Meaning postulates and deference[*]

RICHARD HORSEY

Abstract

Fodor (1998) argues that most lexical concepts have no internal structure. He rejects what he calls Inferential Role Semantics (IRS), the view that primitive concepts are constituted by their inferential relations, on the grounds that this violates the compositionality constraint and leads to an unacceptable form of holism. In rejecting IRS, Fodor must also reject meaning postulates. I argue, contra Fodor, that meaning postulates must be retained, but that when suitably constrained they are not susceptible to his arguments against IRS. This has important implications for the view that certain of our concepts are deferential. A consequence of the arguments I present is that deference is relegated to a relatively minor role in what Sperber (1997) refers to as reflective concepts; deference has no important role to play in the vast majority of our intuitive concepts.

1 Introduction

In his 1998 book *Concepts: Where Cognitive Science Went Wrong*, Jerry Fodor makes a plausible case for conceptual atomism. On this view, nearly all lexical concepts (that is, concepts expressed by monomorphemic lexical items) have no internal structure. He rejects alternative views—that concepts are definitions or prototypes or 'theories'—largely for the following reason: they hold that concepts are constituted at least in part by their inferential relations. He calls any theory of this kind an Inferential Role Semantics (IRS). Fodor has a general argument which he takes to indicate that inferential relations cannot be content constitutive, and hence that any version of IRS is untenable. Basically, his argument is that IRS is incompatible with a fundamental aspect of thought – its compositionality. He also argues that IRS leads to an unacceptable form of holism.

In rejecting content-constitutive inferential relations, Fodor is led to reject meaning postulates, which he himself proposed as a way to capture intuitions of analyticity in the face of strong arguments against definitional accounts. In addition to his worries about

[*] I would like to thank Deirdre Wilson for many useful discussions on the substance of this paper, and for her very helpful comments on several earlier drafts.

holism, Fodor considers that given Quine's strictures on analyticity no principled distinction can be drawn between meaning postulates and encyclopaedic knowledge.

I argue that Fodor is wrong to reject meaning postulates. First, I show that when suitably constrained, meaning postulates do not lead to meaning holism. Second, by drawing on evidence that there are different kinds of concepts processed by distinct inferential modules, I demonstrate that there is a clear theoretical and empirical basis for drawing a distinction between meaning postulates and encyclopaedic knowledge. Furthermore, I show how meaning postulates could plausibly arise out of fundamental properties of our cognitive architecture: if, as is widely believed, we are psychological essentialists with respect to particular classes of concepts, then certain meaning postulates are nomologically necessary. These arguments will prove to have important implications for the view, widely held since Putnam (1975) and Burge (1979), that certain of our concepts are deferential in nature.

Let us start by looking at Fodor's arguments against IRS.

2 Inferential Role Semantics

Fodor claims to have a general argument against all forms of IRS, which runs approximately as follows: once you identify the content of a concept with its inferential role, you have to make a distinction between those inferences which are content-constitutive and those which are not. The only way to avoid having to make this distinction would be to associate content with inferential role *tout court*. But Fodor notes that this option immediately faces the following problem.

Thought, like language, is compositional. So the content of the concept BROWN COW is composed out of the contents of the concepts BROWN and COW, together with their mode of syntactic combination. Inferential roles, however, are not in general compositional. Suppose I believe that brown cows are dangerous, so that I am predisposed to make the inference BROWN COW → DANGEROUS. This forms part of the inferential role of BROWN COW, but is not part of the inferential role of either BROWN or COW (assuming I do not believe either that brown things in general are dangerous, or that cows in general are dangerous). The conclusion which Fodor draws is that content cannot be associated with inferential role *tout court*.

Associating content with inferential role *tout court* is, of course, a species of holism, and Fodor has independent worries about holism which he has spelled out at length (Fodor & Lepore 1992). In the present context, the worry is this. Concepts are "the sorts of things that lots of people can, and do, *share*" (Fodor 1998: 28 ff, original emphasis), and Fodor therefore proposes a publicity constraint to this effect on theories of concepts.

But there is good reason to think that holism leads to incommensurability, and hence that associating content with inferential role *tout court* violates the publicity constraint.

The only other option for IRS, then, is to associate the content of a concept with some proper subset of its inferential role. But on what basis can this subset be distinguished? Traditionally, practitioners of IRS have tended to fall back on some notion of analyticity, and claimed that (all and only) analytic inferences are content constitutive.[1] The very notion of analyticity has been challenged by philosophers, however, and because of these worries Fodor rejects this possibility too.

If these arguments are correct, it follows that no version of IRS is tenable: content cannot be associated with inferential role *tout court*, but neither can it be associated with any proper subset of the inferential role. Let us, then, consider the status of analyticity more closely.

3 Analyticity

It has been noted by philosophers since at least Kant that statements such as "all bachelors are male" are very different from statements such as "all bachelors are untidy"—statements of the former kind are true or false in virtue of logic and the meaning of the parts alone (*analytic*, as one says); statements of the latter kind are true or false in virtue of empirical facts (*synthetic*, as one says). This is the essence of the analytic–synthetic distinction.

Things have never quite been the same, however, since Quine published his classic paper *Two Dogmas of Empiricism* (Quine 1953), in which he argued that no principled analytic–synthetic distinction could be made. The majority of philosophers seem to have been convinced by Quine's arguments, and while the *intuition* of analyticity remains, it has generally been abandoned as a theoretical notion.

Quine's arguments were based on his conception of our beliefs as forming a web of logical interconnections. Any statement, he claimed, could be made unrevisable provided drastic enough adjustments were made elsewhere (e.g. appeal to hallucinations to explain recalcitrant empirical data); conversely, no statement is immune from revision

[1] Boghossian (1993) points out that IRS *cannot* merely fall back on analyticity to distinguish those inferences which are constitutive of the meaning of a concept, since this is plainly circular: analyticity is itself explicated in terms of meaning. Rather, IRS would have to employ a notion of analyticity *reductively construed* (i.e. explicated in terms of some other property necessarily equivalent to analyticity).

(e.g. it has been proposed that the logical law of excluded middle[2] should be abandoned in order to simplify quantum mechanics).

So if we accept Quine's arguments, and reject the analytic–synthetic distinction, we must also give up content-constitutive inferences. But is it *possible* to give up *all* such inferences? It would seem, first blush at least, as if we need to hold on to at least some of them. Consider, for example, Chomsky's (1988: 31–34) example of 'persuade'. If I persuade John to go to university, I must cause him to decide to go to university; if John at no point decides to go to university, then I have not persuaded him. But note that we cannot merely define 'persuade' as 'cause to decide' since I could cause John to decide to go to university without thereby persuading him (for example, if I caused him to decide through force, or by accident). It is, however, *surely* part of the meaning of 'persuade' that '*x* persuade *y* to *z*' → '*y* decide to *z*'. Similarly, it is *surely* part of the meaning of 'kill' that '*x* kill *y*' → '*y* die'.[3] Traditionally, one-way entailment rules known as meaning postulates have been used to capture such semantic relations.

Furthermore, almost everyone (Fodor included) agrees that we need meaning postulates to capture the meaning of logical words such as 'and'. It would seem that the meaning of logical words such as 'and' *just is* their inferential role. Since this is all there is to the meaning of these items, we would not seem to be able to give up meaning-constitutive inferences in these cases.

4 Meaning postulates

Fodor first proposed meaning postulates[4] as a way of capturing intuitions of analyticity about certain meaning relations in the face of strong evidence that definitions had little part to play in the meaning of lexical concepts.

For example, if RED → COLOUR is taken to be analytic, then on a definitional account RED must decompose into COLOUR + X. But what could possibly stand in for X? Nothing, it seems, could have the meaning RED BUT NOT COLOURED. Fodor's idea was

[2] The law of excluded middle states that for any proposition *p*, it is logically necessarily that $p \vee \sim p$.

[3] Of course, the explanation for this is *conceptual*, not *linguistic*; 'kill' and 'persuade' in other languages behave similarly. The correct generalisation, then, is '*x* KILL *y* → *y* DIE' (following convention, concepts are referred to using small capitals).

[4] Fodor, Fodor & Garrett (1975: 519), although the term itself comes from Carnap (1947), who used it in a somewhat different sense.

that RED could be viewed as non-decomposable, and that the analyticity could be captured by the stipulation of a one-way entailment (or "meaning postulate") RED → COLOUR.

In more recent work, however, Fodor has been uncomfortable with the notion of a meaning postulate. For one thing, as we have seen, he takes Quine's arguments regarding analyticity very seriously, believing that a principled analytic–synthetic distinction cannot be maintained. In the absence of such a distinction, Fodor believes that meaning postulates and encyclopaedic knowledge cannot be distinguished, and hence that meaning postulates must be rejected as a theoretical device. The same arguments that Fodor used against IRS also hold against meaning postulates, so if Fodor rejects IRS he must also reject meaning postulates.

The problem is this. Back in the days when life was simple, and philosophers could be counted on to tell which meaning relations were analytic, the distinction between meaning postulates and empirical (i.e. encyclopaedic) knowledge presented no problem. But if the analytic–synthetic distinction cannot be maintained, then how does one decide *which* meaning relations are to be captured by meaning postulates and which by empirical knowledge? Fodor (1998: 111 f) argues as follows. Consider two minds which differ only in that WHALE → MAMMAL is a meaning postulate for one, but 'general knowledge' for the other (an example from Partee 1995). Surely, says Fodor, there are no further differences between these minds entailed.

So Fodor is forced to say that since no principled distinction can be made between meaning postulates and encyclopaedic knowledge, then entailments of the sort traditionally captured by meaning postulates must be seen as encyclopaedic knowledge. While Fodor cannot deny that 'kill → cause to die' is necessary, he denies that it is *semantic*; instead, he claims that it is a fact about killing (Fodor, p.c.).

But even if we were to accept this kind of story about 'kill', we have not completely eliminated meaning postulates.[5] We are still left with the logical vocabulary. At first, this might not appear to be too much of a problem. Fodor rejects meaning postulates, remember, because he thinks no principled distinction can be made between meaning postulates and encyclopaedic knowledge. In the case of the logical vocabulary, however, a principled distinction *can* be made: there is no encyclopaedic knowledge associated with items of the logical vocabulary, so their inferential relations must be concept

[5] In fact, I do not accept Fodor's position regarding KILL. As Deirdre Wilson has pointed out to me, if it is a property of killing that it causes death, it makes sense that it is also a property of KILL that it implies CAUSE TO DIE, since we represent properties of the world. That, after all, is what representation is all about.

constitutive. This is fine, but it just passes the buck. The argument rests on there being a clear distinction between the logical and non-logical vocabulary, and Fodor owes us a story about how to draw this distinction.

But such a distinction is notoriously difficult to draw, and in fact Fodor and others have in the past presented plausible arguments that the distinction should not be drawn (Fodor et al. 1980: 269 ff; Katz 1972). Fodor et al. note that logic is commonly thought of as providing a reconstruction of our intuitions about the validity of arguments such as that in (1), but not of arguments such as that in (2).

(1) John left and Mary wept, therefore Mary wept
(2) John is a bachelor, therefore John is unmarried
(Fodor et al. 1980: 269)

The traditional view in logic has been that this is because argument (1) is valid in virtue of the meaning of 'and', which is part of the logical vocabulary, whereas argument (2) is valid in virtue of the meaning of 'bachelor', which is part of the non-logical vocabulary. But why, given that we have the intuition that both (1) and (2) are valid arguments, should we draw this kind of distinction?

Katz (1972: xix–xx) also points out that if we accept a distinction between logical and non-logical vocabulary, then we have a problem with pairs of sentences such as (3) and (4).

(3) Animals are mortal
(4) Animals do not live forever

The deductive relations of these sentences are identical, but if we distinguish between logical and non-logical vocabulary then they cannot be determined in the same way, since their logical forms are different. As Katz points out, there are indefinitely many such pairs.

What's more, there are all sorts of words which appear to have properties of both the logical and non-logical vocabulary. Consider the example of 'inside' (cf. Sperber & Wilson 1995: 105).

(5) a. *x* is inside *y*
b. *y* is inside *z*
(6) *x* is inside *z*

(5)–(6) is an example of a valid inference. The relation 'is inside' has certain formal properties: it is *transitive* (as just exemplified), *irreflexive* (nothing is inside itself), and

asymmetric (if *x* is inside *y*, then it is never the case that *y* is also inside *x*). These properties may be captured by a set of inference rules in the same way that the properties of a logical connective such as 'or' are captured. And yet it is patently not the case that 'inside' is a part of the logical vocabulary; at least part of its meaning derives from its content (it is a full-fledged concept, with associated encyclopaedic information).

The point about the logical vocabulary is that its content *just is* its inferential role. But the content of 'inside' *cannot* just be its inferential role; if it were, then it would be synonymous with 'below', which has identical logical properties (like 'inside', 'below' is transitive, irreflexive, and asymmetric).[6]

Fodor could, of course, still maintain a distinction: he could say that the logical vocabulary contains just those words whose meaning is exhausted by their inferential role. All other words would be non-logical, and their logical properties would be governed by empirical knowledge rather than inference rules.

But this would seem to be a very difficult position to defend. It would imply that the inferential role of non-logical concepts was not in any way constitutive of their content. This would mean that someone could be said to have the concept INSIDE, say, even if they were not disposed to accept the validity of the argument in (5)–(6). This would apply to other non-logical concepts as well. On this view, someone could have the concept DOG without knowing that dogs are animals, or even that dogs are physical objects. This does not seem plausible. Similarly, it is all very well to say that KILL → CAUSE TO DIE is a property of killing rather than a property of 'killing' or KILLING, but this implies that we could have the concept KILL without the predisposition to infer from '*x* KILL *y*' to '*y* DIE'. Would we really want to say that someone who did not know that killing carried the implication that there was a death nevertheless had the concept KILL?

Furthermore, the inference from '*x* KILL *y*' to '*y* DIE' is *necessary*, and is perceived as such; that is, the inference applies mandatorily. If the inference rule is stored as empirical knowledge, however, this provides no account of why this should be the case. It is well known that we do not have infallible access to our empirical knowledge. Why would it be any different in the case of this particular piece of empirical knowledge?

It would seem, then, that we must hold on to some content-constitutive inferences for the so-called 'non-logical' words.

---

[6] Conversely, it cannot be the case that the logical properties of 'inside' are general to spatial prepositions, and hence do not need to be specified for each individual preposition, since not all spatial prepositions have the same logical properties ('beside', for example, is non-transitive, irreflexive, and symmetric). It would thus seem that information about their logical properties has to be stored with the individual concepts in these cases.

5 Salvaging meaning postulates

We do appear to be in rather a mess at this point. If we accept Quine's arguments regarding the analytic–synthetic distinction, we would seem to be forced to reject content-constitutive inferences such as meaning postulates. But, at the same time, some inferences do indeed seem to be content constitutive (and not just in the case of a small set of logical words). What are we to do?

Perhaps there is some way out of this. Analyticity is a semantic notion, and it can thus be kept distinct from epistemic notions such as aprioricity (though, to be sure, the distinction has not always been rigorously maintained). It is also possible, however, to conceive of analyticity *psychosemantically*—that is, in purely mental terms. To do this is to draw a distinction between whether we consider an inference to be valid come-what-may and whether it is, in fact, valid come-what-may (i.e. analytic). Such a distinction is discussed in Horwich (1992).[7]

Horwich considers the possibility that the language faculty may contain certain meaning postulates (such as 'bachelor → unmarried man' or '*x* caused *y* → *x* preceded *y*') which are "transmitted to that area of the brain in which beliefs are stored" (1992: 100). Horwich suggests that it is then a simple matter to characterise a notion of analyticity which is completely determinate, and so immune from Quine's arguments.

Notice, however, that this is very different from the way the term 'analyticity' is normally construed: in particular, it is conceivable that we could have meaning postulates which were invalid (in fact, this is very probably the case). Imagine, for example, that we had a meaning postulate of the form 'whale → fish'. This would be psychosemantically analytic (and on some level we would believe the proposition 'whales are fish'), but not in fact analytic—the inference is invalid, and the corresponding belief is false.

By invoking a determinate notion of what we are calling psychosemantic analyticity, Horwich is able to keep meaning postulates without having to associate the content of a concept with its inferential role *tout court*. Content is associated with those inferences which are psychosemantically analytic, and this avoids the problems of holism which Fodor raised.

There are some problems, however, with Horwich's account. First, the language faculty doesn't seem to be the obvious place to locate these facts. It is certainly not a fact

[7] See also Boghossian (1994: 119–120) for some relevant discussion of this point.

about the *English words* 'cause' and 'effect' that causes precede their effects.[8] Rather, it's a fact about the world—it's a fact about *causing* and *effecting*. The question is then *where* and *how* this fact is represented in the brain.

It seems we must say that this is not a linguistic fact but a conceptual fact. Similarly, it is not a fact about the English word 'kill' (or about corresponding words in other languages) that killing someone entails that they die. It is a fact about the world, captured by properties of our concepts KILL and DIE, and stored under the appropriate concepts.

Let us consider, then, how meaning postulates might be represented. I take as a starting point the general model of concepts presented in Sperber & Wilson (1995: 85–93). According to this model, a concept has a label (or 'address'), under which three types of information may be stored: lexical information (information about the natural language word which expresses the concept), logical information (content-constitutive deductive rules which apply to logical forms containing the concept), and encyclopaedic information (information about the entities which fall under the concept). Within this framework, there are two places where the information captured by meaning postulates could be stored, corresponding to two different modes of storage. First, this information could be stored in the encyclopaedic entry for a concept, as suggested by Fodor, in which case the information would be stored propositionally. Alternatively, the information could be stored in the logical entry for a concept, as suggested by Sperber & Wilson (1995), in which case it would be stored as an inference rule (as we have seen, Fodor would reject this as a possibility).

We have seen that there are strong arguments in favour of the view that there are content-constitutive inferences. But accepting that there are such inferences would still neither explain where our intuitions of analyticity come from, nor their justification. This is a familiar epistemological problem with a priori belief (for a discussion, see Cowie 1999: Ch. 1). It is the second problem with Horwich's account.

This problem is evident in Horwich's claim that meaning postulates contained in the language faculty are "transmitted to that area of the brain in which beliefs are stored". This fairly innocent-looking statement covers up a thorny issue: the fact that a meaning postulate is stored in the language faculty does not explain our intuitions of analyticity. After all, we do not have corresponding intuitions about other 'rules' contained in the language faculty. We are confident about our intuition that "bachelors are unmarried men", but we are not at all confident in the same way about statements such as "a

[8] Indeed, there is plenty of evidence that children have knowledge of cause and effect long before they have acquired the words 'cause' and 'effect'. Leslie & Keeble (1987), for example, present evidence that infants as young as six months old have knowledge of causation.

variable must be c-commanded by an operator". The problem, then, is this: proposing that the language faculty contains a set of meaning postulates does not explain why we hold the beliefs corresponding to those meaning postulates, nor our intuitions about their analyticity.

One possible way round this is suggested in Sperber & Wilson (1995: 84). They stress the distinction between the logical relation of *implication* and the semantic relation of *entailment*. They explain the distinction as follows. A relation of implication holds between *P* and *Q* iff *Q* can be derived via deductive rules from *P*. A relation of entailment holds between *P* and *Q* iff there is no conceivable state of affairs in which *P* would be true and *Q* false. Now, if we accept this distinction then we can go some way towards explaining why we hold the beliefs corresponding to meaning postulates. Sperber & Wilson's explication of entailment provides a plausible candidate for the cognitive mechanism underlying our intuitions: if we can conceive of no state of affairs in which the antecedent of a meaning postulate was true while the consequent was false, then we would come to believe the necessity of the postulate.[9]

We must still consider the question of whether these are substantive claims. Recall Fodor's objection: would there be any difference between a mind with a meaning postulate and a mind with just the corresponding encyclopaedic knowledge? The answer seems to be 'yes'—there are all sorts of possible differences between these minds. As logical inference rules rather than general knowledge, meaning postulates could be applied faster and more efficiently or reliably (perhaps there is a module dedicated to processing such rules, or perhaps the formal properties of such rules allow the central systems to apply them more quickly).

The point was also made earlier that since the inferences captured by meaning postulates are drawn mandatorily, it is unlikely that they are stored as empirical knowledge (since we do not have infallible access to our empirical knowledge). In addition, there are various pieces of psychological evidence which suggest that rules such as 'φ CAT $\psi \rightarrow \varphi$ ANIMAL ψ'[10] are indeed "logical" principles and not general knowledge. One recent survey of this evidence is Margolis (1998). Margolis considers the acquisition of natural kind concepts within a basically Fodorian framework of concepts. He envisages something like the following scenario.

---

[9] This is clearly not the whole story, and would only explain our intuitions of necessity. Since analyticity does not reduce to necessity in the general case, more would have to be said.

[10] Rules such as these are taken to be substitution rules, and are not intended to be read as material implications. In particular, '$\rightarrow$' should be understood as 'rewrites as'; φ, ψ denote strings. Some restrictions must be placed on the set of strings over which φ, ψ can range, discussed in Cormack (1998: §1.2).

First, Margolis notes that conceptual atomism faces a problem in dealing with fakes. Most people distinguish natural kinds on the basis of superficial properties. But these superficial properties are not always reliable indicators of kind membership (consider a skunk disguised as a cat). There is psychological evidence that people have some sort of "psychological essentialism", according to which category membership for certain kinds is determined by hidden properties which cause their superficial properties (for a review of the evidence, see Medin & Ortony 1989, and Gelman & Coley 1991). If a person endowed with psychological essentialism were to discover that what they had taken to be a cat (on the basis of a set of superficial properties which are normally reliable indicators of kind membership, what Margolis refers to as a "syndrome") was actually a disguised skunk, then they would infer that it lacked the essential property which gives rise to the cat-syndrome, and would cease to apply CAT to it. Margolis then proposes the following model of natural-kind acquisition:

i. Young children believe that certain categories are natural kinds and that these categories are subject to a principle of essentialism;
ii. Essentialism says that it is not superficial properties but an essential property which determines category-membership;
iii. Young children are also predisposed to respond to the types of properties that enter into a kind-syndrome and which are thus highly indicative of a kind;
iv. There are, in fact, syndromes for some natural kinds.

So: a child sees some cats and certain of their properties trigger an essentialist principle. What the child focuses on, however, are certain cognitively salient superficial properties of the cats; this leads to the child acquiring certain beliefs about cats which represent the cat-syndrome. Because the child is a psychological essentialist, however, she takes it that what is important for kind membership is not these superficial properties, but some (unknown) essential property which causes them. The child has acquired a state of mind that causally links her to cats in such a way that she has the concept CAT.

In support of this account, Margolis notes the following:

i. There is a fair amount of empirical evidence that children around the age of three or four do indeed have an essentialist disposition. They seem to be prepared to override gross perceptual similarity in simple induction tasks, and to understand that an object's insides may differ from its outsides, and that what's inside may for certain kinds of thing be more pertinent.

ii. Animacy is a reliable indicator of one class of natural kinds, and children could detect animacy on the basis that animate objects do not require an external force to put them into motion. A number of such principles (for animals, plants, and perhaps substances and other kinds) could explain why essentialism is triggered in the case of these kinds, but not others.
iii. There is converging evidence for the importance of recognising a basic level of perceptual categories. At this basic level, shape tends to correlate with kind. Moreover, children have been shown to be guided by shape similarity (and possibly other heuristics such as texture and function) in categorisation tasks. This is evidence that children are indeed predisposed to respond to the sorts of properties which are indicative of kind membership (i.e. the properties that enter into kind-syndromes).

In light of this, let us return to the question in hand: what possible differences there would be between a mind in which a given inferential relation was a meaning postulate, and a mind in which it was general knowledge. Consider the inference 'φ CAT $\psi \rightarrow \varphi$ ANIMAL ψ'. If we take the arguments presented by Margolis seriously, one obvious difference suggests itself: only a mind for which 'φ CAT $\psi \rightarrow \varphi$ ANIMAL ψ' was a meaning postulate could acquire CAT as a natural kind concept.

Consider: A cat is a good candidate for an entity which is clearly a member of the animal kind. It is middle-sized and highly animate, and displays a clear syndrome (that is, all and only cats appear to us to be cats; compare this to the case of aphids, which are neither middle-sized nor obviously animate). If children are indeed endowed with an essentialist principle in respect of animal kinds, it seems that cats would certainly trigger this essentialist principle and they will be led to believe that cats have a hidden essence.

In fact, there is a considerable body of evidence suggesting that an essentialist principle operates in a number of different cognitive domains (see Gelman, Coley & Gottfried 1994 for a discussion). These domains include (inter alia) naïve biology (Atran 1990), theory of mind (Wellman 1990), and possibly physical causality (Shultz 1982). In other domains, such as artefacts, it appears that no essentialist principle applies (no one claims that artefacts are assigned membership of categories on the basis of whether they possess a hidden causal essence, although it might be possible to argue that artefacts have some kind of functional essence). This strongly suggests that we have distinct essentialist principles which apply in distinct domains, rather than a single domain-general principle.[11]

[11] There is, of course, the possibility that there is a single essentialist principle, but that not all domains invoke it. See, however, Hirschfeld (1994) for arguments against this possibility.

If this is the case, we could generalise Margolis's approach to acquisition so that it covered all domains where an essentialist principle applied. To do this, we would need to assume that different properties of entities trigger essentialism in different domains. Once essentialism is triggered, the child is able to focus on salient perceptible properties of a class of objects, whilst at the same time taking it that what is important for kind membership is some unknown essential property. Our minds would have to be constructed so that the perceptible properties which trigger essentialism are different in different domains.

But if our minds are constructed in this way, then it is a nomological *law* that cats (say) trigger an essentialist disposition. And it is a nomological law that if we have domain-Y essentialism with respect to Xs, then Xs strike us as being of kind Y. So it is a nomological law that cats strike us as animals (and hence that we believe them capable of internally-directed movement, and that they require sustenance in order to survive, and so on). This seems to me to be a very good argument for claiming that 'cats are animals' is psychologically analytic, and hence for proposing a meaning postulate along the lines of 'φ CAT $\psi \rightarrow \varphi$ ANIMAL ψ'.[12] The same would be true for the corresponding meaning postulates in other domains.

There is a possible objection to the argument I have just sketched. While it may be true that acquiring CAT on the basis of experience of the cat-syndrome (either through experience of real cats, or through experience of those cat-representations which manage to preserve the cat-syndrome) leads one to be a cat essentialist and to therefore have the meaning postulate 'φ CAT $\psi \rightarrow \varphi$ ANIMAL ψ', this is not the only conceivable way of acquiring CAT. Suppose that CAT were acquired not on the basis of the cat-syndrome, but from representations of cats which did not preserve the syndrome (cat tales, for example). A person acquiring CAT in this way would not necessarily be a cat essentialist, since they would not have had the right sorts of experiences to trigger essentialism in this case. This suggests that CAT need not *entail* ANIMAL, and hence that there is no corresponding meaning postulate.

---

[12] One could also propose a meaning postulate along the lines of 'φ CAT $\psi \rightarrow \varphi$ LIVING KIND ψ' or 'φ CAT $\psi \rightarrow \varphi$ NATURAL KIND ψ'. It has been argued, however, that children do not have a general concept NATURAL KIND, or even LIVING KIND, and that their most general categories correspond to ANIMAL and PLANT (see Carey 1985). Even if children do have a general NATURAL KIND concept, however, principles of economy might suggest that the necessity of the inference from CAT to NATURAL KIND is captured not by a single meaning postulate, but rather by 'φ CAT $\psi \rightarrow \varphi$ ANIMAL ψ' together with 'φ ANIMAL $\psi \rightarrow \varphi$ NATURAL KIND ψ'. Otherwise, each animal concept C would have to contain both the meaning postulate 'φ C $\psi \rightarrow \varphi$ ANIMAL ψ' and 'φ C $\psi \rightarrow \varphi$ NATURAL KIND ψ'.

My response to this will be to claim that, in fact, we cannot acquire CAT other than on the basis of the cat-syndrome. The basis for this claim is the distinction between intuitive and reflective concepts put forward by Dan Sperber (1997).

6 Intuitive and reflective beliefs

The resources we need to postulate to account for belief/desire psychology would seem to consist, at minimum, of a 'belief box' and a 'desire box'. We could also postulate other 'boxes' to account for other propositional attitudes, but Sperber & Wilson (1995: 73–75) argue that given the indefinite variety of possible attitudes, it is preferable to handle these through metarepresentation. According to this proposal, doubting that aardvarks like milk would consist of having in one's belief box a representation such as 'I doubt [that aardvarks like milk]'. It is possible to handle in this way any attitude which one has the conceptual vocabulary to explicitly represent.

In particular, one can represent a whole range of attitudes which, although not strictly attitudes of belief, would tend to support the truth of their propositional objects (Sperber 1997 gives examples such as, 'It is a scientific fact [that a glass of wine a day is good for the heart]'). Sperber refers to these as 'credal attitudes'. He generalises the situation as follows. A belief with metarepresentational content may provide a validating context (V) for the embedded representation (R). In such a situation the individual has two credal attitudes, one with content V(R) (in virtue of the representation V(R) occurring in the their belief box), and one with content R (in virtue *not* of the representation R itself occurring in their belief box, but in virtue of R occurring in a validating context). Sperber refers to the former attitude as *intuitive belief* and the latter attitude as *reflective belief*.

There is a certain sense in which reflective beliefs encode something of their etiology in the form of their validating context. This need not always be the case, however: the attitude of belief itself could be entertained metarepresentationally, by having a representation of the form 'I believe that *P*' in one's belief box. A plausible assumption would be that there are disquotational mechanisms which would apply to such beliefs, extracting the content, and creating an intuitive belief by adding it to the belief box unembedded, provided the proposition being added does not contradict any already contained in the belief box. Reflective beliefs of the kind 'I believe that *P*' (as opposed to simple beliefs that *P*) would then seem to be superfluous. There are, however, situations in which they could be important. Consider the case where *P* is inconsistent with other intuitive beliefs; suppose, for example, that *P* is the proposition 'Cats are Martian robots'. As argued earlier, it may be a fact of our cognitive make-up that cats

strike us as animals, and it may thus be impossible to override this intuitive belief. This would make it impossible to add the proposition 'Cats are Martian robots' to our belief box. But there is nothing to prevent an individual holding 'Cats are Martian robots' as a reflective belief. It may be that this is represented in the simplest way, in the form 'I believe [that cats are Martian robots]'.[13]

This particular example may seem outlandish, but this kind of situation arises quite commonly whenever our conscious, rational beliefs clash with our intuitions. Modern science throws up a host of examples: the quantum properties of objects clash with our intuitive beliefs about the behaviour of objects, to take one obvious example.

7 Intuitive and reflective concepts

The distinction between intuitive and reflective beliefs implies a corresponding distinction between intuitive and reflective concepts. According to Sperber, intuitive concepts constitute the vocabulary of the intuitive inference mechanisms, that is, the combined vocabulary of our perceptual processes and spontaneous inferential processes. Our remaining concepts—those which are not perceptual, nor derived from percepts by spontaneous inference, are reflective concepts. As Sperber points out, it is clear that the repertoire of reflective concepts is richer than the repertoire of intuitive concepts. This follows from the fact that, in principle, any intuitive concept may be used in representing a reflective belief, whereas the converse is not true.

There is mounting evidence that concepts come in different kinds, and that these different kinds have different properties (among the properties that have been proposed are that they are processed by distinct modules, that they are derived from distinct templates, or that they license distinct sets of inferences). Such evidence has been presented by Atran (1987), Bloom (1996), and Sperber (1994), among others.

The basic idea is that concepts of a given kind are derived from a generic template for that kind, and that representations which include concepts derived from this template are processed by a dedicated module (see Sperber 1994). If this mechanism is to be effective, the child must have some method of determining which kind of concept it is learning. In the case of cats, it might be something like this: the child sees some cats, and notices that they move without an external force being applied. This causes the ANIMAL

[13] There is a question about how such a belief could arise. Clearly it cannot arise from embedding the proposition in the context 'I believe…', since this would imply that it had been an intuitive belief, which seems unlikely. Rather, it could have arisen in a validating context such as 'The guru said…' (possibly during a delusional state), with the etiology subsequently being lost.

template to be initialised (which may itself be derived from a more general LIVING KIND template) producing the CAT concept.

The important point for our purposes is that, whatever feature of cats it is that leads to the animal template being initialised on exposure to cats, mere cat tales or other cat-representations which do not preserve the cat syndrome will not possess it. A child may hear mention of cats, and this may lead to them acquiring beliefs such as 'cats like milk', 'cats eat mice' (or even 'cats are animals', since the template-initialisation mechanism is presumably mediated by percepts, not beliefs). All these beliefs would be reflective, however, because the individual does not yet have a CAT concept, and so lacks the resources to intuitively represent these beliefs. In order to represent these beliefs reflectively, the child has a reflective concept of cats, but this is a mere 'placeholder'. It is not until the child is exposed to the cat-syndrome that an intuitive concept can develop.[14]

We can now see that the earlier objection—that a natural kind concept could be acquired other than on the basis of exposure to the syndrome—is false when taken as a statement about intuitive concepts. Intuitive concepts of natural kinds can only be acquired on exposure to the appropriate syndrome. Furthermore, to acquire the intuitive concept in these circumstances, the individual must be an essentialist in respect of the kind, and so must have a meaning postulate of the form 'φ INDIVIDUAL $\psi \rightarrow \varphi$ KIND ψ' ('φ CAT $\psi \rightarrow \varphi$ ANIMAL ψ' or 'φ DAISY $\psi \rightarrow \varphi$ PLANT ψ', say).

Let us see where we've got to. We started with the following problem: While there are powerful arguments that certain inferences (meaning postulates) must contribute to the content of concepts, there are also powerful arguments that content cannot be constituted, even in part, by such inferences. By invoking a notion of 'psychosemantic analyticity' we were able to resolve this apparent dilemma, showing that meaning postulates could indeed be content-constitutive. Considerations of psychological essentialism then led us to the conclusion that meaning postulates are, in fact, nomologically necessary and arise out of fundamental properties of our cognitive architecture.

This view conflicts with a popular story about how we acquire and deploy concepts in two problematic classes of case: when entities falling under a concept do not display a syndrome (elms and beeches, say), and when we have not been exposed to that

[14] A question then arises as to how we explain the continuity between the placeholder concept and the intuitive concept. That is, how do we explain the fact that information stored under 'CAT' is presumably transferred to the intuitive CAT concept, when we acquire it? This problem arises whenever we acquire an intuitive concept after we already have the concept reflectively. Such a situation may not be common, but would presumably occur.

syndrome (aardvarks, say). In both cases we may have certain knowledge about the entities which fall under the concept, be able to reason about them, and be able to discuss them, in which case we would want to say that we have *some* concept of these entities, but one which is clearly deficient in some sense. The popular story goes like this.

8 Deference

An individual does not acquire ELM and BEECH on exposure to elms and beeches, since they do not display a syndrome sufficiently distinct from that displayed by other deciduous trees (at least, this is true from the perspective of a modern urbanite). And yet individuals do acquire these concepts, albeit deficient versions which do not enable direct classification. Normally, a concept is causally linked to the entities which fall under it. The same is true in these more problematic cases, it is claimed, except that the causal link is a little less direct. There are members of the individual's community who do have direct causal links to the entities falling under the concept ('experts', as one says), and the other members of the community have a causal link to the experts (in the form of a disposition to defer to them on matters of classification). Or, to put things another way: we are cat detectors and mouse detectors; we are not acid detectors or base detectors, but we can construct a detector (litmus paper), and so we can detect acids and bases; we cannot detect elms or beeches, and we do not need to construct detectors, since we already have detectors in our environment—experts (cf. Fodor 1994). On this view, to acquire a concept just *is* to become causally linked to the entities which fall under it in the right way. So to acquire a concept deferentially one has merely to acquire a disposition to defer to an expert.

The other kind of case is when we have not had exposure to a syndrome. We hear aardvark-tales, without encountering any of the beasts, or even encountering representations of them which could provide knowledge of the syndrome. We have no knowledge of the aardvark syndrome, and we may therefore have no classificatory ability in respect of aardvarks. We are thus not causally connected to aardvarks in the right way. But in our community there are individuals who have been exposed to the aardvark syndrome, and who are thus causally connected to aardvarks. So if we have a disposition to defer to them we can complete our causal link.

This story has great appeal, and some version of it has been embraced by philosophers as diverse as Putnam (who first proposed it) and Fodor. Despite its appeal, however, I will argue that it is not psychologically plausible. While it is important as a general

cognitive strategy, the role of deference in the acquisition and deployment of concepts is far less important than has been generally recognised.

Consider first intuitive concepts (we will discuss reflective concepts in a moment). Presumably, such concepts must be derived from some template, in order for them to be processed by a particular module. This would seem to be *just what is meant* by the claim that certain concepts are intuitive—we are able to use a concept intuitively just because there is a module supporting a certain class of inferences available to process the concept.

Now, notice that the greater the amount of information which is hardwired by a template, the less scope there is for deference. If I do not know whether elms are natural kinds or artefacts, I cannot have a deferential concept for them, since I cannot initialise a template, and so cannot have any (intuitive) concept for them at all. This follows from the plausible assumption that template initialisation relies on perception rather than on any rational process—merely being told (however convincingly) that aardvarks are animals will not serve to initialise the animal template; only exposure to the aardvark syndrome can do this.

Deference would also appear to be largely irrelevant in the early acquisition of concepts. Note that it is in cases where there is no obvious syndrome (either because the kind is not locally instantiated, or because it lacks a syndrome) that we may be forced to rely on experts to do our categorising. This is not usually the case at the basic level, however, and it is at the basic level of locally instantiated[15] objects that the bulk of early acquisition takes place.[16]

Consider classification in more detail, in the context of the distinction between intuitive and reflective concepts. Classification of cats requires little reflective thought in typical environments. Classification of objects which may be useful to a person marooned on a desert island, however, is a reflective activity (given that this is a non-intuitive concept). So is classification of distal cats via someone else's use of a telescope. It seems clear that we can classify intuitively or reflectively. And surely the classificatory function of an intuitive concept is *intuitive* classification. That's part of what intuitive concepts are *for*. If we come face-to-face with a tiger, we want to classify

[15] Here I take 'instantiated' to include representations which give some access to the syndrome, such as drawings, cartoons, photographs, and so on.

[16] Woodfield (forthcoming) also argues that it is impossible to acquire a novel concept via deference, since in order to defer to an expert on Xs, one must know *which concept* one is deferring on (i.e. one must already have a concept which picks out Xs).

it as such, and pretty quickly. We don't want to reflect on anything, and we certainly don't want to defer to any experts.

It is in *reflective* classification that we defer to experts—apart from anything else, finding the right expert is in itself a reflective activity. If this is right, then since the repertoire of reflective concepts potentially includes all the intuitive concepts, there should not be any restriction on deference to experts. In particular, we should find cases of deference even when we have good knowledge of a syndrome, *or even when we are experts ourselves*—and not just when we lack the ability to classify objects falling under a concept. And indeed, this is what we find. Imagine a person who is blindfolded, or who has become temporarily blinded. They may defer to experts as to which concept an entity falls under (was that a cat or a skunk which just rubbed against my leg?) without thereby implying that they lack knowledge of the corresponding syndrome. They may even be an expert themselves. Also, we defer on Xs when we have *no intuitive concept at all* of Xs. This is the case with any poorly-understood concept, *or when we do not understand the concept at all*. The latter case is illustrated as follows. Suppose that I do not know what the word 'iconoclast' means. I can ask someone: 'What does *iconoclast* mean?', and I can have the corresponding thought. What concept stands in for 'iconoclast' in the mental representation of this thought? It cannot be ICONOCLAST, since we are assuming that I do not know the meaning of 'iconoclast'. Rather, we might follow Sperber[17] and suppose that we have a 'placeholder' concept, "ICONOCLAST", a reflective concept with no (or little) associated encyclopaedic information, which is linked to the word 'iconoclast'.

The point is that deference is a general strategy to cope with the epistemically limited situations we find ourselves in: if you don't have a particular piece of relevant information, there's a good chance someone else does—defer to them. And because it is a general cognitive strategy, deference isn't restricted to determining the referents of a certain class of concepts (those in respect of which we do not have a classificatory ability). In particular:

i. We can have a good classificatory ability with respect to Xs and still defer;
ii. We can defer on Xs without having any intuitive X-concept at all.

Deference won't help us to acquire or sustain an intuitive concept, since it does not allow us to initialise a template for the concept. Rather, deference is used with reflective

[17] See, for example, Sperber (1997: 74–75).

concepts (even those which we can classify very effectively in normal circumstances), as part of a general cognitive strategy.

Deference was originally proposed to solve a particular problem: how is it that we seem to have certain concepts, and yet we cannot classify the entities which fall under those concepts? The problem disappears if we make a distinction between intuitive and reflective concepts. If we cannot recognise Xs, then we do not have an intuitive concept of Xs. Deference will not help us. If, however, we find ourselves in an epistemically limited position—either because we cannot invoke our classificatory ability (maybe we are blindfolded), or because we do not have a suitable classificatory ability—then we may defer to 'experts'. There is no sub-class of deferential concepts. Rather, deference is a general strategy which may be applied to any reflective concept (and to no intuitive concept).

References

Atran, S. (1987) Ordinary constraints on the semantics of living kinds. *Mind and Language* 2: 27–60.

Atran, S. (1990) *Cognitive Foundations of Natural History: Towards an Anthropology of Science*. Cambridge: Cambridge University Press.

Bloom, P. (1996) Intention, history, and artifact concepts. *Cognition* 60: 1–29.

Boghossian, P. A. (1993) Does an inferential role semantics rest upon a mistake? *Mind and Language* 8: 27–40.

Boghossian, P. A. (1994) Inferential role semantics and the analytic/synthetic distinction. *Philosophical Studies* 73: 109–122.

Burge, T. (1979) Individualism and the mental. In French, P. A., T. E. Uehling & H. K. Wettstein (eds.). *Studies in Metaphysics* [Midwest Studies in Philosophy, vol. IV]. 73–121. Minneapolis: University of Minnesota Press.

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

Carnap, R. (1947) *Meaning and Necessity*. Chicago: University of Chicago Press.

Chomsky, N. (1988) *Language and Problems of Knowledge*. Cambridge, Mass.: MIT Press.

Cormack, A. (1998) *Definitions: Implications for Syntax, Semantics, and the Language of Thought*. New York: Garland.

Cowie, F. (1999) *What's Within? Nativism Reconsidered.* Oxford: Oxford University Press.

Fodor, J. A. (1994) *The Elm and the Expert: Mentalese and its Semantics*. Cambridge, Mass.: MIT Press.

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

Fodor, J. A. & E. Lepore (1992) *Holism: A Shopper's Guide*. Oxford: Blackwell.

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

Fodor, J. D., J. A. Fodor & M. F. Garrett (1975) The psychological unreality of semantic representations. *Linguistic Inquiry* 4: 515–531.

Gelman, S. A. & J. D. Coley (1991) Language and categorisation: The acquisition of natural kind terms. In Gelman, S. & J. Byrnes (eds.). *Perspectives on Language and Thought: Interrelations in Development*. 146–196. Cambridge: Cambridge University Press.

Gelman, S. A., J. D. Coley & G. M. Gottfried (1994) Essentialist beliefs in children: The acquisition of concepts and theories. In Hirschfeld & Gelman (eds.) (1994). 341–365.

Hirschfeld, L. A. (1994) Is the acquisition of social categories based on domain-specific competence or on knowledge transfer? In Hirschfeld & Gelman (eds.) (1994). 201–233.

Hirschfeld, L. A. & S. A. Gelman (eds.) (1994) *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge: Cambridge University Press.

Horwich, P. (1992) Chomsky versus Quine on the analytic–synthetic distinction. *Proceedings of the Aristotelian Society* 92: 95–108.

Katz, J. (1972) *Semantic Theory*. New York: Harper & Row.

Leslie, A. M. S. & Keeble (1987) Do 6-month old infants perceive causality? *Cognition* 25: 265–288.

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

Medin, D. & A. Ortony (1989) Psychological essentialism. In Vosniadou, S. & A. Ortony (eds.). *Similarity and Analogical Reasoning*. 179–195. Cambridge: Cambridge University Press.

Partee, B. (1995) Lexical semantics and compositionality. In Gleitman, L. & M. Liberman (eds.). *An Invitation to Cognitive Science*, 2nd edition, vol. 1. 311–360. Cambridge, Mass.: MIT Press.

Putnam, H. (1975) The meaning of "meaning". In Gunderson, K. (ed.). *Language, Mind, and Knowledge* [Minnesota Studies in the Philosophy of Science, vol. VII]. 131–193. Minneapolis: University of Minnesota Press.

Quine, W. V. O. (1953) Two Dogmas of Empiricism. In *From a Logical Point of View*. 20–46. Cambridge, Mass.: Harvard University Press.

Shultz, T. R. (1982) Rules of Causal Attribution. *Monographs of the Society for Research in Child Development* 47 (1).

Sperber, D. (1994) The modularity of thought and the epidemiology of representations. In Hirschfeld & Gelman (eds.) (1994). 39–67.

Sperber, D. (1997) Intuitive and reflective beliefs. *Mind and Language* 12: 67–83.

Sperber, D. & D. Wilson (1995) *Relevance*, 2nd edition. Oxford: Blackwell.

Wellman, H. M. (1990) *The Child's Theory of Mind*. Cambridge, Mass.: MIT Press.

Woodfield, A. (forthcoming) Reference and deference. To appear in *Mind and Language* 15.4 (September 2000).