the substantial view of concepts, the content of what is thought is a kind of concept, and only someone in the grip of the vehicle/content confusion would think that what is thought is a mental representation. Mental representations might be vehicles that *express* thoughts, but a mental representation is not the *content* of what is thought. Representations have characteristic effects, but no one is in danger of coming under the causal influence of a thought content, as opposed to a

representation of it, or a proponent of it. So the concepts that are the variety of things that are thought are none of them mental representations, and neither are their constituent concepts. Furthermore, as Christopher Peacocke (1992) emphasizes, once the substantial theory of concepts is up and running, it seems obviously coherent to suppose that many or most concepts are beyond our grasp, so that we will never bear any interesting mental relation to them.

However, psychologists who theorize about concepts are directly concerned with putatively mental entities, namely mental representations (including mental images and the like) that play a central role in our activities of classifying and characterizing and making immediate inferences about objects. A psychological account of our concepts in this sense is an account of just how it is that we, or our cognitive systems, systematically exploit certain sorts of general information to use the relevant mental representations to classify and categorize objects, and make inferences about them.

In fact, when psychologists do focus on mental representations they show little or no interest in the rules which determine their possible world extensions. What then is the relation of the psychological literature to the philosopher's interest in concepts? Is it simply a confusion of tongues to suppose that psychologists and philosophers are concerned with the same subject matter?

We believe that once the import of the psychology is made clear, a different philosophical view of concepts, one which attributes less knowledge to concept-users as such, becomes much more plausible. On this 'concepts as terms-in-use' view, a concept is a term, and a term is just an interpreted string in a language; a string which has, thanks to its conventional use, an extension, i.e. a range of items to which the term applies. Accordingly, it is only use, perhaps only conventionally governed use, which gives terms the 'life' that arbitrary strings lack. There need be no more to specifying a term than characterizing its syntactic form and giving a rule that determines its extension. There need be nothing *behind* terms — at least nothing like concepts understood as meanings or senses or extension-determining rules — that speakers *grasp* and which then guides their use of the relevant terms. Mutual correction while participating in a convention for using a term selects for a variety of effective heuristics or criteria that then guide individuals in their use of that term. To 'know how to use a term' is just to have some such effective criterion, and this will often fall far short of even a tacit understanding of the conditions of application of the term. Hence, on the concepts as terms-in-use view, there is no reason to suppose that speakers can render explicit the conditions of application of their terms simply by reflecting on how they would use them in a variety of circumstances. For such actual and counterfactual usages reflect only the speakers' empirical criteria, which may fall short of any grasp of application conditions. Thus 'armchair' philosophical analysis, as traditionally conceived, can at best produce fragmentary and inconclusive results. It can be no more than the articulation of the criteria we presently use in deciding whether to apply a term.

Problems with Psychological Theories of Concepts?

By starting with the philosopher's substantial notion of a concept, and adopting a translation of the following sort

> "Concept" in the psychological literature means pretty much what philosophers who have accepted significant parts of the substantive theory have meant by it

one can quickly make mincemeat out of the psychological literature. Jerry Fodor has already done much of this work for us in his *Concepts* (1998), where he points out that so construed, most, if not all, psychological theories of concepts fail the straightforward requirements of compositionality and publicity. To put it in the broadest possible terms, psychologists seem most interested in discovering our criteria or ways of telling when a given individual has a property or falls within a specified kind, and our criteria or ways of telling which inferences about members of a kind are the ones to make. But it will not be in general true that our ways of telling that something falls under a complex concept *F&G* is a joint application of our ways of telling whether something falls under the concept *F* and our ways of telling whether something falls under the concept *G*. Moreover, there is massive individual variation in our ways of telling whether something falls under the concept *F*; we may only be able to look it up in a book, you may have written the book and conducted the complex experiments needed to determine that it is F, and yet — so the substantial theory has it — all of us can share the same concept *F*.

Someone more enamored than Fodor of the conceptual theory of the a priori – say the Christopher Peacocke of A *Study of Concepts* (1992) and more recently "The A Priori" (2004) – might continue in the same general vein, by noting that, if we simply focus on our ways of telling which inferences about members of a kind are the ones to make, we will fail to distinguish between those inferences underwritten by our grasp or possession of the concept of the kind in question, and those inferences that appeal to collateral knowledge about members of the kind. Yet psychologists studying concepts happen to show no interest in this distinction. Instead, they focus on what generalizations and inferences are cognitively fundamental, in the sense of being (i) early developing and (ii) central to our a posteriori conceptions of the kinds in question.

We believe that the perception that these are — in any sense — defects in the psychological literature derives from the naive translation principle, and not from the psychological theories themselves. We will briefly review the relevant literature, emphasizing as we go just how recent discoveries concerning generics bear on that literature, then propose a different translation scheme, and finally examine the ways in which the psychological theories might put pressure on philosophical theories of concepts.

Some Psychological Theories of Concepts

The focus and the questions that drive psychological theorizing differ significantly from those which organize philosophical discussions of concepts. This is partly obscured by the fact that almost all reviews of the psychological literature on concepts begin with the so-called 'classical view', which is easy to misconstrue as simply a version of the substantial philosophical theory of concepts. On the classical view, when applying most lexical concepts (a lexical concept is a concept that is expressed by a single word) subjects actually exploit represented necessary and sufficient conditions for falling under the concept. The standard illustration of the classical view is the concept *bachelor*, which, it is claimed, is composed of the concepts *unmarried* and *man*, such that anything is a bachelor just in case it is an unmarried man. Of course, not all concepts can be decomposable in this way; there must be some or other stock of primitive concepts, out of which all other concepts are ultimately composed. Nevertheless, concept learning, on this view, involves combining such primitive concepts to form complex ones, and classifying an item under a decomposable concept really is supposed to be a matter of checking whether the item satisfies the necessary and sufficient conditions specified by the decomposition. Thus, someone who possesses the concept *bachelor* will classify items as bachelors *based on* their gender and marital status, in such a way that she will classify something as a bachelor *just in case* it is a man and unmarried. Mutatis mutandis for concepts such as *lion*, *dog,* and *table.*

Notice that, despite the initial appearances, the classical view of the concept bachelor is not just the (false but) banal remark that someone is a bachelor if and only if he is an unmarried man; it is instead the rather heady empirical thesis that we actually use the concepts *unmarried* and *man* in deciding whether to count something as a bachelor. It is distinctively a thesis about the criteria we actually use, not, or not just, a thesis about the application conditions of our concepts.

Since the 1970s, the classical view, understood as a thesis about the criteria we actually use, has been quite roundly rejected. Much of the reason for its rejection has to do with the discovery by Eleanor Rosch and her colleagues of so-called *typicality effects* (e.g., Rosch, 1973, 1978; Rosch & Mervis, 1975). For many categories, some members of a category are perceived as being more typical examples of the category than others, and it turns out that how typical a category member is actually predicts a very wide range of experimental results. For example, people are quicker to categorize typical members, and are more confident and consistent in their categorization of typical members. When learning a novel concept, people learn to categorize the typical members first, and they learn the concept faster when presented with typical members in the learning phase. There are also myriad effects of typicality on language learning and use, on reasoning, and so on so forth. (For some very helpful reviews, see Laurence & Margolis, 1999; Murphy, 2002; Smith and Medin, 1981).

Positing that people represent and exploit necessary and sufficient conditions does not explain typicality effects. Knowing that something is a bachelor just in case it is unmarried and a man does not give any information about what makes for the typical versus the atypical bachelors; for they are all alike in respect of being unmarried men. The pope is an unmarried man, but quite an untypical bachelor. Something else has to be posited to explain typicality effects; but once we have posited this something else there may be no empirical reason to also posit the representation of necessary and sufficient conditions. That is, suppose we posit some sort of mental representation that explains the powerful effect of typicality on people's categorization practice, and suppose we satisfy ourselves that this posited representation explains the experimental findings concerning people's dispositions to categorize items and generalize and reason about them. Then we will have accounted for all the data without positing operations on subjects' representation of necessary and sufficient conditions. From the perspective of explaining and predicting the target empirical results, knowledge of necessary and sufficient conditions then looks like an idle wheel — there are no results whose explanation demand positing the representation of necessary and sufficient conditions. The classical view has fallen into disrepute because many investigators believe that something like this has turned out to be true.

There may be some confusion concerning the rejection of the classical view: the evidence does not, it is sometimes said, *prove* that for most of our lexical concepts subjects do not represent and exploit necessary and sufficient conditions. That is correct; the findings of Rosch and her colleagues do not *prove* that there cannot be necessary and sufficient conditions lurking somewhere in our mental representations. However this form of resistance misses the real thrust of the psychological data; the issue is not that certain empirical results prove the absence of represented necessary and sufficient conditions, but rather that there is simply *no evidence for* their representation in the case of most lexical concepts.[5] Consider, for example, an experiment conducted by Jerry Fodor and his collaborators, in which they asked whether one could find any differences in processing time that would indicate that one concept is composed in part of another. If, e.g., the concept *murder* is composed in part by the concept *kill* (as has been claimed), then it should take longer to process *murder* than *kill*, since processing the former involves processing the latter as a proper part. However, this prediction of the classical view is not borne out (Fodor, Garrett, Walker, & Parkes, 1980). Again, no evidence in favor of the classical view emerged from that experiment. It is for this sort of reason that psychologists have moved beyond the classical view of concepts.

Driven by the need to explain typicality effects, many psychologists have embraced the prototype theory of concepts, where prototypes are statistical functions over properties, which assign weights to features based on how likely a category member is to have that feature, or conversely, based on how likely something with that feature is to be a category member. There are a number

of different proposals that fall under the heading of the prototype theory (see Murphy, 2002, for an extensive review), but they generally appeal to features that are in some way statistically related to category membership. For example, the prototype for *dog* might include features such as *barks*, *has four legs*, *has a tail*, *wears a collar* and so on. These features are not candidates to figure in universal necessary and sufficient conditions since not all dogs have these features; an unfortunate creature can still be a dog even if it has three legs, no tail, no collar and no voice. However, the basic idea behind prototype theory as applied to dogs is that if one is confronted with an animal and wishes to determine whether or not it is a dog, one will use this animal's features, or lack thereof, in a complex sub-personal calculation based on the weights of the various features in the prototype of *dog*. (The details of this calculation differ a great deal depending on the particular version of the theory.)

The weight that a feature receives in the prototype is generally taken to be determined by two sorts of statistical facts, namely the *prevalence* of the property among *dogs* (so *barks, has a tail*, etc would receive high weights since most dogs have these features), and/or the *cue validity* of the property; that is, how likely it is that something with that feature is a dog. Thus even though, perhaps, most dogs don't wear collars, the probability of something being a dog if it wears a collar is high, so *wears a collar* might receive a significant weight in the prototype. The more highly weighted features an individual has, the more typical an exemplar of the kind it will be; prototype theory thus places typicality effects first-and-foremost among the data it aims to explain.

Of course, one could try to express the prototype view in terms of some set of necessary and sufficient conditions; for example

> x is a bachelor if and only if his lifestyle and behavior resembles an appropriate paradigm, e.g. the one presented by Sean Connery in the early James Bond movies.

Clearly, when expressed this way, the prototype theory for *bachelor* fails miserably. This is just one sign that prototype theory is just not intended as an account of the application conditions — as, in effect, an analysis — of the term "bachelor". Another sign is that, as this very dated example suggests, the paradigms can change over time without thereby producing any change in the application conditions of the term.

Prototype theory has many adherents, and many well-motivated critics. While we may often rely on statistically weighted features in categorization, particularly in rapid, perceptually-based categorization, it seems that this cannot be the whole story. Imagine, for example, that you are presented with a raccoon. A perverse scientist then comes along and alters the creature, dying its fur so that it takes on the markings that are typical of a skunk, and even goes so far as to implant a sac of smelly liquid that the creature can use to spray smells when it is under stress. How would you categorize this creature? It now has all the typical features of a skunk, yet overwhelmingly, from elementary school on up,

people say this is still a raccoon (Keil, 1989). This finding is hard for standard prototype theories to accommodate. Furthermore, it seems increasingly clear that typicality ratings are not solely driven by statistical facts; crucially, the causal status of features also matters. Imagine that two features are equally prevalent among members of a kind, but that one is understood as generally being the cause of the other. Suppose then that an instance of the kind has one feature but not the other. Since the statistical facts are the same in both cases, prototype theory would seem to predict that typicality ratings of the individual would not be affected by *which* feature is lacking. However, individuals exhibiting the effect but not the cause are rated as less typical than those exhibiting the cause but not the effect (Ahn, Kim, Lassaline, & Dennis, 2000).

These results and many others suggest that our ways of categorizing things, and reasoning about things in categories, involve a richly structured knowledge base that is responsive to causal-explanatory factors as well as statistical factors (e.g., Carey, 1985, 2009; Gelman, 2003; Gopnik & Meltzoff, 1997; Keil, 1989). As we will use the term here, this is the outlook typical of the so-called theory-theory of concepts.[6] Since theory-theory, so construed, posits that our concepts or criteria for categorization and generalization are sensitive to theory and causal-explanatory structure, there is not too much to be said beyond that about the *general* features of our concepts. Rather, it may be useful to go on to consider concepts within each broad domain, e.g., natural kind concepts, artifact concepts, social concepts, mental state concepts, mathematical concepts and so on so forth. For example, a view known as *psychological essentialism* seems to provide a great deal of insight into how our natural kind concepts are structured from a very young age (e.g., Gelman, 2003; Leslie, in press b). However, this view is very likely not applicable to artifact concepts or mental state concepts, and certainly not to mathematical concepts. That sort of domain sensitivity should not be seen as a failing of theory-theory, but as an appropriate response to the complex and myriad ways we have of categorizing things, and of generalizing on the basis of those categories.

Probably the view that fits best with the mass of empirical material on concepts is a hybrid of theory-theory and some elements drawn from prototype theory. There are terms like "red" or "dog" which we can apply rapidly and without reliance on theory, at least in some circumstances. It is natural to think that this goes by way of sub-personal processing of sensory and perceptual information with subsequent comparison with paradigmatic or prototypical sensory and perceptual profiles. The characterization of such prototypical profiles — in particular whether and to what extent they use prevalence and cue validity — is a complex piece of empirical psychological theorizing, yet to be completed. Obviously, subjects who use such prototypical profiles are not thereby using the yet-to-be-completed psychological theory. When it comes to the immediate sensory or perceptual application of terms like "red" or 'dog", subjects exploit a range of prototypical sensory and perceptual profiles, but they do not exploit a theory of those prototypical profiles.

In what follows, we will take this qualification to theory-theory as read, and emphasize instead the explicitly theoretical elements in our criteria for applying terms.

Concepts and Generalizations

One feature that the foregoing psychological theories of concepts have in common is that they all make some reference to properties that are possessed by member*s* — plural — of the target category; they all involve focus on forms of *generalization* concerning the category and its properties. According to the classical theory, the relevant generalizations are (modalized) universal generalizations; the prototype view treats them as probabilistic generalizations; and on the theory-theory, they are complex and theory-laden general beliefs. These observations suggest a possible alternative route to studying the nature of our classificatory and inferential heuristics: we should look to the empirical investigation of our earliest and most fundamental generalizations.

Suppose, for example, that it was possible to identify and describe our most basic way of forming general judgments about kinds or categories — of moving from information concerning individual members of a category to judgments concerning the category in general. It would be quite surprising if this fundamental manner of generalization was not centrally connected with the heuristics for categorization and inference concerning kinds or categories. Thus a natural and conservative empirical hypothesis would be that our conceptual heuristics in large part consist of such fundamental types of generalizations.

Recent interdisciplinary research suggests an intriguing possibility along these lines; namely that our basic way of generalizing information issues in *generic generalizations*, which are articulated in language via generic sentences such as "tigers have stripes", "lions have manes", and "mosquitoes carry malaria" (e.g., Gelman, 2010; Hollander et al., 2002; Leslie, 2007, 2008, in press a; Leslie & Gelman, 2012; Leslie et al., 2011; Mannheim et al., 2011; Tardif et al., 2011). Such generic sentences exhibit a puzzling truth-conditional profile, as a few familiar examples quickly illustrate. Consider, for example, "lions have manes" — this strikes most people as obviously true, yet only mature male lions have manes. Thus, there are perfectly normal lions (i.e. female ones) who lack manes, and yet the generic seems true. Further, there are *more* male lions than there are maned lions (since some males are immature or lack manes for genetic or environmental reasons), yet the generic "lions are male" is widely rejected. Perhaps more puzzling are generics such as "mosquitoes carry malaria", which are accepted despite the fact that only about one percent of mosquitoes carry the virus. Yet, generics such as "books are paperbacks" are robustly rejected, even though over eighty percent of books are paperbacks (for more discussion of generics, see Carlson & Pelletier, 1995; Cohen, 1996; Leslie, 2007, 2008; for

empirical investigation of people's judgments of these sorts of generics and others, see Prasada, Khemlani, Leslie, & Glucksberg, in press).

Most importantly for our purposes here, generic generalizations are obviously not equivalent to universal generalizations, as is illustrated by the truth of "lions have manes" and "mosquitoes carry malaria". Even generics such as "tigers are striped" and "dogs have four legs" tolerate exceptions in a way that their universal counterparts do not. "All tigers are striped" is falsified by a single stripe-free albino tiger, and similarly for "all dogs have four legs"; the generics are more robust, however, and remain true in the face of exceptions. Thus, if generic generalizations constitute our most fundamental way of making general judgments about categories, this would seem to raise a further empirical challenge for the classical view. A proponent of the classical view would have to argue that the general information employed in our classificatory heuristics does not originate from our most basic way of forming general judgments. The information that we use to identify members of a category would not come by way of our basic means of forming general judgments about the category. This is not incoherent, but given the overwhelming absence of empirical evidence in favor of the classical view, it amounts to positing another unmotivated defensive epicycle.

But why should we think that generic generalizations are more cognitively fundamental than universal ones? Some of the data in favor of this hypothesis comes from the study of language acquisition. As noted, generics have a very complex truth conditional profile; providing an account of when generic sentences are true or false is a quite demanding task (see e.g., Carlson & Pelletier, 1995; Cohen, 1996; Leslie, 2008). In contrast, it is very easy to provide an account of when universally quantified statements are true ("all Ks are F" is true iff the set of Ks is a subset of the set of Fs). In light of this, one might expect that universals would be easier for young children to acquire and process than generics; however, this is precisely the opposite of what we find. Generics are produced and understood by preschool-aged children, and the data collected to date suggest that these young children have a remarkably adult-like understanding of generics. For example, preschoolers who know that only 'boy' lions have manes will accept "lions have manes" but reject "lions are boys" — despite implicitly understanding that there are at least as many 'boy' lions as there are maned lions (Brandone, Cimpian, Leslie, & Gelman, 2012; see also Gelman & Raman, 2003; Gelman, Star, and Flukes, 2002; Graham, Nayer, & Gelman, 2011; for a summary of available evidence on generic acquisition, see Leslie, in press a).

Preschoolers are generally competent with the quantifier "all" *when it is applied to a specific set of individuals* (e.g., Barner, Chow, & Yang, 2009). For example, if preschoolers are shown six crayons and asked "are all *of these* crayons in the box?" they are usually able to answer the question correctly. Most of the work on quantifier acquisition has focused on such

situations; however it should be clear that these sorts of limited, non-projectable pseudo-generalizations are not the sort that are involved in conceptual heuristics. The question, then, is how young children fare with open-ended, category-wide universals — not "all of these crayons", but "all crayons". Several studies indicate that they have considerable difficulty processing universal quantifiers in such kind-wide generalizations. Most intriguing, though, is that when preschoolers are confronted with such kind-wide universals, they do not simply provide random, incorrect answers, instead *they treat the universals as though they were generics*. That is, preschool children not only consistently evaluate generics just as adults do, they also evaluate kind-wide universals as generics (Hollander et al., 2002; Leslie & Gelman, 2012; Tardif et al., 2011; for a detailed review, see Leslie, in press a). In addition to English-speaking children, such findings have also been documented among Mandarin Chinese- and Quechua-speaking children; similar results have also been found with other quantifiers (Brandone, Gelman, & Hedglen, submitted; Hollander et al., 2002; Mannheim et al., 2011; Tardif et al., 2011).

Importantly, these findings are just what one would expect on the hypothesis (Leslie 2007, 2008, in press a) that generics, unlike universals (or "some"- or "most"-quantified statements), articulate cognitively fundamental generalizations. If the cognitive system has a basic, default way of forming general, open-ended judgments then it may sometimes fall back on this means of generalizing when asked to process a more taxing and sophisticated generalization. This tendency would be most pronounced in young children, who would be expected to struggle with the more taxing generalizations. Not only do young children apparently not struggle with generic generalizations, they substitute their understanding of the generic when asked to consider category-wide quantified generalizations.

If generics truly do article cognitively fundamental, default generalizations then one would expect that these effects might not be limited to young children. Adults might also be susceptible to the error of treating quantified statements as generics. Indeed, under a variety of circumstances, adults do show a robust tendency to accept universally quantified statements such as "all ducks lay eggs", despite knowing that male ducks do not lay eggs (where the tendency to accept the universal was *not* due to participants interpreting the universal as quantifying over only females, or over sub-kinds of ducks; Leslie et al., 2011; see also Meyer, Gelman, & Stilwell, 2010). This finding would be explicable if adults were not always evaluating the universal claim, but were instead, like preschoolers, sometimes substituting their evaluation of the corresponding generic. Further confirming evidence can also be found in the study of adult reasoning errors. For example, Steven Sloman (1993, 1998) investigated adults' evaluations of arguments that involve the quantifier "all", and found that their evaluations did not conform to the logic of universal quantification. His participants judged that arguments such as (A) are strictly stronger than arguments such as (B), despite judging that reptiles are indeed animals:

(A) All animals use norepinephrine as a neurotransmitter; therefore all mammals use norepinephrine as a neurotransmitter
(B) All animals use norepinephrine as a neurotransmitter; therefore all reptiles use norepinephrine as a neurotransmitter

This pattern of judgment is simply mistaken given the logic of the universal quantifier; however, if we replace the universals in the arguments with generics, then the judgments of the participants would be very reasonable. Since generics tolerate exceptions, the claim "animals use norepinephrine as a neurotransmitter" can be true even if some animals are exceptions to the claim. If one judges then that reptiles are more likely than mammals to be exceptions to the generic, then argument (A) is indeed stronger than argument (B). Hence these results are as one would expect if adults have a tendency to evaluate universals as generics.

Note that adults also judge that universals such as "all ravens are black" are more likely to be true than universals such as "all young jungle ravens are black", despite understanding that the latter are a subset of the former (Jönsson & Hampton, 2006). Again, this is incoherent if one is really dealing with universally quantified statements; however, if one were instead evaluating these universals as generics, this would be a reasonable judgment, since for all one knows young jungle ravens may be exceptions to the generic "ravens are black". These results are thus naturally read as lending support to the hypothesis that adults are treating these universals as generics. As a further piece of converging evidence from another experimental paradigm, it has been found that both preschoolers and adults *recall* previously presented quantified statements as generics (Leslie & Gelman, 2012).

The hypothesis that generics, unlike quantified statements, articulate cognitively fundamental, default generalizations thus has a fair amount of empirical support at this time. As a final observation in favor of the hypothesis, we might note that quantified statements require a phonologically articulated element, namely the quantifier itself. That is, we say "*all* tigers are striped" or "*most* tigers are striped"; however, in the case of the generic, there is no corresponding articulated element (e.g., "*gen* tigers are striped"). This is not an isolated fact about English, but rather it would appear that few, if any, natural languages have a dedicated, articulated generic operator (Carlson & Pelletier, 1995; Dahl, 1985).

The generics-as-default-generalizations hypothesis offers an explanation for this otherwise puzzling fact: if one wishes to interact efficiently with a system, and the system has a basic, default way of proceeding or performing a task, then one need only issue an explicit instruction to the system if one wishes it to *deviate* from this default way of proceeding. To convey the idea in more intuitive terms, if one is dealing with a child who, say, by default does not pick up her toys, one only needs to say something if one wishes the child to *deviate* from her default and actually pick up her toys. If one does not wish the child to pick up her toys on a given occasion, it would be a waste of breath to say "don't pick up your toys!" since this is what will happen even if one remains silent. Thus,

quantifiers may be articulated in language because one needs to *tell* the cognitive system, as it were, to deviate from its default, generic mode of generalizing, and instead generalize in the universal manner or the existential manner, and so on and so forth (for more details, see Leslie, 2008, in press a). Generics, by virtue of expressing the default mode of generalization, require no such phonological marking.

The Nature of Generic Generalizations

Suppose that, in the light of the foregoing, we were to go further and adopt the *generic encoding hypothesis*; namely that the heuristics which most fundamentally guide our use of terms are properly formulated in generic terms.[7] How would this impact the extant psychological theories of concepts? That depends on just how we as theorists should understand generic generalizations. There would be minimum impact on the classical view if generics were to be understood either as *ceteris paribus* generalizations as in ''other things being equal, all ravens are black" or as universals over what is normal as in "all normal ravens are black". Likewise, the prototype theory would find support if generics could be explained either in terms of the conditional probability of the kind in question having the feature in question or in terms of the cue validity of the feature; that is, its predictive value as an indicator of the kind.

Universally quantified generalizations are easily analyzed, since "all Ks are F" is true just in case the Ks form a subset of the Fs. Yet, as we have seen, generic generalizations of the form "Ks are F" are not universals, since they tolerate exceptions. Can such generics be analyzed as equivalent to claims that are quantified with "most", "usually", or "generally"? This might seem promising for generics such as "tigers are striped" or "ravens are black". There are, however, generics that can be true even though most members of the kind *lack* the property: consider "mosquitoes carry West Nile virus", "lions have manes", "sharks attack swimmers". Paraphrasing these true generics with "most", "usually" or "generally" results in false claims, suggesting that this is not the correct analysis. Conversely, a range of false generics become true when paraphrased with "most", "usually" or "generally": "most school teachers are female", "usually, books are paperbacks" and "generally, humans are right-handed" are all correct claims, yet the corresponding generics ("school teachers are female", "books are paperbacks", "humans are right-handed") would seem to be false. Clearly this latter point also rules out the idea that generics are equivalent to any logically weaker claims, such as statements quantified with "some".

In the semantics literature, it is often proposed that generics are in some sense or other equivalent to claims about what is normal for members of the kind. There are subtle variants on these accounts (see, e.g., Pelletier & Asher, 1997), but they share the core notion that generics tell us something about the

properties of all of the normal members of the kind. Again, this seems promising for generics such as "tigers are striped", since albino tigers may be counted as in some way abnormal. More challenging are generics such as "lions have manes" and "ducks lay eggs", since there is nothing abnormal about maneless female lions, or eggless male ducks. The standard response to such examples is to suggest that these generics involve covert restriction to the relevant sex (or other natural sub-kind) — that is, "lions have manes" is just elliptical for the claims "male lions have manes", which is arguably amenable to an 'all normal' analysis.

However, if it sufficed for the truth of a generic that all the normal members of one sex have a property, we would expect that a range of clearly false generics would instead be true. For example, it is surely true that all normal male lions are male; why then does the generic "lions are male" strike us as clearly false? Further, it is normal for male ducks to *not* lay eggs, yet "ducks don't lay eggs" would seem to be false (for further discussion of the point, see Leslie, 2008). Recent empirical evidence also raises some additional difficulties for any approach that appeals to implicit restrictions. Andrei Cimpian and his colleagues investigated whether people think that a kind in which a property is had only by one sex (that is, where an effective domain restriction to a single sex is possible) is a better satisfier of a generic such as "Xs have manes" than a kind in which half the members, regardless of sex, have the property (that is, where no such effective restriction to single sex is possible). If generics of the form "Xs have manes" are only accepted because people are implicitly restricting the domain to a sub-kind whose members normally possess the property, then they should think that a kind in which the property is had by one sex is a better satisfier of the generic – that is, a better satisfier than a kind in which half the males and half the females have the property. However, no such preference was found (Cimpian, Gelman and Brandone 2010; see also Khemlani, Leslie, & Glucksberg, 2009, for distinct empirical considerations against the notion of implicit domain restriction).

If generics such as "lions have manes" do not implicitly mean "male lions have manes", then they would seem to count against the analysis of such generics in terms of what is normal for members of the kind. There is after all nothing abnormal about the maneless female lions. Further challenges are also posed by generics such as "mosquitoes carry West Nile virus" and "sharks attack swimmers". Very few members of the kinds in question have the relevant properties, and it is hard to maintain that this tiny minority represents what is normal, while, e.g., the virus-free mosquitoes are somehow abnormal. Examples such as these suggest that generics need not always involve properties that are normal for the kind, or even for any of its sub-kinds. Nor does it seem plausible to suppose that these generics are, respectively, equivalent to "other things being equal, all mosquitoes carry the West Nile virus" and "other things being equal, all sharks attack swimmers". There is another possible approach to the analysis of generics, one which might be favored by prototype theorists trying to accommodate the generic encoding hypothesis. Perhaps some generics such as "tigers have stripes" are accepted because the property is highly prevalent among

members of the kind (i.e., because *most* tigers have stripes), while other generics such as "mosquitoes carry West Nile virus" are accepted because if something carries West Nile virus, it is very likely to be a mosquito — that is, because the *cue validity* associated with the generic is high. If such an account of generics were successful, then this would constitute evidence in favor of the prototype theory of our classificatory heuristics. Since prototype theory standardly involves features that are either highly prevalent or highly diagnostic (or some combination of the two) it would then fit well with the generic encoding hypothesis, and it could naturally assimilate the considerable evidence in favor of that hypothesis.

Problems arise for such an account along several dimensions. First and most importantly, even the *combination* of high prevalence and high cue validity does not suffice for the truth of a generic. Consider examples such as "books are paperbacks" or "tigers are Bengal tigers", (As it happens, there are six remaining subspecies of tigers in the world, with Bengal tigers being more numerous than all the other subspecies combined). In both cases, the majority of the members of the kind have the property in question, and furthermore the property is highly diagnostic of being a member of the kind; still these generics seem to be false. Second, recent empirical work suggests that peoples' judgments of low-prevalence generics (such as "mosquitoes carry malaria") are not especially sensitive to their estimates of cue validity. If low-prevalence generics were accepted only because the associated cue validity was high, then one would expect that people's judgments of these generics would scale to some extent with their estimates of cue validity. However, this was found not to be the case (Prasada et al., in press).

If the notions of normalcy, prevalence and cue validity do not allow us to give an account of generics, where does that leave us? As noted above, it would be reasonable to suppose that the best account of our classificatory heuristics should mesh with our best account of generics, since the former involves features that are generalized to kinds and categories, while the latter are language's way of letting us articulate our most cognitively basic generalizations. That is the basis of our advocacy of the generic encoding hypothesis. We have also just seen that generics are simply not equivalent to universals (which means the generic encoding hypothesis does not mesh with the classical view), nor do they seem especially amenable to treatment in terms of probabilistic notions such as prevalence and cue validity (which means that the generic encoding hypothesis does not mesh with the prototype theory). Perhaps, then, in accord with the theory-theory, we need to take into account theoretically rich, content-based factors in giving an account of generics.

Consider, for example, the low-prevalence generics that have been discussed throughout this section: "mosquitoes carry West Nile virus", "sharks attack swimmers", "ticks carry Lyme disease". We might add to the list as follows: "pit bulls maul children", "tigers eat people", "lead toys poison children", and so on so forth. Once we are on the lookout for *content-based* factors — as opposed to merely formal statistical relations like prevalence or cue validity — we can make something of the fact that these generics all share a common feature: the

predicates in question all express properties that make their possessors dangerous or threatening; they involve the sort of property about which we would wish to be forewarned. Indeed, it is natural to suppose that it would be beneficial for us if our most basic way of generalizing was sensitive to such a higher-order feature of the properties predicated of the relevant kinds. Perhaps, then, generics are accepted even at low prevalence levels if the property in question has this type of higher-order feature (see Leslie, 2007, 2008, in press c for detailed discussion; for empirical support see Cimpian, Brandone, & Gelman, 2010).

If generic generalizations are indeed sensitive to the nature of the property being generalized, might that account for other 'troublesome' generics? Suppose that, as part of our theoretical knowledge — as this is understood by the theory-theory — concerning animal kinds, we register that animals kinds tend to be very similar to each other at a certain level of abstraction: by and large, animal kinds have characteristic noises, characteristic modes of locomotion, characteristic diets, characteristic salient physical features, characteristic methods of reproduction/gestation, and of nurturing the resultant young. Outside the domain of animal kinds, artifact kinds have characteristic functions, kinds of professionals have characteristic social roles, and so on.

Suppose that this general structural knowledge is exploited by our cognitive system, so that for each of the relevant characteristic dimensions for each type or kind with which we are presented, we first seek to fill in the relevant values. We might call these values the *characteristic properties* of the kind. Examples of characteristic properties of kinds would include salient, distinctive physical features (e.g., "lions have manes"), methods of reproduction and nurturing the young for animal kinds (e.g., "ducks lay eggs", "pigs suckle their young"), and for artifact kinds, functions (e.g., "Orange-Crusher-2000s crush oranges" — which could be a true generic even if every Orange-Crusher-2000 is destroyed before it is used).

There is thus a sense in which these generics are 'answering an implicit question' about, e.g., how ducks reproduce, or what the function of Orange-Crusher-2000s is. They do not reflect merely statistical claims about how prevalent the property is among members of the kind. Indeed, if a generic attributes a characteristic property to the kind, then the generic may be accepted even if few members of the kind have the property in question (see Leslie, 2008 for more details).

Given that generics are indeed sensitive to rich, content-based factors, it may be possible to explain why some generics are accepted even though the property is not prevalent among members of the kind. What, though, of generics such as "books are paperbacks", "school teachers are female", or "Canadians are right-handed"? People tend to reject such generics, despite judging that the property in question is highly prevalent (Prasada et al., in press). One proposal is that generics may be sensitive to the nature of the *exceptions* to the generic claim (Leslie, 2007, 2008). That is, amongst the members of the kind that fail to have the

predicated property, it may matter *how* they fail to have the property; in particular whether they simply lack the property, or whether they have an equally salient, concrete, positive property *instead*. "Ticks carry Lyme disease" is accepted even when the prevalence estimate is low, but the non-infected ticks are known to simply *not carry Lyme disease;* they do not have a specific alternative property instead. However, the books that are not paperbacks are instead hardcover. Similarly, people who are not right-handed are instead *left-handed*; elementary school teachers who are not female are *male*. Intuitively, one might feel as though one would be 'overlooking' these hardcover books, these left-handed individuals, and these male school teachers if one accepted the generics in question. Thus, a promising proposal is that generics may be also sensitive to the nature of the exceptions to the generalization.

Of course, the study of generics does not provide decisive evidence regarding the nature of our heuristics and criteria. However, if the best account of generic generalizations meshes with theory-theory but not other accounts, this could be naturally construed as providing support for the theory-theory approach to categorization and inference. In our terms, the theory-theory fits best with the generic encoding hypothesis; the empirical claim that the heuristics or, more generally, the criteria we use in applying terms frequently take a generic rather than a universal form. Therefore, in what follows, we will assume that the theory-theory offers the best account of our heuristics of categorization, and that the generalizations involved are articulated in language via generic sentences, whose interpretations are governed by rich, content-based factors.

This tentative vindication of the theory-theory, because of its capacity to incorporate the generic encoding hypothesis, may initially seem to raise the stocks of the Canberra plan. To the casual eye the Canberra plan, focusing as it does on the theory, or alternatively the set of platitudes, that guides us in the use of a term may seem to best comport with the theory-theory of concepts. As we shall see, this first impression is badly wrong; the theory-theory is plausible only as an account of the heuristics or criteria we employ in using a term, while the Canberra plan is intended as a general account of how we should specify the application conditions of a term. Moreover, as we will see, the generic character of many of our platitudes raises quite systematic problems for the Canberra plan, at least when it is applied to arenas where the relevant generic platitudes are not backed up by the corresponding universal truths (as they typically *are* in logic and mathematics).

Some Suggestions about "Translation"

As even this brief review of the psychological literature makes clear, when relating philosophical and psychological theories of concepts, we need a better translation scheme than the following

> "Concept" in the psychological literature means pretty much what philosophers who have accepted significant parts of the substantial theory have meant by it.

Otherwise, philosophers might all too easily follow Jerry Fodor's lead in despairing of psychological theories of concepts, while psychologists will be immediately led to think that the philosophical discussion is derisory.

The naïve translation principle implies that psychologists should be interested in the application conditions of concepts, but of course they are not. They show no interest in studying those who would be best placed to know the application conditions of concepts, namely clear-headed taxonomists and experts in philosophical analysis. Indeed, given the naïve translation manual, it would have be said that psychologists exhibit what can only be seen as a curious and indefensible fixation; they study the cognitive performance of untutored adults and even of young children! So they might discover, for example, that children and untutored adults withhold the term "fruit" from tomatoes, olives and peppers. But that just entails that such people are not full masters of the concept *fruit*, which does indeed apply to these ovaries of flowering plants. That is, these subjects simply do not know the application conditions of the term "fruit". Why then is *their* performance at all relevant in the study of concepts? Is this not an *appalling* waste of time and money?

That line of questioning, which is on its face rather silly, would be quite sensible if the naïve translation were adequate. So it is not adequate. We should do better.

Often psychologists simply mean by "concept of Ks" what ordinary speakers mean, namely a contextually relevant part of someone's conception or total theory of Ks. In *Doing Without Concepts* (2009), Edouard Machery has recently argued that in so far psychologists mean just that, and in so far as they appeal to *different* parts of a subject's conception of Ks when they explain her perception of Ks, memory of Ks, classification of Ks and inferences concerning Ks there is then nothing uniform to go under the heading of "a psychological theory of concepts".

However, there seems to us to be more relevant uniformity in the psychological literature on concepts. Mostly psychologists mean by "concept of Ks" psychologically real clusters of heuristics, or more generally criteria, which guide us in classification, characterization and inference concerning Ks. The crucial interpretive point is that most psychologists working on concepts are not concerned with what an analysis would deliver, namely application conditions of concepts, but with heuristic criteria for classification, characterization and inference. If this is right then we should also avoid the opposite error that could be driven by the naïve translation principle, namely the idea that we could use the experimental methods of psychology, or the methods of opinion surveys, to directly study not just the criteria or heuristics we use, but the application conditions of our concepts, thereby letting the empirical work do our analyses for us, as it were.

Three Main Theses

This overview of the relation between philosophical and psychological theories of concepts naturally suggests the following thesis:

> Philosophers who believe in concepts/conceptual knowledge/the distinctively philosophical a priori should treat the psychological theories outlined above, along with other psychological theories in the same line of country, as in significant part accounts of the *heuristics or more generally the criteria* we actually use to apply terms to things.

This, in its turn, suggests a second thesis, concerning the crucial relevance of the psychology to philosophical theories of concepts:

> Philosophers should then ask whether and to what extent their theories of concepts, when combined with the best psychological accounts of the heuristics we actually use to apply terms, will entail massive underdetermination in respect of just which concepts we are using those terms to express.

Thirdly,

> To the extent that there is such massive underdetermination of concepts and their associated application conditions by the criteria we use in applying terms, philosophers should lose confidence in the usefulness of conceptual analysis, understood as the making explicit of the conditions of application of concepts by means of the method of cases. To that same extent, we should be pessimistic about the range of the distinctively philosophical a priori.

Dogs, Wolves and German Shepherds

In order to illustrate the last two theses, suppose that in actual fact our heuristics for applying the term "dog" are exhausted by the following down-and-dirty criteria:

> Is it an animal?
>
> Does it have one of the characteristic looks, smells, coat textures, etc, of one of the familiar kinds of dogs?
>
> Is it the offspring of an animal with one of the characteristic looks, smells, coat textures, etc, of those one of those familiar kinds of dogs?

As things actually go in suburban environments these three criteria may be good heuristics for collecting together observed instances of the kind dog, the kind we now know to be the species *Canis lupus familiaris*. However, this is due

to the contingent fact that the canines around us are almost all of them from that species. In fact *Canis lupus familiaris*, as the name suggests, is a subspecies of the kind grey wolf, *Canis lupus*, which includes wolves that are not dogs, but which happen to look very like German Shepherds. Wolves typically don't roam in suburban neighborhoods. So, what we in fact track by means of these heuristics are dogs, but in another environment the same heuristics might lead us to track a larger group, namely the grey wolves.

What concept in the philosophical sense should we then assign to our word "dog": the concept that comports with membership in *Canis lupus familiaris* or the concept that comports with membership in *Canis lupus*? Notice that some background intention to use "dog" to refer to *the same sort of animal* as the ones around us won't really help here. First, we may not in fact have that intention, and yet presumably we would still have a concept associated with "dog". Second, the intention itself does not eliminate *Canis lupus* from contention, for there is no good reason to suppose that all the grey wolves considered together are not *the same sort of animal*. In general there are too many *sorts of animals* for underdetermination to be reduced in this way.

Moreover, even on the assumption that these heuristics enable us in our environment to use "dog" to pick out members of *Canis lupus familiaris*, the heuristics themselves will be poor candidates to figure in the analysis of the concept *dog*. For it is not a priori and necessary that

> If x is an animal which has one of the characteristic doggish looks, smells, coat textures, etc., and is the offspring of that ilk then x is a dog.

Other animals can be made to look like dogs. A coiffed squirrel can be made to look like a chihuahua, and so can the squirrel's immediate forebears. The lesson is that the heuristics are simply *indicators* of when something around here is a dog. It is a bad verificationist error to mistake such heuristics for the application conditions of the relevant concept.

A similar argument goes through on the assumption that these heuristics in fact enable us, at least in our typical environment, to pick out the kind *Canis lupus*. Animals that are not grey wolves can be made to look, smell and feel like grey wolves. So can their immediate forebears. Satisfying the criteria for applying a concept — or better, for applying a mental representation or a term — is not satisfying the application conditions (in the philosopher's sense) for the concept associated with the representation or term.

Indeed, once the criteria/conditions distinction is made, why should we think that application conditions rather than our criteria are what is articulated by our judgments about cases? Once that distinction is made clear, the method of appealing to our judgments as to whether we should apply or withhold a term in a variety of imaginary cases is obviously a way of manifesting our criteria or ways of telling whether the term applies. It is not obviously a way of manifesting

our 'implicit grasp' of the application conditions of terms. Indeed, the empirical explanation of our use of terms, by way of one or another psychological theory of concepts, seems to *drive out* the supposed explanation by way of implicit grasp of application conditions. But the supposed explanation was doing the crucial work of justifying the method of cases as the royal road to the analysis of our concepts.

A Temptation to be Resisted

Some may still be tempted by the thought that *if* our heuristics for applying the term "dog" are in fact just exhausted by the down-and-dirty criteria

> Is it an animal?
>
> Does it have one of the characteristic familiar doggish looks, smells, coat textures, etc.?
>
> Is it the offspring of an animal with one of those characteristic looks, smells, coat textures, etc?

then it just follows that the concept we express by "dog" is the concept of an animal which has one of the characteristic domestic doggish looks, smells, coat textures, etc., and which is the offspring of an animal with one such perceptual profile. Accordingly, they will say that given how we use the word "dog" it is a priori and necessary that

> x is a dog iff it is an animal with one of the characteristic patterns of domestic doggish looks, smells, coat textures, etc., and it is the offspring of that ilk.

Here we no longer have underdetermination, understood in terms of there being multiple kinds — *Canis lupus familiairis, Canis lupus* — that are equally good candidates to fall under the concept *dog*. Instead we have lack of specificity in the concept expressed by "dog"; it turns out to be a concept under which the more general kind *Canis lupus* falls.

The problem then is to explain how it could have been rational on our part to accept the scientific discovery that

> All dogs are members of the kind *Canis lupus familiaris*.

At the very least, accepting this will involve a conceptual change, by changing from the topic we had in mind when we use "dog" to a much more specific topic. That is odd, for it suggests that we could reasonably reject such scientific discoveries out of hand because they confuse a kind with one of its sub-kinds in precisely the same way as the following claim does:

All human beings are men.

Obviously, such a rejection of the scientific discovery that all dogs are members of the kind *Canis lupus familiaris* is wildly out of place, therefore the temptation that prompts it is not to be followed.

What then is wrong with the temptation to make the conditions of application of a concept exactly as unspecific as the heuristics which govern its use? Once again, it confuses useful heuristics or criteria for the employment of a term with a priori and necessary conditions for the term's correct application.

The usefulness of a heuristic is invariably situation-dependent. In the suburbs of New York, the heuristics are a fair guide to the presence of a member of *Canis lupus familiaris*. In the Arctic hinterlands, where there are lots of wolves, they are not. The conditions of application of the concept *dog* are not in this way situation-dependent.

Are There Any Concepts in the Sense of Things to be Analyzed?

Philosophical reflection on the psychological literature yields the following straightforward challenge:

> Premise 1: The concept that subjects express in their use of a term "T" is not fixed by the heuristics that subjects employ in applying "T".
>
> Premise 2: All that subjects could have introspective or armchair access to, even in the best case of articulating what they know about Ts by means of an large inventory of real and imagined possible cases in which "T" is applied and withheld are (i) the various deliverances of their heuristics in the real and imagined possible cases, and (ii) by way of inferences to the best explanation of such deliverances, the details of their own heuristics.
>
> Conclusion: The method of trying to extract a conceptual analysis, a statement of the necessary and sufficient conditions of application of a concept, from armchair intuitions about cases typically will not work. For if there is a concept that subjects express in their use of a term "T", the relevant heuristics will typically underdetermine which concept that is.

This challenge has only been illustrated, not vindicated, by the toy example of our heuristics for "dog". For it is open to someone to maintain that when we really take *all* the heuristics or criteria that we employ in applying "T" into account then the conditions of application of the concept expressed by 'T' *are* determined or fixed (or fixed modulo the further facts about which concepts, properties or extensions are privileged attractors).

That response will be taken up in detail later in the paper, in the context of discussing the general program of philosophical analysis known as the Canberra plan. Obviously, the actual facts about the kind of heuristics or criteria we really

are employing are highly relevant to the claim that the totality of the heuristics or criteria we use determines the conditions of application of the relevant term or concept. The overall challenge from psychology can only be met — if it can be met at all — by looking at what psychologists have discovered about these matters.

As we noted above, recent psychological results are now converging on the suggestion that many of the heuristics or criteria we use in applying terms involve beliefs that are best expressed in generic form. If this is so then we believe that it can be shown that it is very unlikely that the totality of our heuristics or criteria fix the application conditions of the relevant terms.

Isn't "Reference Fixing" Analysis Still Viable?

Once upon a time, there *was* a privileged way of telling whether something was Neptune; all we had to determine was whether it was the planet which caused perturbations in the orbit of Uranus. And, arguably, this way of telling was semantically tied to the name "Neptune" in the sense that this name was in fact introduced to denote whatever planet was causing the already observed perturbations in Uranus. According to the well-known tale, Alexis Bouvard observed odd perturbations in the orbit of Uranus and hypothesized that these were caused by the gravitational influence of an unknown planet, which was then dubbed "Neptune". Then, in 1846, Johann Galle discovered the planet in question. Some would then say that it is a priori that

(1) Neptune is the cause of perturbations in Uranus

and hence that it is a priori and necessary that

(2) Neptune is the actual cause of perturbations in Uranus

Finally, someone might say, we now have just what we wanted from an analysis. We have identified a privileged way of telling whether something is Neptune, a way of telling that is semantically tied to the term "Neptune", and one which can be used to guide us in determining the application conditions of the term across all possible situations. All we have to do is look to see whether the thing in the possible situation is identical with the actual cause of perturbations in Uranus.

The idea can be generalized if we indulge ourselves in various 'as-if' stories about the introduction of our terms. Suppose the term "water" was introduced to denote the same stuff as the potable liquid found in rivers, lakes and streams, and that falls from the sky when it rains. Then, according to the line of thought in question, it is a priori and necessary that

(3) Water is the stuff that is the actual potable liquid found in rivers, lakes and streams, and that falls from the sky when it rains.

Once again, we have something that can play the role of an analysis of the concept *water*, for if (3) is a priori and necessary then it specifies the application conditions of "water" across all possible situations.

Why then is there any problem about analysis? Analysis is just the articulation and rigidification of the reference fixers of our terms!

However, as our brief survey of the psychological facts suggests, ordinary users of terms do not themselves privilege such reference fixers among their ways of telling. Indeed, as in the case of "water", they often apply terms on the basis of prototypical perceptual profiles, profiles which may resist any discursive characterization that could appear on the right hand side of an analysis, even a reference fixing analysis. Do they then have different concepts from those who explicitly treat (2) and (3) as a priori?

The friend of reference-fixing analyses need not say this. He or she can say that, as Kripke and Putnam have shown, ordinary users of the terms "Neptune" and "water" respond to possible cases as if they did treat (2) and (3) as a priori and necessary. So although they have anticipated one important upshot of the psychological literature, namely that ordinary users of terms employ at most fallible heuristics or criteria in their use of those terms, Kripke and Putnam have nonetheless shown us the way to provide thoroughly modern conceptual analyses: namely, find the reference-fixer for the term in question and rigidify the reference-fixing condition. It is Frank Jackson, more than anyone else, who has done the most to advance this route to refurbishing the credentials of analysis.

The main difficulty for the thought that reference fixers can provide analyses is that, although in some cases reference fixing descriptions may have played a 'semantic' role precisely in serving, *at a particular time*, to fix the reference of a term, it is a notable fact that even in these very cases the relevant rigidified reference fixers need not in fact fix the *subsequent* application conditions of terms across all possible situations. That is, (2) and (3) and their various analogs are not a priori and necessary.

To see that (2) and its ilk are not a priori imagine that after Johann Galle's observation of Neptune, Alexis Bouvard subsequently noticed odd perturbations in the orbit of Neptune, though of a less dramatic sort than those he found in Uranus. Bouvard then hypothesizes that these too are the gravitational effects of still another planet, one that remains to be observed. The question arises as to how to name the planet that is doing this to Neptune, and it comes to be called "Pluto". Galle then subsequently observes Pluto. If history had gone that way, then if (2) is a priori and necessary, the same would hold of

(4) Pluto is the actual planet that is causing perturbations in Neptune.

However, (4) is certainly not a priori and necessary, even in the scenario imagined. For our little story about Galle's second success is compatible with what actually happened in the early 21st century, when shockingly, Pluto was demoted from the status of a planet, because it turned out to be more similar to large

asteroids than to planets. If (4) were indeed a priori and necessary that could not have happened. Instead the only option would to have been to conclude that Pluto never existed. Likewise, we could discover that Neptune is a huge, well-disguised, alien spaceship, so (2) is not a priori and necessary. The crucial lesson is this: an established pattern of usage can override original reference fixers. So long as we are not dealing with what Putnam called 'one-criterion' terms, as, perhaps, in the artificial example of the name "Julius" introduced just to denote the inventor of the zipper, we can find examples where the term survives the falsity of its initial reference fixing description. Reference fixing descriptions for terms do not in general hold a priori of the items that the terms pick out.

This is even more obvious in the case of so-called natural kind terms such as "water". (For an extensive discussion of crucial, but often suppressed, complexities surrounding natural kind terms see Leslie (in press b).) We can discover that the water cycle is much more complicated than we supposed, so that the stuff that falls from the sky changes its chemical structure as it approaches the surface of the earth, with the consequence that there is no natural unity in the stuff that falls from the sky and the stuff in lakes and streams. This is not the discovery that water does not exist, as it would have to be if (3) is a priori. Again, an established pattern of usage can override almost any original reference fixer.

There is an appealing response that Frank Jackson and David Chalmers offer to these kinds of counter-examples (Chalmers & Jackson, 2001; Jackson, 1994, 1998). They say that, in order for us to have these very intuitions about the possible cases in question, we must be using some other criterion for being Pluto, or Neptune or water. Otherwise, how can we identify the possible situation before us as one in which we have Pluto, though (4) is false, or as one in which we have Neptune, though (2) is false, or as one in which we have water, though (3) is false?

That may seem like a compelling question, at least if we take seriously what Kripke (1980) called the 'vernoscope picture' of our relation to possible worlds. On this picture, a possible world is presented to us neutrally, leaving open whether it contains Pluto, or Neptune, water or whatever target phenomenon or feature. It is then up to us to bring to bear our a priori and necessary criteria for telling whether it contains Pluto, or Neptune or water or whatever other target we have in mind. An articulation of our ultimate a priori and necessary criteria for telling will then be the proper analysis of the term for the target. So, say Jackson and Chalmers, counterexamples to analyses are themselves indications that we implicitly possess a priori and necessary conditions for telling, i.e. that we know the true analysis.

However, the vernoscope picture, as Kripke (1980) emphasized, systematically misrepresents the nature of our thoughts about possibilities. We happen to know that certain things are possible, and this knowledge of possibilities already comes laden with subject matters, e.g. Pluto, or Neptune, or water. Possible

worlds neutrally described are not our basic epistemic starting points in arriving at knowledge of possibilities; they are simply useful devices for defining validity in modal languages.

In particular, in urging the argument above, did we actually find ourselves in such a situation of surveying a possible world neutrally described so that we then had to bring to bear a priori and necessary criteria to determine whether that possible world contains Pluto, or Neptune, or water? No, for the purposes of the argument above, we simply found ourselves in this epistemic situation: the following individual claims are very plausible

> Pluto could exist even if (4) were false.
>
> Neptune could exist even if (2) were false.
>
> Water could exist even if (3) were false.

Clearly we do not need to be relying upon even an implicit analysis of "Pluto", "Neptune" and "water" to be in that, fairly minimal, epistemic situation. That should be clear because the possession of an implicit analysis is a very strong condition; an implicit analysis would decide every case. For the purposes of the argument above, we needed only to decide three cases. Obviously, something much more fragmentary and thus much less committal than an implicitly possessed analysis could suffice for that. After all, our grounds for the three claims above were simply that established usage can simply override the reference fixers of old. In saying that we need simply be relying on some reasonable criteria for counting as Neptune, Pluto and water, criteria that do not themselves amount to an analysis, and which may not themselves figure in any plausible analysis of the application conditions of the terms "Neptune", "Pluto" and "water".

Laura Schroeter has offered a very useful way of understanding the general approach of Jackson and Chalmers when it comes to the analysis of our terms. She writes

> However exactly we choose to cash it out, it seems that Jackson's and Chalmers's account is ultimately grounded in the conceptual [or term-using] dispositions the subject would form after ideal reflection on hypothetical cases. The rigidified definite descriptions Chalmers and Jackson offer as an analysis of "water" are correct just in case they accurately summarize the ideal conceptual dispositions the subject would converge on, irrespective of which world she considers as actual (2004, p. 449).

The thought is that *something* is guiding our use of terms in thinking about merely possible situations. Imagine a huge list of the positive and negative verdicts about whether the term under consideration applies in a fully representative range of possible cases — verdicts we would give under ideal conditions; that is, conditions such as no distractions, plenty of computational ability, no straightforward irrationality. If we could codify the pattern found in such a list, we would have

an analysis of the concept expressed by the term; equivalently, we would have an account of the application conditions of the term. If not, we will not have an analysis, but there is no more to know about the application conditions of the term in question than the list of verdicts. What else, Jackson and Chalmers might ask, could determine the application conditions of the term?

Having worked through the toy example of dogs, we hope that the reader can now recognize this kind of idea. Once again, it amounts to the conviction that what we rely upon in making judgments about cases, namely our criteria, themselves determine the application conditions of terms.

What else could determine the application of our terms? Recall the concepts as terms-in-use view. Mutual correction while participating in a convention for using a term selects for a variety of effective heuristics or criteria that then guide individuals in their use of that term. To 'know how to use a term' is just to have some such effective criterion, and this will often fall well short of even a tacit understanding of the conditions of application of the term. Accordingly, on the concepts as terms-in-use view, there is no reason to suppose that speakers can render explicit the conditions of application of their terms simply by reflecting on how they would use them in a variety of circumstances. For such actual and counterfactual usages reflect only the speakers' criteria, which may fall short of any significant grasp of application conditions. (Remember how Sean Connery was once more central to the criteria for counting as a bachelor than were the properties of being male and being unmarried.) Thus philosophical analysis, as traditionally conceived, can at best produce fragmentary and inconclusive results, *even under conditions of ideal reflection*.

One way to press this point home is to characterize just what knowledge is required to fix the application conditions of our terms, and how far that knowledge transcends anything that could be fixed even by full knowledge of the criteria we actually use in applying our terms. As the examples of Neptune, Pluto and water themselves show, scientific discoveries can entail surprising facts about the application conditions of our terms. Neptune may not be the cause of the very perturbations which were cited in the fixing of the reference of the term "Neptune"; "Pluto" need not refer to a planet, and "water" may not apply to rain.

In fact, as we shall now see, in order to have the knowledge required to fix the application conditions of our terms, we would either have to know (i) the true total theory of the world (encompassing all the potential sources of surprises about the application conditions of our terms) or (ii) conditionals which specify *for each epistemically possible total theory of the world* just what the application conditions of our terms would be relative to that total theory. Surely both (i) and (ii) massively transcend what can be extracted from our criteria, even under ideal conditions, at least as those conditions have so far been defined.

To see just why the Jackson and Chalmers view that our dispositions to classify possible cases are rich enough to determine the application conditions of our terms leads to one or another of these positions, consider the

following example, which simply recalls the traditional Quinean worries about the inter-animation of application conditions and empirical theory. Suppose that a subject has never understood or even heard of special relativity, and suppose that its important claims do in fact figure in the total true theory of the world. When asked to clearheadedly consider a world in which there is no relevant two-place temporal relation between events that can count as a candidate to be simultaneity, our subject will wrongly conclude that there is no simultaneity in that world, because he has not had the benefit of the relevant scientific understanding of the widespread explanatory advantages associated with special relativity — the very ones that justify thinking of simultaneity as frame-dependent. The general lesson is this; if actual science as we know it can reveal surprising facts about the extensions of our terms, for example that "simultaneity" applies in a world in which there is no relevant two place temporal relation between events that can count as simultaneity, then absent the knowledge of the total true science of our world, we will be liable to make systematic errors about the extensions of our terms when presented with possible cases.

Of course, our subject can be presented with an explanation of special relativity, and then asked "What is the extension of our term "simultaneity" under those circumstances?" Then he may get it right. But suppose he is given a description of a possible case in a possible world, without knowing just how the ultimately correct scientific theory of our world reveals, as it likely will reveal, surprising facts about the extensions of our terms. Suppose, as is likely, that this forever unknown theory would be even more revisionary and surprising than special relativity. As the history of science's impact on our use of terms already reveals, if we are in ignorance of the true theory of the world, then we should treat our reactions to 'vernoscope' presentations of possible worlds — presentations in which it is left open whether the target things in question are in them — as hypotheses, not as data. They could well be overturned by subsequent scientific developments.

On the Jackson and Chalmers view that our dispositions to classify possible cases are rich enough to determine the application conditions of our terms, there are two ways of assimilating the familiar point about the unknown surprises that the true theory of the world might deliver. On the first, in order to be sure that our 'ideal' reactions to possible cases get the extensions of our terms right, we need to be in ideal conditions *that include knowledge of the surprising implications of the extensions of our terms due to the ultimately correct theory of our world.* Only then can we properly evaluate the extensions of our terms in the possible situations presented to us.

Let us suppose that for each term "T" there is such a thing as the list of those reactions we would have to the full range of possible cases, under the following conditions: we have no distractions, plenty of extra computational ability, no straightforward irrationality and *we know the surprising implications of the extensions of our terms due to the ultimately correct theory of our world.* What are we to make of that list?

First, it looks unlikely that for many terms there will be any useful finite codification of the infinite number of reactions on the relevant list. So, we may have here a final account of what would make an analysis true, without there ever being any analysis, i.e. any actual statement of an analysis, which is true. Second, given the nature of the idealization required, it would be utterly forced to say that for each term we even *tacitly know* its list, let alone the codification of the list if there is one. Third, and most importantly, we have arrived once more at a central theme of this paper; the real problem with concepts, and with the conceptual a priori and with analyses, is that they have no demonstrated use. Certainly, given the idealization in question, it is not our knowledge of them (explicit or tacit) which guides our use of terms in ordinary life, where we manifestly do not know *the surprising implications of the extensions of our terms due to the ultimately correct theory of our world.* It is rather our heuristics, our criteria, which guide our use of terms. Whereas the psychological approach to concepts as clusters of heuristics has an explanatory relation to a subject-matter, the philosophical treatment of concepts that goes by way of such highly idealized outputs does not.

The same conclusion emerges on the alternative way of assimilating the point about the surprises concerning the extensions of our terms that the true theory of our world might force. Imagine the following response to the worries just raised.

> We agree that the comprehensive scientific theories we accept can involve surprising consequences for the use of our terms. What is true for users of a term "T" is that under ideal conditions of the more restrictive sort, involving only no distractions, plenty of extra computational ability, and no straightforward irrationality, the users will be disposed to give a verdict for each epistemically possible true total theory of the actual world whether, under that theory, "T" applies in this or that possible case.

This massively multiplies the computational complexity required to generate the relevant list, and so intensifies the worry that there will be no codification of the list. But it also means that the items on the list will themselves be conditional in form; for example, when it comes to the concept of simultaneity, the relevant conditional will look like

> If special relativity holds up in the true total theory of our world then. . .

So now the obvious problem is this. Thus articulated, our conceptual knowledge leaves us in the following situation: we should make no judgment about any cases with respect to whether they fall in the extension of the term in question or not *until we know the total true theory of our world.* Hence, on this alternative way of going, our main point holds a fortiori; if this is the content of our conceptual knowledge, it not only has no demonstrated use, its lack of usefulness is itself demonstrable.

Jackson and Chalmers might insist that there is indeed one remaining use for talk of a priori or conceptual knowledge here, however idealizing it may be. For they each identify physicalist reductionism with a very distinctive thesis, namely with so-called 'type A' physicalism. This is the thesis that the deliverances of ideal reflection about all relevant cases will in principle support rich enough intermediate a priori premises to provide a derivation of the truths about all that there is, from the truths about what there fundamentally is at the physical level. Chalmers denies this thesis and Jackson now defends it, but they both understand physicalist reductionism in this same way. The fact that the deliverances of ideal reflection may not be codifiable is not crucial here, for that only means that *we cannot carry out the derivation*.

As against this highly theoretical use of the conceptually based a priori, we have only the following plausibility argument. Consider what ideal reflection, on this way of going, is. It is no more than having no distractions, plenty of extra computational ability, and being subject to no straightforward irrationality. Consider the input that ideal reflection takes in, namely the dispositions of use associated with the heuristics and criteria we employ in deciding how to use terms. Now consider the massive output needed to meet the standards of type A physicalism: for each term in our language which is not part of the logical or basic physical vocabulary we need a conditional which specifies, for each epistemically possible total theory of our world, just what the possible world application conditions of the term would be relative to that epistemically possible total theory.

Is it really credible that ideal reflection using the input yields *anything like* this output? Is this not a massive, and extremely adventurous, empirical psychological speculation?

In any case, it was an important line of inquiry to ask whether the externalist semantics of Kripke and Putnam reintroduced something like analysis by the back door. We take ourselves to have argued for a negative judgment on this more restricted point; reference fixing descriptions do not (except in very artificial cases, where there is one, simple and theory-free, criterion) hold a priori, and in arguing for that conclusion we ourselves need not be evidencing an implicit grasp of an analysis.

Total Theory as a Criterion

A theory of Ts, that is, a theory of the things to which the term "T" applies, can be readily re-interpreted as a heuristic for applying the term "T". Simply proceed as follows: first put a variable "X" for every occurrence of "T" in the statement of the theory. You then have a vast open sentence. You now look around the world for the things which satisfy the open sentence. Your heuristic is: apply "T" to *those* things and those things only.

Furthermore, if we abstract away from the actual psychological limitations which make it reasonable to use only smallish subparts of what one believes about

Ts when determining whether we have a T before us, then one's *total* theory of Ts at some given moment should be the theory from which one derives one's T-sensitive heuristic. For suppose you omit something you do in fact believe about Ts; then you are not abiding by the requirement of total evidence, the requirement that you bring all relevant beliefs to bear on the question of whether you have a T before you.

Of course, your total theory of Ts is likely to evolve over time, and so your ideal heuristic will also change over time. What changes here are not the application conditions of the term "T", but your ways of telling whether to apply it. Imagine a child who starts out with the following theory of dogs

> If x is an animal which has one of the characteristic doggish looks, smells, coat textures, etc., and is the offspring of that ilk, then x is a dog.

and then becomes a biology major, adding to her earlier theory of dogs the thesis that

> Dogs are members of the subspecies *Canis lupus familiaris*, distinct from the members of the wider species *Canis lupus* in the following ways: A, B, C.

Obviously, she now has *better* criteria for telling when something is a dog; she is less likely to confuse dogs and grey wolves. Has her concept *dog* changed as a result? Has she moved from a concept with the application conditions determined by the smaller childhood theory to a concept whose application conditions are determined by the larger adult theory? As we noted earlier, there is an everyday use of the term "concept", on which one's concept of Ks is just some contextually relevant part of one's total theory of Ks. In this sense, in certain contexts, it will be right to say that her concept has changed. However, this is manifestly not the philosophers' use of "concept", which ties concepts essentially to conditions of application. Once we have that use in mind, the thing to say is that she has just learnt more about the things the concept *dog* applies to; the extension of the term "dog" could have remained the same for her, even as she acquires more and more knowledge about the nature of the items in that extension.

Suppose instead that there were, earlier and later, two concepts in play; not in the trivial everyday sense of there being two conceptions in play, but in the philosophical sense of two extension-determiners in play. Then the biology major would have made a conceptual error in simply conjoining her later knowledge with her earlier knowledge and treating them as knowledge concerning the same kind of animal. To assume that our total theory determines the application conditions of its concepts, makes what we call "learning" typically count as conceptual confusion.

This is obviously not an objection to the total theory-theory of concepts, so long as we understand that as merely an empirical account of theory-based criteria we use in applying concepts. But if we instead treat the total theory-theory

of concepts as an account of the application conditions of concepts, then the urgent objection will be "How is significant learning — specifically learning that narrows or broadens our conception of the range of entities a term is to be applied to — possible?" Moreover, if the total theory-theory account of application conditions were true, someone could not learn a theory in the obvious way; that is, by having it explained to them using words he or she already understands. For the total theory-theory account of application conditions implies that in order to know what the words that occur in the total theory mean, one must already know the total theory.

The total theory-theory account of the application conditions of our terms faces a second objection: it wrongly represents cases of genuine disagreement as cases in which the disputants are talking past each other. Suppose, as David Lewis (1970, 1994) argues, that our total folk theory of mental states represents them as inner sates that occupy certain characteristic causal roles, in particular with respect to other such inner states, sensory input and behavioral output. On the face of it, the folk theory is at odds with behaviorism, which represented being in a mental state merely as exhibiting certain patterned connections between sensory input and behavioral output. The folk theory is even more obviously at odds with epiphenomenalism, which denies that mental states are causes, and with the parallelism of Leibnitz, which keeps the mental and the physical realms causally isolated, but running in harmony with each other, so that they behave as if they were interacting.

But now suppose that we adopt the total theory-theory view of the application conditions of our terms, and accordingly consider the four theories just discussed — folk theory, behaviorism, epiphenomenalism, parallelism — as different total theories that thereby establish different application conditions for mental state terms such as "belief", "desire", "pain" and so on. Now the manifest disagreement about the nature of the mental among the defenders of such theories is lost. There are simply four different term defining theories, and their proponents are simply talking past each other.

The familiar response to both the objection from learning and the objection from disagreement involves a restriction of the theory of Ts to a sub-theory that involves just those criteria that are analytic as opposed to empirical. ("Analytic" in this context can be taken to mean both a priori and necessary.) The thought is that, quite generally, part of the theory ordinary speakers use in applying a term T will involve a priori and necessary criteria, where this entails, crucially, that these criteria are immune to empirical revision, and so should persist throughout any amount of learning. Likewise, disputants with differing total theories of a domain can still genuinely be in dispute over a common subject matter. For that subject matter can be defined by a common shared core of a priori and necessary criteria.

Even setting aside general philosophical skepticism about the analytic/synthetic distinction, the response seems inadequate. It faces a dilemma: are the analytic *criteria* for a term "T" themselves sufficient to fix the application

conditions of T, or not? The supposition that they typically are appears to have been empirically embarrassed; that is, the accumulated evidence against what Rosch called the classical view is precisely evidence to the effect that speakers in using terms do not know, and so do not employ, analytic criteria — a priori necessary and sufficient conditions — that fix the application conditions of those terms.

So at best the analytic criteria, in so far as speakers possess them and associate them with terms, will somewhat constrain, but not fix, the application conditions of terms. A number of difficulties now arise on this horn of the dilemma. First, in the case of many of our terms, there is reason to doubt that we actually possess *any* analytic criteria. Consider the biology major who comes to hold the consolidated theory about dogs:

> If x is an animal which has one of the characteristic doggish looks, smells, coat textures, etc. and is the offspring of that ilk then x is a dog. Dogs are members of the subspecies *Canis lupus familiaris*, distinct from the members of the wider species *Canis lupus* in the following ways: A, B, C.

None of this is analytic, i.e. utterly insulated from being overturned by surprising empirical discoveries. As Hilary Putnam made clear long ago, we could discover that the things we call "dogs" are not members of the subspecies *Canis lupus familiaris*, and indeed are not animals, but instead are cleverly disguised bionic robots collecting information about our domestic habits. Moreover, recall the coiffed squirrels that look like Chihuahuas. What actually is analytic for speakers when it comes to the term "dog"?

Second, consider the cases where we might suppose that some criterion possessed by speakers is at least a pretty good candidate to be "analytic". What are we then to make of subsequent learning of empirical matters of fact that intuitively work to restrict the extension of the term? How is that really possible? Suppose for example that for some user of the term "arthritis" the criterion "arthritis is a disorder" is an analytic criterion, and so one that constrains, but does not fully specify, the application conditions of his term "arthritis". Suppose he now learns the synthetic medical fact that arthritis is inflammation of the joints. How exactly *can* he have learned this, given the present theory of the application conditions of terms? Prior to the supposed learning, he associated one analytic criterion with arthritis, namely its being a disorder. So, on the hypothesis that only one's analytic criteria constrain the application conditions of one's terms, he is using a term which is no more specific in its application conditions than the term "disorder" or the phrase "some disorder or other". On the hypothesis that only the analytic criteria one associates with a term can constrain its application conditions, the most he can learn is that some disorder or other is inflammation of the joints. But clearly he is in a position to learn more than this, namely that *arthritis* is inflammation of the joints.

Have we crucially under-estimated what is analytic for our subject? Maybe it is something like "arthritis is the disorder that my speech community refers to by the term "arthritis"". Now perhaps we can explain, consistent with the hypothesis that only one's analytic criteria constrain the application conditions of one's terms, how our subject can learn that specifically *arthritis*, and not just some disorder or other, is inflammation of the joints. One familiar worry for this line of response is that many speakers who use terms do not have views about how their speech community uses those terms. That is not very telling, since a defender of the view that we have analytic criteria, and that only they constrain the application conditions, could say that rational speakers would come to recognize the truth that *arthritis is the disorder that my speech community refers to by the term "arthritis"*, just by understanding that truth.

A more telling objection is this: upon reflection it is obviously a synthetic or empirical matter of fact that one's speech community refers to arthritis by "arthritis". For just how far one's speech community extends is an empirical matter, and, crucially, it is an empirical matter just how good or bad the members of one's speech community are at spelling and pronunciation. It is manifestly not a priori that many of them are not severely impaired when it comes to spelling and speech production. They may, many or most of them, systematically fail to produce the right graphemes and phonemes, and so not refer to arthritis by "arthritis".

These points aside, we should say that our target is not the idea that we have *some* analytic or a priori criteria for applying *some* concepts. Our point is that the empirical literature makes this look like a special kind of case. This is especially so if the generic encoding hypothesis is true. Our earliest forms of generalization are generic in nature, so the little theories we use as criteria for applying terms are from the beginning generic in form. Even as adults, our thought is rife with generics, and we can relatively easily be gotten to evaluate universal generalizations as generics (Leslie et al., 2011). Thus it is very likely that our general criteria for the use of the concepts are generic in form. Here is the crucial observation in this context: the only generics that are plausibly taken to be a priori are those for which the corresponding universal generalization is a priori. These are the sorts of cases we have in mind:

> Yellows are brighter than browns.
>
> Primes are divisible by themselves.
>
> Irrational numbers are not expressible in the form m/n where m and n are rational.
>
> Conjunctions entail their separate conjuncts.

These are the special cases; and our point is that even in these special cases it remains a question whether the application conditions of the constituent concepts

can be *defined* by way of the backing universal generalizations. (For what it is worth, our best guess is "yes", for some simple mathematical and logical concepts, like *prime* and *conjunction*, and "no", for almost everything else.)

For all these reasons, we should reject any simple identification of the application conditions of the term "T" with the criteria — even the analytic criteria — derived from our best theory of Ts. The theory-theory of *criteria* is tenable; it is the favored psychological theory of concepts, and as we have seen, it meshes with the generic encoding hypothesis better than pure prototype theory; but the theory-theory of application conditions, at least in the simple form just considered, is not tenable.

The question remains as to whether there is some sophisticated tweaking of the theory-theory of application conditions that is tenable. Hence the relevance of the so-called Canberra plan, which we take to be just such a theory.

Lewis on Defining Terms

John Hawthorne and Huw Price coined the term "the Canberra plan" in 1996 to denote what was then a growing tendency among Australian philosophers, several of them then located in Canberra, to apply David Lewis's method of defining terms quite generally as a part of a revival of the program of philosophical analysis. Lewis' own analyses of the concepts of mind (1970), value (1989) and color (1997) are taken to be paradigms of the Canberra plan. Michael Tooley (1987) provided an analysis of causation which can be easily assimilated to the plan. Johnston's (1992) discussion of color is sometimes cited as an example of the plan, since it begins with central beliefs about color that we would find hard to give up, but unlike Lewis's (1997) somewhat similar discussion of color, the general apparatus of the plan is in fact absent. David Braddon-Mitchell and Robert Nola have recently edited an important collection of papers, *Conceptual Analysis & Philosophical Naturalism* (2009), a collection which is on the whole extremely sympathetic to the plan.

David Lewis is rightly viewed as the father of the Canberra plan, and his "Psychophysical and Theoretical Identifications" (1972) and "How to Define Theoretical Terms" (1970) are taken to be its founding documents, with an important addition arising from Lewis's emphasis in "New Work for a Theory of Universals" (1983) and "Putnam's Paradox" (1984) on natural properties as default referents. One way to see Lewis's proposal, and the Canberra plan more generally, is as a very sophisticated revision — one developed over fifteen years — of the general idea that the criteria derived from our theory of Ts are indeed the application conditions of the term "T".

Lewis makes three crucial revisions to the simple version of this idea, the version we have already found to be unworkable. The Canberra plan is the idea that, given these revisions and some associated qualifications, the criteria derived from our common theory of Ts are the application conditions of the term "T".

Precisely what revisions to the simple criteria/conditions identification are required, according to Lewis? First, instead of focusing on analytic criteria, which, even if they exist, are in fact very sparse on the ground and obviously insufficient to determine the application conditions of terms, Lewis resorts to what in various places he calls "platitudes". Thus in "Psychophysical and Theoretical Identifications" he writes

> Collect all the platitudes that you can think of regarding the causal relations of mental states, sensory stimuli, and motor responses . . . Add also all the platitudes to the effect that one mental state falls under another—"toothache is a kind of pain", and the like. Perhaps there are platitudes of other forms as well. Include only platitudes which are common knowledge among us — everyone knows them, everyone knows that everyone else knows then, and so on. (1972, p. 256)

Notice that the items of common knowledge considered individually need not be a priori; they can be highly informative. As an example, Lewis (1997) has the platitudes concerning *red* include such things as that British postboxes are red.[8] It is clear also that Lewis allows that platitudes can be false, for he immediately adds

> Form the conjunction of these platitudes; or better, form a cluster of them—a disjunction of all conjunctions of most of them. (That way it will not matter if a few are wrong.) This is the postulate of our term-introducing theory (1972, p. 256).

The platitudes embedding a term "T" used by our speech community make up the knowledge or supposed knowledge common to the members of our speech community. They are the relevant things 'that are known' in the sense that we take ourselves to know them, and we take them to be widely known, at least implicitly, throughout our speech community. Since these platitudes are common (supposed) knowledge, a certain kind of informed and reflective thinker can access them from the armchair, i.e., without any *further* empirical investigation.[9] Though some have tried to place further constraints on the Canberra plan, most notably that the platitudes be genuinely a priori or immune from empirical revision, Lewis's own more liberal account is more realistic, given the actual psychological literature. For each of our terms, we have a rich set of relevant things we take 'to be known'; but outside of logic and mathematics there is little that looks a priori.

Second, as the last quotation makes clear, in "Psychophysical and Theoretical identifications" Lewis adopts a cluster theory of application conditions. Instead of requiring that all of the platitudes taken together determine the application conditions of our terms, he assigns that work to a disjunction of conjunctions, each of which contains collections of most of the platitudes. As he notes,

this has the attractive consequence that the term "T" can have a reference even when many (though not most) of the platitudes involving it turn out to be false.

Third, in "New Work for a Theory of Universals" and "Putnam's Paradox" Lewis crucially departs from the idea that the reference of our terms is fixed solely by our total set of platitudes involving them. He recognizes that such an account, especially when it takes an appropriately clustered form, may well underdetermine the references of our terms in recognizable ways. Moreover, as Putnam (1980) pointed out, there will be intuitively wrong interpretations of "T" that are nonetheless compatible with making the "T"-involving platitudes true. Thus Lewis is led to endorse an external constraint on the reference of terms — one that turns on a notion that he finds indispensible for ontology, namely the notion of one property or class of individuals being more natural than another property or class of individuals. A property or class is perfectly natural when it is suited to figure in a fundamental ontology; when things share such a property they exhibit a respect of perfect similarity. The more natural a property or class, the more the sharing of a property, or membership in the class, makes for genuine similarity. For example, the property of being green makes for a genuine similarity among the things that have it, and so is more natural than the property of being grue, i.e. the property of being green and observed before the year 2000 or blue and observed after the year 2000, which does not make for a genuine similarity among the things that have it. Moreover, the class of electrons is more natural than the class of particles; the electrons collectively exhibit a higher degree of genuine similarity than does the wider class of particles.

Lewis's external constraint on reference fixing may now be put as follows; the more natural a property or class, the more it works as a default reference-attractor; that is, where there are two interpretations of "T" that do as well in making our "T"-involving platitudes true, or mostly true, we should prefer that interpretation which assigns "T" the more natural property or class. Notice that this way of putting it ranks the satisfaction of most of our "T"- involving platitudes lexically above the appeal to naturalness.

Alternatively, but still in the spirit of Lewis, one could adopt a principle which says that the preferred assignment of an extension to "T" is one that optimizes both the making true of platitudes and the degree of naturalness of the extension of "T". Depending on which way one goes, there will be a difference in what will count as the best analysis of a term or concept. For example, it is now a platitude, among philosophers at least, that justified true belief must somehow be 'de-Gettierized' in order to count as knowledge. If knowledge is a conjunctive property, then the identification of it with the tripartite conjunctive property — true and justified and believed — will hew to the naturalness constraint better than the identification of it with the four part conjunctive property true and justified and believed and de-Gettierized. Of course, the second identification will make one more platitude true, namely the very platitude that amounts to the standard philosophical intuition in Gettier cases. On the lexical ranking version of the naturalness constraint, this latter fact means that the second,

and philosophically favored, interpretation wins the day. On the other hand, the optimizing form of the naturalness constraint would treat this as a trade-off situation, and might deliver the result that our term "knowledge" can equally well be associated with either the three-part or the four-part conjunction. (Hereafter, the difference between the lexical and the optimizing forms of the naturalness constraint will not make much difference.)

General Empirical Problems with the Plan

By the Canberra plan we here specifically mean the idea that Lewis has thus outlined a quite general method of analysis. (Variants on this straightforward characterization will be considered as we go.) Lewis himself seemed more cautious than the planners so defined; after all, his own counterfactual analysis of causation does not go by way of the 'Ramsey/Lewis' method. In part, this was because Lewis (1973, 2000) believed that there are two very different forms of causation, causation by 'biff', roughly, by the transference of energy, and causation by absence, which are not in fact *platitudinously* identified as two forms of causation. (As it turns out, some friends of the Canberra plan such as David Liebesman (2011) have chided Lewis on just this point.)

If we follow the planners in viewing Lewis's account as a quite general method of analysis, we then have an account which (modulo considerations of naturalness) tightly connects the application conditions of a term to a certain privileged criterion that users of the term implicitly possess, namely the criterion which is determined by the disjunction of conjunctions of most of the platitudes involving the term. For Lewis, and for the planners more generally, "implicitly possess" now has a definite empirical content. Many or most of the users of the term will recognize the relevant platitudes as such upon appropriate reflection.

One immediate empirical problem with any specific application of the plan is deciding whether there are in fact enough platitudes on offer in order to fix the application conditions of the term in question. Being platitudinous within a speech community is not like being a priori or being analytic. Those latter properties are — at least according to theories which give them any significant role — properties a member of the speech community can in principle recognize claims to have simply by considering the claims themselves. In contrast, being platitudinous within a speech community is a substantial psycho-social property; I am only in a position to treat some claim as a platitude in this sense, and so incorporate it into a Lewis-style analysis, if I both take myself to know it and take it to be widely known, even if only implicitly, throughout my speech community. This latter condition is on its face an empirical belief about the claim's general uptake in my speech community, and there are reasons to think we have a tendency to overestimate how many of our deep convictions are obvious to others. The first worry then is this: for any given term there do seem to be more platitudes connected with it than there are genuine a priori or analytic claims

involving the term, so that a platitude analysis initially seems more promising as an account of what determines application conditions; however, it turns out to be an empirical question just which claims are platitudes, and there are reasons to believe we may often be mistaken about such questions. So we need to get out of the armchair and do some empirical social psychology in order to perform our "analysis" by platitudes. So the analysis is not in any real sense a priori. That point could be given an extra twist: since we need to get out of the armchair anyway, why not instead investigate the worldly phenomenon associated with the term? Is that not invariably of more philosophical interest than the investigation of the adventitious *social status* of claims?

Moreover, a claim's being platitudinous within a speech community can be a temporary property of that claim; many of the platitudes in which a term figures can suddenly be given up. Suppose investigators seem to discover that Neptune is not a planet but a massive space ship, a space ship that does not orbit the Sun, but simply occupies positions consistent with that, during just the periods when the space ship is visible from the Earth. The discovery gets widely put about. Should we then have any confidence that there are claims about Neptune which still are platitudinous in the relevant sense? (Recall that it need not be known by the members of my speech community that "Neptune" denotes Neptune; they need not be readers or writers to be speakers, and they may badly mispronounce the names of planets.) Does "Neptune" cease to have a reference in such conditions? It seems not. Reference seems a more persistent feature than the adventitious social status of claims.

A third problem arises from the residual truth in prototype theory, at least when it is restricted to what some philosophers have called "partly recognitional terms". These are terms like "red" or "dog", which we apply at least in part on the basis of paradigmatic or prototypical sensory and perceptual profiles, profiles whose linguistic characterization is a complex empirical psychological matter. We are not saying that an extensionally adequate linguistic characterization of the prototypical look of red or of the prototypical looks of dogs can never be given. The point is that even if we had such characterizations they would not be platitudinous in the defined sense. They would be complex bits of empirical psychological theorizing. The same is true of the very claim that the term "dog" is partly recognitional; philosophers can plausibly speculate that this is so, but it is not a platitude in the relevant sense. It is a bit of tentative, if plausible, psychological theorizing that goes beyond 'what is known' in our speech community. However, if the term "dog" is in fact partly recognitional then it is fairly likely that its extension is at least partly determined by our dispositions to treat certain looks as prototypical dog-looks. The upshot will then be that a crucial determiner of the extension of our term "dog" will not be incorporated into an analysis via platitudes. That method of analysis will thus be susceptible to being systematically thrown off for partly recognitional terms like "dog" or "red". (Notice that this point would hold a fortiori if we departed from Lewis,

as Michael Smith does, and restricted the platitudes to those claims that are ostensibly a priori. As observed in footnote 8, Smith agrees.)

Some may say that this just shows that philosophers have no business analyzing partly recognitional terms. However, many terms of peculiarly philosophical interest may well turn out to be partly recognitional. Color terms are an obvious example. The attributive adjectives "good" and "bad" may be partly recognitional. You need to have visual recognitional capacities to identify a good tennis swing as such. You need to have visual recognitional capacities to identify any badlooking thing as such, including morally-badlooking things such as the torturing of a cat. And as Albert Michotte's (1963) work reflects, there are a range of recognizable perceptual prototypes for causation, at least of the biff variety. Even six-month-old infants are sensitive to this profile, as Alan Leslie and Stephanie Keeble (1987) demonstrated.

Generics and the Plan

The Canberra plan, inspired by Lewis's account of the definition of terms, proposes a general method for the analysis of concepts: identify the term or terms which express the target concept, collect the platitudes involving the term that expresses the target concept, and then find the most natural satisfier of at least most of those platitudes. (If there is no satisfier of most of the platitudes, the term will then be empty.) Although the planners confine their attention to philosophically interesting concepts, there is, in fact, no such in-principle limitation built into the plan. If the plan works, then it should work quite generally. If it fails quite generally, then absent special pleading for the concepts of philosophical interest, we should expect it to fail in the philosophical cases as well.

One kind of special pleading we regard as very unpersuasive, given the psychological evidence, is that in philosophy our platitudes are invariably universals and not generics. That would be remarkable if true. "Events have causes", "true justified belief is knowledge", "lying is wrong", "people choose those acts that seem to them, given their beliefs, to advance the satisfaction of their desires" — these generics are platitudes, but the various attempts to find universal truths backing each of them has not resulted in anything that is likewise platitudinous.

Only if the plan works quite generally for concepts in whatever subject area we choose can it be seriously put forward as a method of conceptual analysis. *Concept* and *conceptual analysis* are topic neutral or quasi-formal notions that apply across all subject matters. The moral is that if the plan fails for many subject matters then whatever the plan is achieving in the philosophical cases it is not the analysis of concepts.

The plan does fail quite generally when interpreted as a philosophical analysis of concepts in any interesting sense of "concepts". First, let's set aside

an uninteresting sense. As noted earlier, a term of the form "concept of Ts" is often used to pick out contextually relevant part of someone's or some group's conception of Ts, as in "their concept of men is so old-school". A conception of Ts can be articulated or explicitly set out, but this is not analysis in any interesting sense, nor is it a distinctively philosophical enterprise, even when the term T is widely employed in philosophy. The articulation of conceptions is a matter for history, psychology and sociology; which is not to say that philosophers are to be prevented from pursuing such matters.

Suppose instead that we understand concepts as terms-in-use. If we want to know the extension of a concept understood as a term-in-use, we should consult the lexicographer, who has studied the relevant empirical questions about the extensions of our terms. Still, the plan might be put forward as a general account of what makes it the case that any given term in use has the (possible worlds) extension that it does, namely that the extension is determined by the most natural satisfier of the platitudes which involve the term. But now we face problems. Terms in use combine compositionally to give more complex terms in use. However, many of our platitudes are generic in form, and the structure of generics systematically interferes with compositionality. Furthermore, the requirement of naturalness and the structure of platitudinous generics can work together to assign manifestly incorrect extensions to terms.

The same points apply when we understand concepts by way of the substantial conception; that is, as items speakers grasp and which themselves determine (modulo considerations of naturalness) the extensions of our terms. Concepts, so understood, combine compositionally. Yet many of our platitudes are generic in form, and the structure of generics systematically interferes with compositionality. Moreover, the requirement of naturalness and the structure of platitudinous generics can work together to assign manifestly incorrect extensions to terms. Finally, when the plan relies on disjunctions of conjunctions of platitudes in order to insulate itself from the possibility of platitudes turning out to be false, it thereby breaks with another requirement on conceptual analysis, namely that good conceptual analyses should underwrite inferences like

Bonnie is a mare

to

Bonnie is a horse.

that is, the rare inferences where we *do* seem to possess (what looks like) an a priori justification for proceeding from the first belief to the second.

Generic Platitudes that Hold Only in the Minority of Cases

There can be generic platitudes that are true even though they hold of a minority of the very kind those generics concern. "Lying is wrong" holds true,

and may even be platitudinous, in a world in which people are quite honest and mostly lie *righty*, say to crazed axe-men at the door. Whenever we have generic platitudes that hold of a minority of cases, there is the likelihood that a platitude analysis will produce failures of compositionality. This is most easily seen in the case of sex-typed generics, but the point is quite general.

The possible world extension of the concept *female lion* should be a specific compositional function of the possible world extension of the concept *female* and the possible world extension of the concept *lion*. In particular, the concept *female lion* is intersective; that is, its extension in a possible world is the intersection of the extension of the concept *female* in that world with the extension of the concept *lion* in that world. (Contrast attributively qualified concepts like *small lion*, which though compositional are not intersective in this sense.) The Canberra plan applied to these three concepts *female*, *lion* and *female lion* will generate the required result that the concept *female lion* is intersective only if the platitudes concerning *female lion* are a union of the platitudes concerning *lion* and the platitudes concerning *female.*

Since being a platitude is a substantial psycho-social property there is no general guarantee that this is so. If it is not so, then there will be a failure to find the right kind of compositionality for the concept *female lion*. And the point will generalize for every intersective concept for which the platitudes do not align in the proper way. This kind of problem, if and when it arises, has always been taken to be a disabling objection to a philosophical theory of concepts. Recall that Fodor (1998) used this kind of objection against psychological theories; although we think it was misplaced there, the objection is disabling if it applies to a proposed theory of the application conditions of our concepts. We want concepts to combine compositionally, since the (possible worlds) extensions of our terms combine compositionally and concepts are, minimally, extension-determiners.

The generic encoding hypothesis suggests one way in which the right alignment of platitudes might not be in place. From a very young age, we are interested in what psychologists call "basic-level" kinds, kinds such as *dogs, lions, tigers, tables,* and *chairs*. Basic-level kinds can be contrasted with superordinate kinds (e.g., *mammals*, *furniture*), and subordinate kinds (e.g., *Bengal tigers*, *formal dining tables*). The first count nouns we learn usually denote basic-level kinds, and even into adulthood, they are the first terms we supply to answer the question, "what is this?" (e.g., if shown a picture of Princeton University's mascot, and asked what it is, one would naturally reply "a tiger", rather than "a mammal" or "a Bengal tiger"). Our knowledge concerning such kinds — knowledge passed on to us in childhood in generic form — is likewise encoded in as generic generalizations.

When taken collectively for each such kind, this mostly generic knowledge forms perhaps the best candidate to be the platitudes applicable to the kind. After all, we not only 'know' such things but we 'know' that our parents and teachers 'know' them. It would be very implausible to suggest that we only have a concept of a basic-level kind like the kind lion when and if we arrive at

exceptionless generalizations concerning it, and then come to know that these exceptionless generalizations are widely known in our speech community. For one thing, the exceptionless generalizations will not be widely known in our speech community.

So our platitudes for such basic-level kinds are, many of them, generic in form. Moreover, in the case of animal kinds, some of these generic platitudes involve properties that are only had by one sex. These include generics that describe methods of reproduction, such as "ducks lay eggs" and "horses give live birth", but also characteristic salient physical traits that happen to be had only by members of one sex, such as "lions have manes", "deer have antlers" or "peacocks have fabulous blue-green tails". As emphasized earlier, these generics are just not plausibly interpreted as contextually restricted universals respectively applying to female ducks and horses, or male lions, deer or peacocks. Nor, as noted above, are they encoded as sex-restricted generics such as "male lions have manes" (e.g., Cimpian, Gelman et al., 2010; Leslie, 2007, 2008, Khemlani et al., 2009). Rather, these are generics that are true despite there being a substantial number of exceptions to the generalization. These sex-typed generics are understood and accepted by preschool children; and plausibly are accepted as platitudes even at this age (Brandone et al., 2012). Surely, if anything is a platitude, *lions have manes* is a platitude.

We now have the following concrete situation. The platitudes for the concept *female lion* are not a union of the platitudes for the concept *female* and the platitudes for the concept *lion*. Even though it is a platitude that *lions have manes*, it is not a platitude that *female lions have manes*. Some of our fellow speakers of the language do not regard this as true because they know that female lions do not have manes; others do not regard it as true because they do not know whether it is the male lions or the female lions that have manes. The platitudes associated with an intersective concept such as *female lion* thus are not given by the union of the platitudes associated with *female* and the platitudes associated with *lion*. Note that if platitudes took universal rather than generic form, this particular issue would not arise, for if the only platitudes associated with *lion* were ones that were satisfied by *all lions*, and similarly for *female*, then the platitudes associated with *female lion* could consist of the union of these platitudes. However, it is simply not empirically plausible that our platitudes here take universal rather than generic form.

What if *only male lions have manes* was also a platitude? It is not clear how this could help. The point is simply that *lions have manes* is a platitude associated with *lions* but not with *female lions*, no matter what other platitudes may be also associated with the relevant concepts; therefore the platitudes associated with an intersective concept are not the union of the platitudes associated with both concepts. Further, it is much less plausible that *only male lions have manes* is a platitude, certainly it need not be. (In fact, a recent study found that some twenty percent of adult participants were ignorant of this fact, and a full third did not know that only male goats have horns and only female kangaroos have pouches

(Leslie et al. 2011).) We can't plausibly maintain that these adults who do not know that it is the male lions that have manes therefore do not have the concept lion, and hence do not have the concept female lion.[10]

What then can the friend of the Canberra plan say in order to defend the idea that it is a general method of conceptual analysis?[11]

Clustering to the Rescue?

There is one feature of the Canberra plan that might be thought to be helpful here. Lewis proposed that we 'cluster' our platitudes; that is, we should take the extension-determiner for a term not to be the totality of platitudes which govern it, since this would make the term empty if even one platitude was false, but rather to be a disjunction of conjunctions, each disjunct of which contains most of the platitudes, with each platitude appearing in at least one disjunct. So if there are three platitudes, "P1", "P2" and "P3", governing a term "T" we have in effect the claim that the possible worlds extension of "T" is fixed in the following way

A as it is in W satisfies "T" if and only if either

(i) the open sentence formed from "P1 and P2" by substituting the variable "X" for "T" throughout is true of A as it is in W, *or*
(ii) the open sentence formed from "P1 and P3" by substituting the variable "X" for "T" throughout is true of A as it is in W, *or*
(iii) the open sentence formed from "P2 and P3" by substituting the variable "X" for "T" throughout is true of A as it is in W.

Thus Lewis says that all his account requires is that the platitudes be mostly true of the target extension, which on the famous Lewisian account of adverbs (1975) amounts to the claim that all that is required is that most of the platitudes be true of the target extension. (We confess that we cannot make out anything helpful here in the alternative suggestion, sometimes heard, to the effect that *each* platitude should be true of most of the items in the extension. This fails for the same reason that accounts that assimilate generics to "most"-statements fail. For example, the platitude that lions have manes is not 'mostly true' of lions; most lions lack manes, since only *typically developing adult male* lions have manes.)

The relevance of Lewis's clustering proposal is this: clustering allows us to effectively shed or neutralize platitudes that turn out to be problematic, and thus avoid having them constrain the extensions of the relevant terms. Platitudes can be problematic if they are false, and this is what motivated the clustering account in the first place; for not every false platitude makes for an empty concept. Perhaps platitudes can also be problematic if they interfere with the compositional derivation of intersective concepts from their components. Only

one of the disjuncts —itself a conjunction — needs to hold for the disjunction to hold. So perhaps we can effectively neutralize the impact of the problematic platitude that lions have manes.

A moment's reflection will show that to resort to clustering here is to jump from the frying pan into the fire. First, we need not change our concept *female lion* by getting very interested in female lions, thereby coming to have more platitudinous knowledge of them than we do of lions in general. But when we do this, the disjunction of conjunctions of platitudes associated with female lions will change. Is this not a change in concept, according to the cluster view (just as it is a change in concept on the non-cluster view)? Certainly the extension determining cluster itself has changed.

Second, a single platitude or two may be all that is holding the possible world extensions of two distinct concepts apart, keeping them disjoint as it were. Such is the case, perhaps, with the concepts *Siberian tiger* and *Bengal tiger*, where the relevant true platitudes might be as follows

For Siberian Tigers

P4: Xs are predominantly located in or derive from the territory surrounding Siberia

P5 : Xs are striped

P6 : Xs are large cats

For Bengal Tigers

P7: Xs are predominantly located in or derive from the territory surrounding Bengal.

P8: Xs are striped

P9: Xs are large cats

The 'clustering' approach now produces chaos. Bengal tigers in any possible world will satisfy a disjunction of three distinct conjuncts each containing two of P4, P5, and P6, since they satisfy P5 and P6. Yet no Bengal tigers are Siberian tigers. Siberian tigers in any possible world will satisfy a disjunction of three distinct conjuncts each containing two of P7, P8 and P9, since they satisfy P8 and P9. Yet no Siberian tigers are Bengal tigers.

What if we add this last observation to the list of platitudes; what if we add *Siberian tigers are not Bengal tigers* to the lists? Even this will not help. It is, in effect, a numbers game; for suppose we also add one more platitude that Siberian and Bengal tigers 'agree' on, say "Xs are ferocious". Now there will be five platitudes associated with each concept, and so there will be a disjunct that contains only the three platitudes that both Siberian and Bengal tigers both

satisfy. This disjunct will suffice to give the two concepts the same extension, despite it being a *platitude* that they do not share the same extension.

As an elaboration of this point, let us return for a moment to the case of female lions. Suppose we were able to deal with the earlier difficulty about compositionality by insisting (implausibly) that platitudes must be universals, not generics. The set of platitudes associated with *female lion* could then be the union of the platitudes associated with *female* and with *lion*. But consider the possibility that there are *more* platitudes associated with *lion* than with *female.* Then there will be a disjunct that contains *only* platitudes associated with *lion*, and none with *female*. Clearly, male lions will satisfy this disjunct, and so we have the result that male lions fall in the extension of *female lion*.[12]

It is important to notice that this problem does not lie specifically with the idea that a given disjunct should contain *most* of the platitudes. Even if we raise the requirement to, say, 90% of the platitudes, we need only consider a context in which there are nine times as many platitudes associated with *lion* than with *female*. (This could be the result of a case of extreme female sequestration, across the entire animal kingdom.)

A final difficulty with clustering may now be noted. Suppose we persuade ourselves that there is an a priori and necessary connection between two concepts; given clustering, how can we be sure that the a priori and necessary status of this connection will be preserved? Perhaps a plausible case is the concept *mare* and the concept *horse*, so that it is a priori and necessary that if x is a mare then x is a horse. Intuitively, if anything is due to concepts, this is due to the concept *mare* and the concept *horse*, and so the relevant platitude "if x is a mare then x is a horse" should be part of the roster of platitudes for both concepts. But depending on how many platitudes there are for mares, something, say a perfect inorganic simulacrum of a mare, can fall under *most* of the platitudes on the roster for the concept mare, even though it is not a horse. How then can it be a priori and necessary that if x is a mare then x is a horse?

It should be noted that that many of these objections would still apply even if we limited the platitudes to prima facie a priori claims, as Smith (1994) suggests. Clustering would thus appear to be an inherently problematic strategy. But then how do we avoid the unwanted result that a single false platitude makes the target term empty?

Striking Property Generics and the Requirement of Naturalness

Earlier, we indicated that we wanted to consider some variants of Lewis's view that might be considered further 'tweaks' to which a particular planner might be attracted. Consider the following variant on Lewis's stated theory, one which does not in general associate application conditions with concepts, but only gives a criterion for deciding when one competing account is to be preferred to another: where we have two or more competitors we should prefer, other things

being equal, the account that vindicates *more* of the relevant platitudes. This criterion would generate a preference even in a case in which almost all of the relevant platitudes are false. In such cases, we believe Lewis would have said that there is no good account of the extension determiner of the concept. So it is a significant weakening of Lewis's actual view. Consider also the following significant strengthening of Lewis's view: the extension-determiner expressed by a term encompasses almost all of the platitudes associated with the term. Both the weakening and the strengthening of Lewis's view run into difficulties when we see that many of our platitudes are generic in form, and then come to understand the import of generic platitudes.

Recall that some generics, such as "sharks attack bathers," "manipulators are evil" and "mosquitoes carry West Nile virus", can be true even though very few members of the kind in question have the property.[13] Such cases arise if the property in question makes its bearers dangerous. One tempting first response is to suppose that, if a property has this feature, then a generic is true if just *some* members of the kind have the property in question. However this is too simple: consider "fish attack bathers" and "mammals maul children". Some great white, oceanic white-tipped, bull and tiger sharks attack bathers, and some pit-bulls maul children, yet the more general generics concerning fish and mammals do not seem right.

Leslie (2007, 2008, in press c) proposes that these sorts of generics are only true if the kind in question is a *good predictor* of the striking property, where a kind is a good predictor if its members are *typically disposed* to have the property in question, even if they do not manifest it. "Fish attack bathers" is false because trout and salmon and sea bass have no such disposition. However, we do not often have access to detailed information about unmanifested dispositions; so the question then arises, how do we in fact select a kind to be the locus of a 'dangerous-making property generic' in the absence of this sort of information? A plausible hypothesis is that, by default, we generalize the property to the basic-level kind. These kinds make up the psychologically privileged level of the subjective taxonomy, which includes kinds such as *sharks*, *tigers*, and *lions*. This pattern of generalization allows us, by default, to generalize these properties to kinds that can be easily and efficiently identified (see Leslie, 2008, in press c for more details). However, there is no guarantee that such kinds will not be overly inclusive; their members may not, in fact, be typically disposed to have the property in question. For example, consider "sharks attack bathers". It is plausible to suppose that this generic is, in fact, false, since arguably only great white, oceanic white-tip, bull and tiger sharks have this disposition. This is plausibly a case in which our default practice of generalizing to the basic-level kind leads us to over-generalize the property.

It would, however, seem to be a platitude that sharks attack bathers. It certainly satisfies the intuitive test, namely that it is natural to say, in an authoritative tone, "*it is widely known* that sharks attack bathers!" In fact, of course, this is one of those cases that Lewis deliberately allowed for, a case in

which a platitude, something supposedly widely known, is false. It is only the great white sharks, the oceanic white-tip sharks, the tiger sharks and the bull sharks that ever attack bathers; for example, basking sharks, whale sharks and megamouth sharks feed only on plankton. But this is not widely known, and so not a platitude. Suppose then, for purposes of argument, that great white sharks possess all the properties platitudinously predicated of sharks. After all they are sharks, and clustering aside, that is how you get to count as a shark on the Lewis view. Thus the great white sharks satisfy more of the platitudes on the roster for the concept shark, since they *do* attack bathers. So we now have an unintuitive result, namely that the extension of "sharks" is great white sharks. (By now it should be clear that "unintuitive" in such contexts does not mean: at odds with the proper analysis of the concept *shark*. It just means hard to believe, even given the circumstances imagined.)

Alternatively, suppose that there is one platitude on the roster of the concept *shark* that great white sharks do not satisfy. Then we indeed might have a tie, where both sharks and great white sharks each satisfy all but one of the platitudes on the roster of the concept *shark*, a different one in each case. Then the third distinctive feature of Lewis's account of how to define terms will kick in; when we have such a tie we should take the extension of the term to be the more natural set of things that satisfies the platitudes.

As we noted earlier, for Lewis a property or class is perfectly natural when it is suited to figure in a fundamental ontology; when things share such a property or are in such a class they exhibit at least one respect of perfect similarity. When we depart from the perfectly natural, the controlling rule is that the more natural a property or class, the more the sharing of a property or membership in the class makes for genuine similarity. So the class of electrons is more natural than the class of particles; the electrons collectively exhibit a higher degree of genuine similarity than does the wider class of fundamental particles. Likewise, the great white sharks exhibit a higher degree of genuine similarity than the sharks.

Thus, if we apply Lewis's criterion of naturalness to break the tie between sharks and great white sharks in the case where each satisfy all but one of the platitudes on the roster of the concept *shark*, we then get the unwanted result that the extension of "sharks" is the great white sharks. That unwanted result is the upshot of an unavoidable fact about platitudes, namely that just what the available platitudes are at a time is a contingent psycho-social matter, and a typical feature of naturalness, namely that a so-called "basic-level" kind like *shark* is less natural than its sub-kinds such as *megamouth shark* and *great white shark*.

It may even be possible to generate this sort of unwanted result simply from the facts about naturalness alone. Suppose we have a kind, namely *Panthera tigris*, more typically known as the kind tiger, and suppose that this kind satisfies *all* the platitudes governing the term "tiger". Depending on how the facts of genetics stand there may be a more natural sub-kind or sub-class of the kind tiger, which also satisfies all the platitudes in question. Here is a way of looking for it; identify some mutant tigers of a very distinctive sort, and consider the kind or class that

includes just the tigers that do not have that mutation. It will be very likely that this sub-kind or sub-class exhibits more genetic and phenotypic similarity among its members than does the kind tiger. To take a concrete example; consider the albinos, mutant tigers with a homozygous occurrence of a recessive gene that controls coat color, which has the effect of producing a stripe-free white coat. The albinos lie among the tigers, but thanks to the generic character of platitudes like "tigers have stripes" this does not prevent the kind tiger or the class of tigers from satisfying *all* the platitudes. But now consider the non-albino tigers, they may also satisfy all the platitudes, for the fact that some tigers are albinos is a relatively recherché fact, and so need not be a platitude. The non-albino tigers are not only more phenotypically similar among themselves than are the tigers; they are genetically more similar among themselves than the tigers. The non-albino tigers are, accordingly, the more natural class. Applying the Canberra plan, we get the unwanted result that our term "tiger" applies just to the non-albino tigers.[14]

We could go on in much more detail, but perhaps enough has been said to show that the Canberra plan, understood as a generalization or natural variant on Lewis on defining terms, is not really viable as a method of conceptual analysis.

There is another, much more deflationary way of construing the interesting work the advocates of the plan have done on color, mind and value; on this construal the work has been very useful, but it is not the implementation of a recognizable successor to conceptual analysis, let alone a newfangled method of conceptual analysis. It is just the humdrum, old-fashioned method of starting our inquiry with what we happen to take ourselves to know or firmly believe. That method is more or less compulsory anyway, and its credentials do not improve if it is further adorned with implausible aspirations to thereby provide an analysis, or explain the appearance of the a priori.

Conclusion

There are strong grounds for doubting that even when one takes into account all the heuristics and guiding theoretical commitments that we employ in applying a term "T" they will determine the conditions of application of the concept we express by "T". Indeed, there is good reason to doubt that when one takes all the heuristics that we employ in applying "T" into account, *and* all the facts about the relative naturalness of the extensions that could be captured by those, then all of this together will determine the conditions of application of the concept we express by "T". Just this emerged in the discussion of the Canberra plan.

Furthermore, once we assimilate the relevant psychology and philosophy, the following picture presents itself as plausible overview of the subject area that goes under the title of "concepts".

1. There are terms, both linguistic and mental. Terms *somehow* respectively get to be about individuals, classes, kinds, properties, etc. They *somehow* come to have these worldy items as their extensions. The unsolved mystery of reference or intentionality remains; we have shed no positive light upon it, but have only tried to rule out certain popular conceptions of how it works. Perhaps one crucial determiner of reference is the acceptance correction by others whom we regard as speaking our common tongue, so that we can then cull hopeless hypotheses about the reference of our common terms, and refine useful ones.
2. As a matter of psychological fact, subjects associate criteria with a term; these are heuristics, including perceptual prototypes, and guiding theoretical commitments, which put us in a good position to recognize many items in the extension of the term. These heuristics may be effective only in certain restricted situations, and the guiding theoretical commitments may be, even in large part, false.
3. Even when they are true, the guiding theoretical commitments are likely to be generic in form, allowing for many exceptions which are not counterexamples. (The generic encoding hypothesis.)
4. The psychologically real analog of Frege's notion of sense, namely the set of criteria a subject associates with a term, does not determine the reference of terms. It does not guarantee the compositionality of our concepts. It does not make for publicity, for different speakers using the same term with the same extension can employ different theories and different heuristics as guides.
5. The psychologically real analog of sense may serve to explain how it is that terms with the same possible word extensions can have different cognitive values for a given speaker. It can be informative to learn that vixens are female foxes because it unites the criteria we use for "vixen" with those we use for "female foxes".
6. The method of cases, the method of trying to articulate what we know in using a term by considering how we are inclined to apply it to real and merely possible cases, is at best a way of articulating our criteria. It is not a method for analyzing concepts, as is shown by the fact that our reaction to cases is typically explained by our criteria, and not by some supposed grasp of the application conditions of our concepts.
7. There is no empirically explanatory need to postulate concepts, understood as abstract entities (i) which encode the conditions of application of our terms and (ii) grasp of which guides in the use of our terms. Talk of grasping a concept is either an unhelpful metaphor or a description of a supernatural theory of language use, one rendered otiose by the actual empirical psychology of language use.
8. There is accordingly no good reason to think that there are concepts in any substantial philosophical sense. There are terms, public languages,

the extensions of terms, and criteria for applying terms. There is the something or other that gives terms their particular extensions, and hence their conditions of application. But this something or other is never a concept in the substantial philosophical sense. Concepts in the substantial sense have no demonstrated use.

Notes

1. Special thanks to Shamik Dasgupta, Frank Jackson, Joshua O'Rourke, Nathaniel Tabris, and Gideon Rosen for helpful comments on an earlier draft.
2. Preparation of this article was supported by NSF grant BCS-1226942, awarded to Leslie.
3. 1967, p 22.
4. See, for example, Ramsey (1992), Stich and Weinberg (2001), Laurence and Margolis (2003), Sytsma (2010), and Banicki (2012).
5. It is interesting here to consider the findings of Armstrong, Gleitman, and Gleitman (1983), who found some of the typicality effects even for concepts such as *odd number*. This shows, of course, that typicality effects are *compatible* with representing necessary and sufficient conditions. However, the crucial point is that in the case of concepts such as *odd number* we have independent reason for supposing that we represent necessary and sufficient conditions — namely that we (i.e., typical competent adult users of the term "odd number") can articulate what they are. There is no corresponding case to be made for the majority of concepts. Thus while the in-principle compatibility point concerning typicality effects and necessary and sufficient conditions is illustrated by Armstrong et al.'s findings, it does not address the fact that, for most concepts, we simply have no reason to suppose that typical, competent adults represent and exploit necessary and sufficient conditions.
6. Sometimes "theory-theory" is reserved for the specific view that our concepts, including young children's concepts, are *exactly like* scientific theories. We follow many in the field by using the term more inclusively.
7. It is, perhaps, rather telling that psychological experiments concerned with concepts almost invariably use the generic form to articulate conceptual knowledge.
8. Some friends of the Canberra plan, such as Michael Smith (1994), reject this liberal attitude to platitudes. For Smith, platitudes must be "prima facie a priori", i.e. such that if they are true then their truth is not an empirical matter. However this means that the plan will often fail, since, as we have seen, in general there are just not enough prima facie a priori truths around to determine the application conditions of our terms. Indeed, as in the case of space not in fact being Euclidean makes clear, empirical inquiry is itself highly relevant to the question of whether it is the case that if S is true then its truth is an a priori or instead an empirical matter. We often simply cannot imagine what would undermine our central beliefs. So, the category of what we can legitimately take to be prima facie a priori will shrink as we become more and more aware of the inter-animation of the putatively a priori and the empirical. In fact, Smith's own conclusion is that the plan does fail in both of the cases he is concerned with, namely the analysis

of "red" and in the analysis of "right". And he provides a plausible diagnosis of this failure; in such cases competence with the term consists in part in a disposition to use it to classify cases correctly on the basis of certain patterns of features. There is no reason to believe that there is any prima facie a priori encapsulation of the relevant patterns.

9. Although Lewis drops the term "platitudes" in "Reduction of Mind" (1994) he crucially still talks in terms of putative common knowledge. He writes "We have a very extensive shared understanding of how we work mentally. Think of it as a theory: FOLK PSYCHOLOGY. It is common knowledge among us; but it is tacit, as our grammatical knowledge is. We can tell which particular predictions and explanations conform to its principles, but we cannot expound those principles systematically" (p. 56). Our use of "platitudes" throughout is meant to conform simply to this idea that we can articulate the relevant putative common knowledge in which our terms are entangled, and then use it to define our terms. In fact, Lewis (1997) seems to persist with a platitude-based analysis of color terms.
10. We have argued that the platitudes associated with intersective concepts do not always include the union of the platitudes associated with the component concepts. A related point that Jerry Fodor (e.g., 1998) has often made is that the platitudes associated with the complex concept will also not usually be *exhausted* by this union of platitudes. That is, it could be a platitude that *black bulls are dangerous* without it is being a platitude that *black things are dangerous* or its being a platitude that *bulls are dangerous*.
11. One response to failures of compositionality due to the widespread application of one's favored technique of analysis is instead to identify a base of uncompounded terms, apply one's favored technique of analysis only to them and then simply let the syntactical methods of combination by which the language generates compound terms determine the extensions of compound terms. This, however, would be an unhappy way to patch up the failures of compositionality due to widespread application of the Canberra plan. First, it is likely that many terms in any plausible base will be partly recognitional terms like "red" and "dog" and "cause", for which the platitude-based analysis is ill-suited because of the reasons noted earlier. Second, many of the target terms of philosophical analysis are unlike "physical cause" and more like "free will" in being themselves complex though not simply intersective combinations (or any other simple combination) of their respective components (here "free" and "will").
12. Note that it will not help to insist, implausibly, that each concept have the same number of platitudes associated with it. For consider then the complex concept *adult female lion*. By assumption there is the same number of platitudes associated with all three concepts, so each concept contributes one third of the total number of platitudes. But then again there will be a disjunction that contains only platitudes associated with *adult* and *lion,* and none with *female.* Adult male lions will again satisfy this disjunction.
13. Far and away, most cases of manipulation involve parents and children; you get to count as a manipulator by manipulating a bit now and then, but the overwhelming majority of parents and children who manipulate a bit now and then are not evil. The same holds for lawyers, politicians and the police.

14. For an analog of this sort of point in the case of Lewis's analysis of the concept of personal identity, see Johnston (2010), pp. 79–80 "On the Unhelpfulness of Reference Magnetism".

References

Ahn, W., Kim, N. S., Lassaline, M. E., & Dennis, M. J. (2000). "Causal Status As A Determinant of Feature Centrality." *Cognitive Psychology*, 41, 361–416.

Armstrong, S.L., Gleitman, L.R., & Gleitman, H. (1983). "What Some Concepts Might Not Be." *Cognition*, 13, 263–308.

Banicki, K. (2012). "Connective Conceptual Analysis and Psychology." *Theory and Psychology*, 22 (3), 310–323.

Barner, D., Chow, K., & Yang, S. (2009). "Finding One's Meaning: A Test of the Relation Between Quantifiers and Integers in Language Development."*Cognitive Psychology*, 58, 195–219.

Bealer, G. (1987). "The Philosophical Limits of Scientific Essentialism." *Philosophical Perspectives*, 1, 289–365.

Boghossian, P. (2003). "Blind Reasoning." *Proceedings of the Aristotelian Society, Supplementary volume*, 77, 225–248.

Braddon-Mitchell, D., & Nola, R. (2009). *Conceptual Analysis and Philosophical Naturalism*. Cambridge, MA: MIT Press.

Brandone, A., Cimpian, A., Leslie, S. J., & Gelman, S. A. (2012). "Do Lions Have Manes? For Children, Generics are About Kinds, Not Quantities." *Child Development*, 83(2), 423–433.

Brandone, A., Gelman, S. A., & Hedglen, J. (submitted). "Young Children's Intuitions about the Truth Conditions and Implications of Novel Generics."

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

Carey, S. (2009). *The Origin of Concepts*. New York: Oxford University Press.

Carlson, G.N., & Pelletier, F.J. (1995). *The Generic Book*. Chicago: Chicago University Press.

Chalmers, D. (1996). *The Conscious Mind*. New York: Oxford University Press.

Chalmers, D., & Jackson, F. (2001). "Conceptual Analysis and Reductive Explanation." *The Philosophical Review*, 110, 315–361.

Cimpian, A., Brandone, A. C., & Gelman, S. A. (2010). "Generic Statements Require Little Evidence for Acceptance But Have Powerful Implications." *Cognitive Science*, 34(8), 1452–1482.

Cimpian, A., Gelman, S. A., & Brandone, A. C. (2010). "Theory-Based Considerations Influence The Interpretation of Generic Sentences." *Language and Cognitive Processes*, 25(2), 261–276.

Cohen, A. (1996). *Think Generic: The Meaning and Use of Generic Sentences*. Ph.D. dissertation, Carnegie Mellon University.

Dahl, O. (1985). *Tense and Aspect Systems*. Oxford: Blackwell.

Davidson, D. (1967). "Truth and Meaning." *Synthese*, 17, 304–323.

Devitt, M. (2005). "There is No A Priori." In E. Sosa and M. Steup (eds.), *Contemporary Debates in Epistemology*, pp. 105–115. Cambridge, MA: Blackwell Publishers.

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

Fodor, J. A., Garrett, M. F., Walker, E., & Parkes, C. (1980). "Against definitions." *Cognition*, 8, 263–367.

Fodor, J. A., & Lepore, E. (2002). *The Compositionality Papers*. New York: Oxford University Press.
Frege, G. (1956). "The Thought: A Logical Inquiry." *Mind*, 65, 289–311
Geach, P.T. (1967). "Identity." *Review of Metaphysics*, 21, 3–12.
Gelman, S. A. (2003). *The Essential Child: Origins of Essentialism in Everyday Thought*. New York: Oxford University Press.
Gelman, S. A. (2010). "Generics as A Window onto Young Children's Concepts." In F. J. Pelletier (ed.), *Kinds, Things, and Stuff: The Cognitive Side of Generics and Mass Terms* (New Directions in Cognitive Science v. 12.), pp. 100–123. New York: Oxford University Press.
Gelman, S.A. & Raman, L. (2003). "Preschool Children use Linguistic form Class and Pragmatic Cues to Interpret Generics." *Child Development*, 74, 308–325.
Gelman, S. A., Star, J., & Flukes, J. (2002). "Children's Use of Generics in Inductive Inferences." *Journal of Cognition and Development*, 3, 179–199.
Gopnik, A., & Meltzoff, A.N. (1997). *Words, thoughts, and theories*. Cambridge, MA: MIT Press.
Graham, S. A., Nayer, S. L., & Gelman, S. A. (2011). "Two-Year-Olds use the Generic/Non-Generic Distinction to Guide Their Inferences About Novel Kinds." *Child Development*, 82, 493–507.
Harman, G. (1974). "Meaning and semantics." In M. K. Munitz and P. K. Unger (eds.), *Semantics and Philosophy*, pp. 1–16. New York: NYU Press.
Harman, G. (1975). "Language, Thought, and Communication." In K. Gunderson (ed.), *Language, Mind, and Knowledge: Minnesota Studies in the Philosophy of Science* VII, pp. 279–98. Minneapolis, MN: University of Minnesota Press.
Hawthorne, J., & Price, H. (1996). "How to Stand Up For Non-Cognitivists." *Australasian Journal of Philosophy*, 74 (2), 275–292.
Hollander, M. A., Gelman, S. A., & Star, J. (2002). "Children's Interpretation of Generic Noun Phrases." *Developmental Psychology*, 38, 883–894.
Jackson, F. (1994). "Armchair Metaphysics." In M. Michael and J. O'Leary-Hawthorne (eds.), *Meaning in Mind*, pp. 23–42. Boston: Kluwer Academic Publishers.
Jackson, F. (1998). *From Ethics to Metaphysics: A Defense of Conceptual Analysis*. Oxford: Clarendon Press.
Johnston, M. (1992). "How to Speak of the Colors." *Philosophical Studies*, 68(3), 221–263.
Johnston, M. (2010). *Surviving Death*. Princeton, NJ: Princeton University Press.
Jönsson, M. L. & Hampton, J. A. (2006). "The Inverse Conjunction Fallacy." Journal of Memory and Language, 55, 317–334.
Keil, F. (1989). *Concepts, Kinds, and Cognitive Development*. Cambridge, MA: MIT Press.
Khemlani, S., Leslie, S. J., & Glucksberg, S. (2009). "Generics, Prevalence, and Default Inferences." *Proceedings of the 31st Annual Cognitive Science Society*. Amsterdam: Cognitive Science Society.
Kripke, S. (1980). *Naming and Necessity*. Cambridge MA: Harvard University Press.
Kripke, S. (1982). *Wittgenstein on Rules and Private Language*. Cambridge, MA: Harvard University Press.
Laurence, S., & Margolis, E. (2003). "Concepts and Conceptual Analysis." *Philosophy and Phenomenological Research*, 67 (2), 253–282.
Leslie, A.M., & Keeble, S. (1987). "Do Six-Month-Old Infants Perceive Causality?" *Cognition*, 25, 265–288.
Leslie, S. J. (2007). "Generics and the Structure of the Mind." *Philosophical Perspectives*, 21, 375–403.
Leslie, S. J. (2008). "Generics: Cognition and Acquisition." *Philosophical Review*, 117(1), 1–47.
Leslie, S. J. (in press a). "Generics Articulate Default Generalizations." *Recherches Linguistiques de Vincennes*.

Leslie, S. J. (in press b). "Essence and Natural Kinds: When Science Meets Preschooler Intuition." In T. Gendler and J. Hawthorne (eds.), *Oxford Studies in Epistemology*, 4.

Leslie, S. J. (in press c). "The Original Sin of Cognition: Fear, Prejudice, and Generalization." *The Journal of Philosophy*.

Leslie, S. J., & Gelman, S. A. (2012). "Quantified Statements are Recalled as Generics: Evidence from Preschool Children and Adults." *Cognitive Psychology*, 64, 186–214.

Leslie, S. J., Khemlani, S., & Glucksberg, S. (2011). "All Ducks Lay Eggs: The Generic Overgeneralization Effect." *Journal of Memory and Language*, 65, 15–31.

Lewis, D. (1970). "How to Define Theoretical Terms." *Reprinted in his Philosophical Papers*, pp. 78–95, vol. 1, (1983), New York: Oxford University Press.

Lewis, D. (1972). "Psychophysical and Theoretical Identifications." *Australasian Journal of Philosophy*, 50, 249–259.

Lewis, D. (1973). "Causation." *Journal of Philosophy*, 70, 556–567.

Lewis, D. (1975). "Adverbs of Quantification." In E. L. Keenan (ed.), *Formal Semantics of Natural Language*, pp. 3–15. Cambridge: Cambridge University Press.

Lewis, D. (1983). "New Work for a Theory of Universals." *Australasian Journal of Philosophy*, 61, 343–377.

Lewis, D. (1984). "Putnam's Paradox." *Australasian Journal of Philosophy*, 62, 221–236.

Lewis, D. (1989). "Dispositional Theories of Value." *Proceedings of the Aristotelian Society*, Supplementary volume 63, 113–137.

Lewis, D. (1994). "Reduction of Mind." In S. Guttenplan (ed.) *A Companion to the Philosophy of Mind*, pp. 412–31. Oxford: Basil Blackwell.

Lewis, D. (1997). "Naming the Colours." *Australasian Journal of Philosophy*, 75(3), 325–342.

Lewis, D. (2000). "Causation as Influence", *Journal of Philosophy*, 97, 209–212.

Liebesman, D. (2011). "Causation and the Canberra Plan". *Pacific Philosophical Quarterly*, 92(2), 232–242.

Machery, E. (2009). *Doing Without Concepts*. New York: Oxford University Press.

Mannheim, B., Gelman, S. A., Escalante, C., Huayhua, M., & Puma, R. (2011). "A developmental analysis of generic nouns in Southern Peruvian Quechua." *Language Learning and Development*, 7(1), 1–23.

Margolis, E., & Laurence, S. (1999). *Concepts: Core Readings.* Cambridge, MA: MIT Press.

Meyer, M., Gelman, S. A., & Stilwell, S. M. (2011). "Generics are a Cognitive Default: Evidence from Sentence Processing." In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Boston, MA: Cognitive Science Society.

Michotte, A. (1963). *The Perception of Causality*. Andover: Methuen.

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

Peacocke, C. (1992). *A Study of Concepts*. Cambridge, MA: MIT Press.

Peacocke, C. (1993). "How Are A Priori Truths Possible?" *European Journal of Philosophy,* 1, 175–99.

Peacocke, C. (1997). "Metaphysical Necessity: Understanding, Truth and Epistemology." *Mind,* 106, 521–74.

Peacocke, C. (1998). "Implicit Conceptions, Understanding and Rationality." in E. Villanueva (ed.), *Philosophical Issues* 9: *Concepts*, pp. 121–48.

Peacocke, C. (2005). "The A Priori." In F. Jackson and M. Smith (eds.), *The Oxford Handbook of Contemporary Philosophy*, pp. 739–767. Oxford: Oxford University Press.

Pelletier, F. J. & Asher, N. (1997). "Generics and Defaults." In J. van Benthem and A. ter Meulen (eds.), *Handbook of Logic and Language*, pp. 1125–1179. Cambridge, MA: MIT Press.

Prasada, S., Khemlani, S., Leslie, S. J., & Glucksberg, S. (in press). "Conceptual Distinctions Amongst Generics." *Cognition*.

Putnam, H. (1962). "It Ain't Necessarily So." *The Journal of Philosophy*, 59, 658–671.

Putnam, H. (1973). "Explanation and Reference." Reprinted in his *Mind, Language and Reality: Philosophical Papers*, vol. 2, (1975), pp. 196–215. New York: Cambridge University Press.

Putnam, H. (1975). "The Meaning of 'Meaning'." Reprinted in his *Mind, Language and Reality: Philosophical Papers*, vol. 2, (1975), pp. 215–272. New York: Cambridge University Press.

Putnam, H. (1980). "Models and Reality," *Journal of Symbolic Logic*, 45(3), 464–482.

Quine, W.V.O. (1951). "Two Dogmas of Empiricism." Reprinted in his *From a Logical Point of View*, (1953), pp. 20–47. Cambridge MA: Harvard University Press.

Ramsey, W. (1992). "Prototypes and Conceptual Analysis." *Topoi*, 11, 59–70.

Rosch, E. (1973). "Natural Categories." *Cognitive Psychology*, 4, 328–350.

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

Rosch, E. & Mervis, C.B. (1975). "Family Resemblances: Studies in the Internal Structure of Categories." *Cognitive Psychology*, 7(4), 573–605.

Schroeter, L. (2004)."The Limits of Conceptual Analysis." *Pacific Philosophical Quarterly*, 85, 425–453.

Sloman, S. A. (1993). "Feature-based induction." *Cognitive Psychology*, 25, 231–280.

Sloman, S. A. (1998). "Categorical inference is not a tree: The myth of inheritance hierarchies." *Cognitive Psychology*, 35, 1–33.

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

Smith, M. (1994). *The Moral Problem*. Oxford: Wiley-Blackwell.

Stich, S.P., & Weinberg, J.M. (2001). "Jackson's Empirical Assumptions." *Philosophy and Phenomenological Research*, 62 (3), 637–643.

Sytsma, J. (2010). "The Proper Province of Philosophy." *Review of Philosophy and Psychology*, 1(3), 427–445.

Tardif, T., Gelman, S. A., Fu, X., & Zhu, L. (2011). "Acquisition of Generic Noun Phrases in Chinese: Learning about Lions without An "-S"." *Journal of Child Language*, 30, 1–32.

Tooley, M. (1987). *Causation: A Realist Approach*. New York: Oxford University Press.

Williamson, T. (2003). "Understanding and Inference." *Proceedings of the Aristotelian Society*, 77, 249–293.