5

Prototypes and Compositionality[1]

A Good Apple tree or a bad, is an Apple tree still: a Horse is not more a Lion for being a Bad Horse.

William Blake

Introduction

The definition theory says that concepts are complex structures which entail their constituents. By saying this, it guarantees both the connection between content and necessity and the connection between concept individuation and concept possession. On the one hand, since definitions *entail* their constituents, it follows that whatever belongs to a concept's definition is thereby true of everything, actual or possible, that the concept subsumes. On the other hand, since what definitions entail are their *constituents*, it follows that a definition of a concept specifies its canonical (viz. individuating) structural description. And finally, whatever else concept possession may amount to, you can't have a thing unless you have its parts; hence the connection between concept possession and concept individuation according to the definition story. This metaphysical synthesis of a theory of concept individuation with theories of modality and concept possession was no small achievement. In some respects it has yet to be bettered, as we're about to see.

By and large, it's been the modal properties of definitions that philosophers have cared about since, as previously remarked, the semantical truths that definitions generate recommend themselves for

[1] Terminological conventions with respect to the topics this chapter covers are unsettled. I'll use 'stereotype' and 'prototype' interchangeably, to refer to mental representations of certain kinds of properties. So, 'the dog stereotype' and 'the dog prototype' designate some such (complex) concept as: BEING A DOMESTIC ANIMAL WHICH BARKS, HAS A TAIL WHICH IT WAGS WHEN IT IS PLEASED, . . etc. I'll use 'exemplar' for the mental representation of a kind, or of an individual, that instantiates a prototype; so 'sparrows are the exemplars of birds' and 'Bambi is Smith's exemplar of a deer' are both well-formed. 'Sparrows are stereotypic birds' (/'Bambi is a prototypic deer') are also OK; they mean that a certain kind (/individual) exhibits certain stereotypic (/prototypic)

antisceptical employment. By contrast, it's their being *complex* that primarily makes definitions interesting to psychologists and linguists. With complex things, there's always the hope that their behaviour can be predicted from the behaviour of their parts; with primitive things, since there are no parts, there is no such hope. In particular (for the linguists), if words have definitions, then arguably words have the syntax of phrases "at the semantic level"; so perhaps lexical grammar can be unified with phrasal grammar. Likewise (for the psychologists), if lexical concepts are *tacitly* structurally complex, perhaps they can be brought under the same psychological generalizations that govern concepts that are *manifestly* complex; if the concept BACHELOR is the concept UNMARRIED MAN, then learning or thinking with the one can't differ much from learning or thinking with the other.[2]

So the definition theory was a fusion of disparate elements; in particular, the idea that concepts are complex and the idea that their constitutive inferences are typically necessary are in principle dissociable. And, for better or worse, they have been coming unstuck in the recent history of cognitive science. The currently standard view is that the definition story was right about the complexity of typical lexical concepts, but wrong to claim that complex concepts typically *entail* their constituents. According to the new theory, it's not the *necessity* of an inference but its *reliability* that determines its relevance to concept individuation.

How this is supposed to work, and why it doesn't work the way that it's supposed to, and where its not working the way that it's supposed to leaves us in the theory of concepts, will be the substance of this chapter.

Statistical Theories of Concepts

The general character of the new theory of concepts is widely known throughout the cognitive science community, so the exegesis that follows will be minimal.

Imagine a hierarchy of concepts ordered by relations of dominance and sisterhood, where these obey the intuitive axioms (e.g. dominance is antireflexive, transitive and asymmetric; sisterhood is antireflexive, transitive, and symmetric, etc.). Figure 5.1 is a sort of caricature.

[2] The structural complexity of definitions was of some use to philosophers too: it promised the (partial?) reduction of conceptual to logical truth. So, for example, the conceptual truth that if John is a bachelor then John is unmarried, and the logical truth that if John is unmarried and John is a man then John is unmarried, are supposed to be

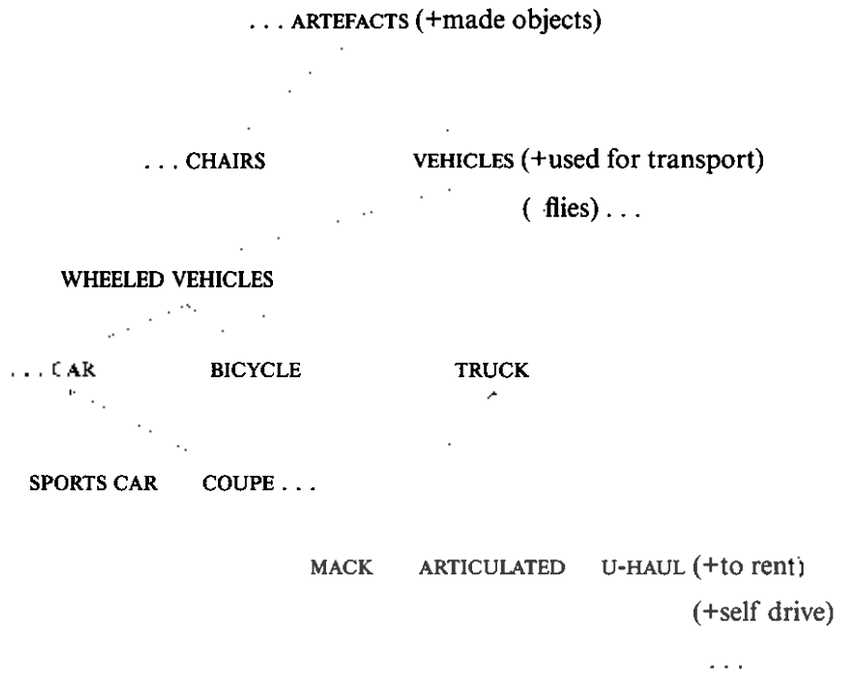

FIG. 5.1 An entirely hypothetical 'semantic hierarchy' showing the position and features of some concepts for vehicles.

The intended interpretation is that, on the one hand, if something is a truck or a car, then it's a vehicle; and, on the other hand, if something is a vehicle, then it's either a truck, or a car, or . . . etc. (Let's, for the moment, take for granted that these inferences are sound but put questions about their modal status to one side.) As usual, expressions in caps ('VEHICLE' and the like) are the *names* of concepts, not their structural descriptions. We continue to assume, as with the definition theory, that lexical concepts are typically complex. In particular, a lexical concept is a tree consisting of names of taxonomic properties together with their features (or 'attributes'; for the latter terminology, see Collins and Quillian 1969), which I've put in parentheses and lower case.[3] In a hierarchy like 5.1, each concept inherits the features of the concepts by which it is dominated.

[3] What, exactly, the distinction between semantic features and taxonomic classes is supposed to come to is one of the great mysteries of cognitive science. There is much to be said for the view that it doesn't come to anything. I shall, in any case, not discuss this issue here; I come to bury prototypes, not to exposit them.

Thus, vehicles are artefacts that are mobile, intended to be used for transport, . . . etc.; trucks are artefacts that are mobile, intended to be used for transport of freight (rather than persons), . . . etc. U-Haul trucks are artefacts that are mobile, intended to be rented to be used for transport of freight (rather than persons), . . . and so forth.

The claims of present interest are that when conceptual hierarchies like 5.1 are mentally represented:

i. There will typically be a *basic level* of concepts (defined over the dominance relations);

and

ii. There will typically be a *stereotype structure* (defined over the sisterhood relations).

Roughly, and intuitively: the *basic level concepts* are the ones that receive relatively few features from the concepts that immediately dominate them but transmit relatively many features to the concepts that they immediately dominate. So, for example, that it's a car tells you a lot about a vehicle; but that it's a sports car doesn't add a lot to what 'it's a car' already told you. So CAR and its sisters (but not VEHICLE or SPORTS CAR and their sisters) constitute a basic level category. Correspondingly, the *prototypical sister* at a given conceptual level is the one which has the most features in common with the rest of its sisterhood (and/or the least in common with non-sisters at its level). So, cars are the prototypical vehicles because they have more in common with trucks, buses, and bicycles than any of the latter do with any of the others.

Such claims should, of course, be relativized to an independently motivated account of the individuation of semantic features (see n. 3). Why, for example, isn't the feature bundle for VEHICLE just the unit set {+vehicle}? Well may you ask. But statistical theories of concepts are no better prepared to be explicit about what semantic features are than definitional theories used to be; in practice, it's all just left to intuition. That's scandalous, to be sure; but fortunately it doesn't matter a lot for the issues that will concern us here. As we're about to see, prototype concepts and basic object concepts exhibit a cluster of reliably correlated properties which allow us to pick them out pretty well even though, lacking a theory of features, we have no respectable account of what their basicness or their prototypicality consists in.

That concepts are organized into hierarchies isn't, of course, anything that definitional theories need deny. What primarily distinguishes the new story about concepts from its classical predecessor is the nature of the glue that's supposed to hold a feature bundle together. Defining features were

supposed to exhibit severally necessary and jointly sufficient conditions for a thing's inclusion in a concept's extension. On the present account, by contrast, whether a feature is in the bundle for a given concept is primarily a question of *how likely it is* that something in the concept's extension has the property that the feature expresses. Being able to fly isn't a *necessary* condition for being a bird (*vide* ostriches); but it is a property that birds are quite reliably found to have. So, *ceteris paribus*, *+flies* belongs to the feature bundle for BIRD. The effect, is to change from a kind of metaphysics in which the concept-constitutive inferences are distinguished by their *modal* properties to a kind of metaphysics in which they're picked out *epistemically*.[4]

Notice that the thesis that concepts are individuated by their inferential roles (specifically by their inferential relations to their constituents) survives this shift. It's just that the individuating inferences are now supposed to be statistical.[5] A fortiori, we're still working within a cognitivist account of concept possession: to have a concept is, at least *inter alia*, to believe certain things (e.g. in the case of BIRD, that generally birds fly). Notice also that the new story about concepts has claims to philosophical good repute that its definitional predecessor arguably lacked. Maybe, as Quine says, conceptual entailment isn't all that much clearer than the psychological and semantic notions that it was traditionally supposed to reconstruct. But if there's something philosophically wrong with statistical reliability, *everybody* is in trouble.

So, then, consider the thesis that concepts are bundles of statistically reliable features, hence that having a concept is knowing which properties the things it applies to reliably exhibit (together, perhaps, with enough of the structure of the relevant conceptual hierarchy to at least determine how basic the concept is).

A major problem with the definition story was the lack of convincing examples; nobody has a bullet-proof definition of, as it might be, 'cow' or 'table' or 'irrigation' or 'pronoun' on offer; not linguists, not philosophers.

[4] Elanor Rosche, who invented this account of concepts more or less single-handed often speaks of herself as a Wittgensteinian; and there is, of course, a family resemblance. But I doubt that it goes very deep. Rosche's project was to get modality out of semantics by *substituting* a probabilistic account of content-constituting inferences. Whereas I suppose Wittgenstein's project was to offer (or anyhow, make room for) an epistemic *reconstruction* of conceptual necessity. Rosche is an eliminativist where Wittgenstein is a reductionist. There is, in consequence, nothing in Rosche's theory of concepts that underwrites Wittgenstein's criteriology, hence nothing that's of use for bopping sceptics with.

[5] Just as it's possible to dissociate the idea that concepts are complex from the claim that meaning-constitutive inferences are necessary, so too it's possible to dissociate the idea that

least of all English-speakers as such. By contrast, the evidence that people know (and agree about) concerning the prototype structure of words and concepts is ubiquitous and robust.[6] In fact, you can hardly devise a concept-possession test on which prototype structure fails to have an appreciable effect. Ask a subject to tell you the first [redacted] that comes into his head, and it's good odds he'll report the prototype for the category [redacted]: cars for vehicles, red for colours, diamonds for jewels, sparrows for birds, and so on. Ask which vehicle-word a child is likely to learn first, and prototypicality is a better predictor than even very good predictors like the relative frequency of the word in the adult corpus. Ask an experimental subject to evaluate the truth of 'a [redacted] is a vehicle' and he'll be fastest where a word for the basic level prototype fills the blank. And so forth. Even concepts like ODD NUMBER, which clearly do have definitions, often have prototype structure as well. The number 3 is a 'better' odd number than 27 (and it's a better prime than 2) (see Armstrong, Gleitman, and Gleitman 1983). The discovery of the massive presence of prototypicality effects in all sorts of mental processes is one of the success stories of cognitive science. I shall simply take it for granted in what follows; but for a review, see Smith and Medin 1981.

So prototypes are practically everywhere and definitions are practically nowhere. So why not give up saying that concepts are definitions and start saying instead that concepts are prototypes? That is, in fact, the course that much of cognitive science has taken in the last decade or so. But it is not a good idea. Concepts can't be prototypes, *pace* all the evidence that everybody who has a concept is highly likely to have its prototype as well. I want to spend some time rubbing this point in because, though it's sometimes acknowledged in the cognitive science literature, it has been very much less influential than I think that it deserves to be. Indeed, it's mostly because it's clear that concepts can't be prototypes that I think that concepts have to be atoms.

[6] For a dissenting opinion, see Barsalou 1985 and references therein. I find his arguments for the instability of typicality effects by and large unconvincing; but if you don't, so much the better for my main line of argument. Unstable prototypes *ipso facto* aren't public (see Chapter 2), so they are *ipso facto* unfitted to be concepts.

Some of the extremist extremists in cognitive science hold not only that concepts are prototypes, but also that thinking is the 'transformation of prototype vectors'; this is the doctrine that Paul Churchland calls the "assimilation of 'theoretical insight' to 'prototype activation'" (1995, 117; for a review, see Fodor 1995*a*). But that's a minority opinion prompted, primarily, by a desire to assimilate a prototype-centred theory of concepts to a connectionist view about cognitive architecture. In fact, the identification of concepts with prototypes is entirely compatible with the "Classical" version of RTM according to which

In a nutshell, the trouble with prototypes is this. Concepts are productive and systematic. Since compositionality is what explains systematicity and productivity, it must be that concepts are compositional But it's as certain as anything ever gets in cognitive science that prototypes don't compose. So it's as certain as anything ever gets in cognitive science that concepts can't *be* prototypes and that the glue that holds concepts together can't be statistical.

Since the issues about compositionality are, in my view, absolutely central to the theory of concepts, I propose to go through the relevant considerations with some deliberation. We'll discuss first the status of the arguments for the compositionality of concepts and then the status of the arguments against the compositionality of prototypes.

The Arguments for Compositionality

Intuitively, the claim that concepts compose is the claim that the syntax and the content of a complex concept is normally determined by the syntax and the content of its constituents. ('Normally' means something like: *with not more than finitely many exceptions*. 'Idiomatic' concepts are allowed, but they mustn't be productive.) A number of people (see e.g. Block 1993; Zadrozny 1994) have recently pointed out that this informal characterization of compositionality can be trivialized, and there's a hunt on for ways to make the notion rigorous. But we can bypass this problem for our present purposes. Since the argument that concepts compose is primarily that they are productive and systematic, we can simply stipulate that the claim that concepts compose is true only if the syntax and content of complex concepts is derived from the syntax and content of their constituents *in a way that explains their productivity and systematicity*. I do so stipulate.

The Productivity Argument for Compositionality

The traditional argument for compositionality goes something like this: There are infinitely many concepts that a person can entertain. (*Mutatis*

connectionist architectures, it doesn't follow that the difference between the architectures is neutral with respect to prototypes. For example, in so far as Connectionism is committed to statistical learning as its model of concept acquisition, it may well *require* that concepts have statistical structure on pain of their being unlearnable. If, as I shall argue, the struc-

mutandis in the case of natural languages: there are infinitely many expressions of L that an L-speaker can understand.) Since people's representational capacities are surely finite, this infinity of concepts must itself be finitely representable. In the present case, the demand for finite representation is met if (and, as far as anyone knows, only if) all concepts are individuated by their syntax and their contents, and the syntax and contents of each complex concept is finitely reducible to the syntax and contents of its (primitive) constituents.

This seems as good an opportunity as any to say something about the current status of this line of thought. Of late, the productivity argument has come under two sorts of criticism that a cognitive scientist might find persuasive:

The performance/competence argument. The claim that conceptual repertoires are typically productive requires not just an idealization to infinite cognitive capacity, but the kind of idealization that presupposes a memory/program distinction. This presupposition is, however, tendentious in the present polemical climate. No doubt, if your model for cognitive architecture is a Turing machine with a finite tape, it's quite natural to equate the concepts that a mind could entertain with the ones that its program could enumerate *assuming that the tape supply is extended arbitrarily*. Because the Turing picture allows the size of the memory to vary while the program stays the same, it invites the idea that machines are individuated by their programs.

But this way of drawing a 'performance/competence' distinction seems considerably less natural if your model of cognitive architecture is (e.g.) a neural net. The natural model for 'extending' the memory of a network (and likewise, *mutatis mutandis*, for other finite automata) is to add new nodes. However, the idea of adding nodes to a network while preserving its identity is arguably dubious in a way that the idea of preserving the identity of a Turing machine tape while adding to its tape is arguably not.[8] The problem is precisely that the memory/program distinction isn't available for networks. A network is individuated by the totality of its nodes, and the nodes are individuated by the totality of their connections, direct and indirect, to one another.[9] In consequence, 'adding' a node to a network changes the identity of all the other nodes, and hence the identity

[8] If the criterion of machine individuation is I(nput)/O(utput) equivalence, then a finite tape Turing machine *is* a finite automaton. This doesn't, I think, show that the intuitions driving the discussion in the text are incoherent. Rather it shows (what's anyhow independently plausible) that I/O equivalence isn't what's primarily at issue in discussions of cognitive architecture. (See Dyrbuhn 1984.)

of the network itself. In this context, the idealization from a finite cognitive performance to a productive conceptual capacity may strike the theorist as begging precisely the architectural issues that he wants to stress.

The finite representation argument. If a finite creature has an infinite conceptual capacity, then, no doubt, the capacity must be finitely *determined*; that is, there must be a finite set of sufficient conditions, call it S, such that a creature has the capacity if S obtains. But it doesn't follow by any argument I can think of that satisfying S depends on the creature's representing the compositional structure of its conceptual repertoire; or even that the conceptual repertoire *has* a compositional structure. For all I know, for example, it may be that sufficient conditions for having an infinite conceptual capacity can be finitely specified in and only in the language of neurology, or of particle physics. And, presumably, notions like *computational state* and *representation* aren't accessible in these vocabularies. It's tempting to suppose that one has one's conceptual capacities in virtue of some act of intellection that one has performed. And then, if the capacity is infinite, it's hard to see what act of intellection that could be other than grasping the primitive basis of a system of representations; of Mentalese, in effect. But talk of grasping is tendentious in the present context. It's in the nature of intentional explanations of intentional capacities that they have to run out sooner or later. It's entirely plausible that explaining what determines one's conceptual capacities (figuratively, explaining one's mastery of Mentalese) is *where* they run out.

One needs to be sort of careful here. I'm not denying that Mentalese *has* a compositional semantics. In fact, I can't actually think of any other way to explain its productivity, and writing blank checks on neurology (or particle physics) strikes me as unedifying. But I do think we should reject the following argument: 'Mentalese *must have* a compositional semantics because mastering Mentalese requires *grasping* its compositional semantics.' It isn't obvious that mastering Mentalese requires grasping *anything*.

The traditional locus of the inference from finite determination to finite representation is, however, not Mentalese but English (see Chomsky 1965; Davidson 1967). Natural languages are learned, and learning is an 'act of intellection' par excellence. Doesn't that show that English has to have a compositional semantics? I doubt that it does. For one thing, as a number of us have emphasized (see Chapter 1; Fodor 1975; Schiffer 1987; for a critical discussion, see Lepore 1997), if you assume that thinking is computing, it's natural to think that acquiring a natural language is learning how to translate between it and the language you compute in

and explicitly represented. Still, there is no obvious reason why translation between English and Mentalese requires having a compositional theory of *content* for either language. Maybe translation to and from Mentalese is a syntactical process: maybe the Mentalese translation of an English sentence is fully determined given its canonical structural descriptions (including, of course, lexical inventory).

I don't really doubt that English and Mentalese are both productive; or that the reason that they are productive is that their semantics is compositional. But that's faith in search of justification. The polemical situation is, on the one hand, that minds are productive only under a tendentious idealization; and, on the other hand, that productivity doesn't literally entail semantic compositionality for either English or Mentalese. Somebody sane could doubt that the argument from productivity to compositionality is conclusive.

The Systematicity Argument for Compositionality

'Systematicity' is a cover term for a cluster of properties that quite a variety of cognitive capacities exhibit, apparently as a matter of nomological necessity.[10] Here are some typical examples. If a mind can grasp the thought that $P \rightarrow Q$, it can grasp the thought that $Q \rightarrow P$; if a mind can grasp the thought that $\sim(P \,\&\, Q)$, it can grasp the thought that $\sim P$ and the thought that $\sim Q$; if a mind can grasp the thought that Mary loves John, it can grasp the thought that John loves Mary . . . etc. Whereas it's by no means obvious that a mind that can grasp the thought that $P \rightarrow Q$ can also grasp the thought that $R \rightarrow Q$ (not even if, for example, $(P \rightarrow Q) \rightarrow (R \rightarrow Q)$. That will depend on whether it is the kind of mind that's able to grasp the thought that R. Correspondingly, a mind that can think *Mary loves John* and *John loves Mary* may none the less be unable to think *Peter loves Mary*. That will depend on whether it is able to think about Peter.

It seems pretty clear why the facts about systematicity fall out the way they do: mental representations are compositional, and compositionality explains systematicity.[11] The reason that a capacity for *John loves Mary*

[10] It's been claimed that (at least some) facts about the systematicity of minds are *conceptually* necessary; 'we wouldn't call it thought if it weren't systematic' (see e.g. Clark 1991). I don't, in fact, know of any reason to believe this, nor do I care much whether it is so. If it's *conceptually* necessary that thoughts are systematic, then it's *nomologically* necessary that creatures like us have thoughts, and this latter necessity still wants explaining.

[11] It's sometimes replied that compositionality doesn't *explain* systematicity since

thoughts implies a capacity for *Mary loves John* thoughts is that the two kinds of thoughts have the same constituents; correspondingly, the reason that a capacity for John loves Mary thoughts does *not* imply a capacity for *Peter loves Mary* thoughts is that they *don't* have the same constituents. Who could really doubt that this is so? Systematicity seems to be one of the (very few) organizational properties of minds that our cognitive science actually makes some sense of.

If your favourite cognitive architecture doesn't support a productive cognitive repertoire, you can always argue that since minds are really finite, they aren't *literally* productive. But systematicity is a property that even quite finite conceptual repertoires can have; it isn't remotely plausibly a methodological artefact. If systematicity needs compositionality to explain it, that strongly suggests that the compositionality of mental representations is mandatory. For all that, there has been an acrimonious argument about systematicity in the literature for the last ten years or so. One does wonder, sometimes, whether cognitive science is worth the bother.

Some currently popular architectures *don't* support systematic representation. The representations they compute with lack constituent structure; a fortiori they lack compositional constituent structure. This is true, in particular, of 'neural networks'. Connectionists have responded to this in a variety of ways. Some have denied that concepts are systematic. Some have denied that Connectionist representations are inherently unstructured. A fair number have simply failed to understand the problem. The most recent proposal I've heard for a Connectionist treatment of systematicity is owing to the philosopher Andy Clark (1993). Clark says that we should "bracket" the problem of systematicity. "Bracket" is a technical term in philosophy which means *try not to think about*.

I don't propose to review this literature here. Suffice it that if you assume compositionality, you can account for both systematicity and productivity; and if you don't, you can't. Whether or not productivity and systematicity *prove* that conceptual content is compositional, they are clearly substantial straws in the wind. I find it persuasive that there are

continental drift explains why (e.g.) South America fits so nicely into Africa. It does so, however, not by *entailing* that South America fits into Africa, but by providing a theoretical background in which the fact that they fit comes, as it were, as no surprise. Similarly, *mutatis mutandis*, for the explanation of systematicity by compositionality.

Inferences from systematicity to compositionality are 'arguments to the best explanation', and are (of course) *non-demonstrative*: which is (of course) not at all the same

quite a few such straws, and they appear all to be blowing in the same direction.

The Best Argument for Compositionality

The best argument for the compositionality of mental (and linguistic) representation is that its traces are ubiquitous; not just in very general features of cognitive capacity like productivity and systematicity, but also everywhere in its details. Deny productivity and systematicity if you will; you still have these particularities to explain away.

Consider, for example: the availability of (definite) descriptions is surely a universal property of natural languages. Descriptions are nice to have because they make it possible to talk (*mutatis mutandis*, to think) about a thing even if it isn't available for ostension and even if you don't know its name; even, indeed, if it doesn't *have* a name (as with ever so many real numbers). Descriptions can do this job because they pick out unnamed individuals *by reference to their properties*. So, for example, 'the brown cow' picks out a certain cow; viz. the brown one. It does so by referring to a property, viz. *being brown*, which that cow has and no other cow does that is contextually relevant. Things go wrong if (e.g.) there are no contextually relevant cows; or if none of the contextually relevant cows is brown; or if more than one of the contextually relevant cows is brown . . . And so forth.

OK, but just how does all this work? Just what is it about the syntax and semantics of descriptions that allows them to pick out unnamed individuals by reference to their properties? Answer:

i. Descriptions are complex symbols which have *terms that express properties* among their syntactic constituents;

and

ii. These terms contribute the properties that they express to determine what the descriptions that contain them specify.

It's because 'brown' means *brown* that it's the brown cow that 'the brown cow' picks out. Since you can rely on this arrangement, you can be confident that 'the brown cow' will specify the local brown cow *even if you don't know which cow the local brown cow is*; even if you don't know that it's Bossie, for example, or that it's *this* cow. That, however, is just to say that descriptions succeed in their job *because* they are compositional. If English didn't let you use 'brown' context-independently to mean *brown*,

Names, by contrast, succeed in their job because they *aren't* compositional; not even when they are syntactically complex. Consider 'the Iron Duke', to which 'Iron' does *not* contribute *iron*, and which you can therefore use to specify the Iron Duke even if you don't know what he was made of. Names are nicer than descriptions because you don't have to know much to specify their bearers, although you *do* have to know what their bearers are called. Descriptions are nicer than names because, although you do have to know a lot to specify their bearers, you *don't* have to know what their bearers are called. What's nicer than having the use of either names or descriptions is having the use of both. I agree that, as a piece of semantic theory, this is all entirely banal; but that's my point, so don't complain. There is, to repeat, no need for *fancy* arguments that the representational systems we talk and think in are in large part compositional; you find the effects of their compositionality just about wherever you look.

I must apologize for having gone on at such length about the arguments pro and con conceptual compositionality: the reason I've done so is that, in my view, the status of the statistical theory of concepts turns, practically entirely, on this issue. And statistical theories are now the preferred accounts of concepts practically throughout cognitive science. In what follows I will take the compositionality of conceptual repertoires for granted, and try to make clear how the thesis that concepts are prototypes falls afoul of it.

Why Concepts Can't Be Prototypes[12]

Here's why concepts can't be prototypes: whatever conceptual content is, compositionality requires that complex concepts inherit their contents from those of their constituents, and that they do so in a way that explains their productivity and systematicity. Accordingly, whatever is *not* inherited from its constituents by a complex concept is *ipso facto* not the content of that concept. But: (i) indefinitely many complex concepts have no prototypes; a fortiori they do not inherit their prototypes from their constituents. And, (ii) there are indefinitely many complex concepts whose prototypes aren't related to the prototypes of their constituents in the ways that the compositional explanation of productivity and systematicity requires. So, again, if concepts are compositional then they can't be prototypes.

In short, *prototypes don't compose*. Since this is the heart of the case against statistical theories of concepts, I propose to expatiate a bit on the examples.

(i) *The Uncat Problem*

For indefinitely many "Boolean" concepts,[13] *there isn't any prototype* even though:

> their primitive constituent concepts all have prototypes,

and

> —the complex concept itself has definite conditions of semantic evaluation (definite satisfaction conditions).

So, for example, consider the concept NOT A CAT (*mutatis mutandis*, the predicate 'is not a cat'); and let's suppose (probably contrary to fact) that CAT isn't vague; i.e. that 'is a cat' has either the value *S* or the value *U* for every object in the relevant universe of discourse. Then, clearly, there is a definite semantic interpretation for NOT A CAT; i.e. it expresses the property of *not being a cat*, a property which all and only objects in the extension of the complement of the set of cats instantiate.

However, although NOT A CAT is semantically entirely well behaved on these assumptions, it's pretty clear that it hasn't got a stereotype or an exemplar. For consider: a bagel is a pretty good example of a NOT A CAT, but a bagel couldn't be NOT A CAT's prototype. Why not? Well, if bagels are the prototypic NOT A CATs, it follows that the more a thing is like a bagel the less it's like a cat; *and the more a thing isn't like a cat, the more it's like a bagel*. But the second conjunct is patently not true. Tuesdays and erasers, both of which are very good examples of NOT A CATs, aren't at all like bagels. An Eraser is not more a Bagel for being a bad Cat. Notice that the same sort of argument goes through if you are thinking of stereotypes in terms of features rather than exemplars. There is nothing that non-cats qua non-cats as such are likely to have in common (except, of course, not being cats).[14]

[13] To simplify the exposition, I'll use this notion pretty informally; for example, I'm glossing over the distinction between Boolean *sentences* and Boolean *predicates*. But none of this corner-cutting is essential to the argument.

[14] This is not to deny that there are typicality effects for negative categories; as Barsalou remarks, "with respect to *birds*, *chair* is a better nonmember than is *butterfly*" (1987: 101). This observation does not, however, generalize to Boolean functions at large. I doubt that

The moral seems clear enough: the mental representations that correspond to complex Boolean concepts specify *not their prototypes but their logical forms*. So, for example, NOT A CAT has the logical form *not*(*F*), and the rule of interpretation for a mental representation of that form assigns as its extension the complement of the set of *F*s. To admit this, however, is to abandon the project of using prototype structure to account for the productivity (/systematicity) of complex Boolean predicates. So be it.

(ii) *The Pet Fish Problem*

Prototype theories want to explicate notions like *falling under a concept* by reference to notions like *being similar to the concept's exemplar*. Correspondingly, prototype theories can represent conceptual repertoires as compositional only if (barring idioms) a thing's similarity to the exemplar of a complex concept is determined by its similarity to the exemplars of its constituents. However, this condition is not satisfied in the general case. So, for example, a goldfish is a poorish example of a fish, and a poorish example of a pet, but it's a prototypical example of a pet fish. So similarity to the prototypic pet and the prototypic fish doesn't predict similarity to the prototypical pet fish. It follows that if meanings were prototypes, then you could know what 'pet' means and know what 'fish' means and still not know what 'pet fish' means. Which is just to say that if meanings were prototypes, then the meaning of 'pet fish' wouldn't be compositional. Various solutions for this problem are on offer in the literature, but it seems to me that none is even close to satisfactory. Let's have a quick look at one or two.

Smith and Osherson (1984) take prototypes to be matrices of weighted features (rather than exemplars). So, for example, the prototype for APPLE might specify a typical shape, colour, taste, size, ripeness, . . . etc. Let's suppose, in particular, that the prototypical apple is red, and consider the problem of constructing a prototype for PURPLE APPLE. The basic idea is to form a derived feature matrix that's just like the one for APPLE, except that the feature *purple* replaces the feature *red* and the weight of the new colour feature is appropriately increased. PET FISH would presumably work the same way.

It's pretty clear, however, that this treatment is flawed. To see this, ask yourself *how much* the feature *purple* weighs in the feature matrix for PURPLE APPLE. Clearly, it must weigh more than the feature *red* does in the matrix for APPLE since, though there can be apples that aren't red,

purple has to weigh *infinitely* much in the feature matrix for PURPLE APPLE because *purple apples are purple*, unlike *typical apples are red*, is a *logical* truth.

So the Smith/Osherson proposal for composing prototypes faces a dilemma: either treat the logical truths as (merely) extreme cases of statistically reliable truths, or admit that the *weights* assigned to the features in derived matrices aren't compositional *even if the matrices themselves are*. Neither horn of this dilemma seems happy. Moreover, it's pretty clear what's gone wrong: what really sets the weight of the *purple* in PURPLE APPLE isn't the concept's prototype; *it's the concept's logical form*. But prototypes don't have logical forms.

Another way to put the pet fish problem is that the 'features' associated with the As in AN constructions are not, in the general case, independent of the features associated with the Ns. So, suppose that the prototype for NURSE includes the feature *female*. Pace Smith and Osherson's kind of proposal, you can't derive the prototype for MALE NURSE just by replacing *female* with *male*; all sorts of other things have to change too. This is true even though the concept MALE NURSE is 'intersective'; i.e. even though the set of male nurses is the overlap of the set of males with the set of nurses (just as the set of pet fish is the overlap of the set of pets with the set of fish). I want to stress this point because prototype theorists, in their desperation, are sometimes driven to suggest that MALE NURSE, PET FISH, and the like *aren't* compositional after all, but it's all right that they aren't, since they are idioms. But surely, *surely*, not. What could be stronger evidence against PET FISH being an idiom or for its being compositional than that it entails PET and FISH and that {PET, FISH} entails it?

It's perhaps worth mentioning the most recent attempt to salvage the compositionality of prototypes from pet fish, male nurses, striped apples, and the like (Kamp and Partee 1995). The idea goes like this: maybe good examples of striped apples aren't good examples of striped things *tout court* (compare zebras). But, plausibly, a prototypic example of a striped apple would *ipso facto* be as good an example of something striped *as an apple can be*. That is a way of saying that the relevant comparison class for judging the typicality of a sample of apple stripes is not the stripes on things at large but rather the stripes on other apples; it's these that typical apple stripes are typical *of*. In effect, then, what you need to do to predict whether a certain example of apple stripes is a good example of apple stripes, is to "recalibrate" STRIPES to apples.

A fair amount of algebra has recently been thrown at the problem of

and Partee 1995; Osherson and Smith 1996). But, as far as I can see, the undertaking is pointless. For one thing, it bears emphasis that the appropriate information for recalibrating a complex concept comes from the world, not from the content of its constituents. If it happens that they paint fire engines in funny shades of red, then typical fire engine red won't be typical red. To decide whether the colour of a certain engine is typical, you'd therefore need to recalibrate RED to FIRE ENGINE; and to do that, you'd need to know the facts about what shades of red fire engines are painted. Nothing about the concepts RED or FIRE ENGINE, per se, could tell you this; so nothing about these concepts, per se, could predict the typicality of a given sample of fire-engine red. In this sense, "recalibrated" compositionality, even if we knew how to compute it, wouldn't really *be* compositionality. Compositionality is the derivation of the content of a complex concept *just* from its structure and the content of its constituents; *that's why* compositionality explains productivity and systematicity.

Still worse, if possible: identifying the relevant reference set for a complex concept itself depends on a prior grasp of its compositional structure. In the case of STRIPED APPLE, for example, the reference set for the recalibration of STRIPE is the striped apples. How do we know that? Because we know that STRIPED APPLE applies to is the intersection of the striped things and the apple things. And how do we know *that*? Because we know the compositional semantics of STRIPED APPLE. Computing typicality for a complex concept by "recalibrating" its constituents thus *presupposes* semantic compositionality; it presupposes that we already know how the content of the concept depends on the content of the concept's constituents. So, recalibration couldn't be what makes concepts compositional, so it couldn't be what makes them systematic and productive. So what is recalibration *for*? Search me.

By the way, these pet fish sorts of arguments ramify in ways that may not be immediately apparent; compositionality is a sharp sword and cutteth many knots.[15] For example, it's very popular in philosophical circles (it's the last gasp of Empiricist semantics) to suppose that there are such things as 'recognitional concepts'; RED and SQUARE, for example, and likewise, I suppose, DOG and TREE, and many, many others. Peacocke 1992 is a *locus classicus* for this thesis, but any philosopher who says there are 'criteria' for the application of a concept is likely to be intending to claim that the concept is recognitional. All told, that includes quite a lot of philosophers and quite a lot of concepts.

A concept is recognitional, in the intended sense, only if the ability to identify its instances in favourable circumstances is among its concept-constitutive possession conditions. Thus, being able to identify squares is part and parcel of having the concept SQUARE; it's constitutive of the content hence of the identity of the concept. So the story goes. Notice that having SQUARE doesn't require the ability to identify any and every square (consider a square as big as the universe). Likewise, somebody could be thoroughly a possessor of the concept BIRD and none the less not know whether to apply it to ostriches (to say nothing of pterodactyls). So the story must be (indeed, is) that having a recognitional concept requires being able to recognize good (clear, paradigmatic, etc.) instances of the concept. You don't have BIRD unless you are inclined to take sparrows and the like to be birds.

But, now, the pet fish/striped apple/male nurse worries return full force. If, in particular, nothing is constitutive of conceptual content unless it composes, then recognitional capacities can't be constitutive of conceptual content. For someone could have the appropriate recognitional capacities with respect to FISH (he sees at a glance that trout, tuna, and the like are fish) and could have the appropriate recognitional capacities with respect to PET (he sees at a glance that poodles, Siamese kittens, and the like are pets), but be quite at a loss to identify even paradigmatic pet fish (e.g. even goldfish) as such. Because *being a paradigm* doesn't compose, recognitional capacities don't compose either. So the same argument that shows that paradigms aren't constituents of content shows that recognitional capacities aren't either; hence that there aren't any recognitional concepts. Compositionality is a sharp sword which cutteth many knots. (Or have I mentioned that?)

The long and short: either concepts qua prototypes aren't compositional or, if they are, their compositionality is parasitic upon concepts qua something other than prototypes. Conceptual contents, however, *must* be compositional; nothing else could explain why concepts are systematic and productive. So concepts aren't prototypes. This is too sad for words. A theory of concepts has two things to explain: how concepts function as categories, and how a finite mind can have an infinite and systematic conceptual capacity. Prototypes do a not-bad job of explaining the first (though, notoriously, they're not so good at penguins and ostriches being birds). Anyhow, they do noticeably better than definitions. But they are *hopeless* at the second job; so I am claiming.

It may occur to you, however, that my evidence for this claim has thus far consisted exhaustively of the enumeration of counter-examples; and it

fish; but it's a profound methodological principle (owing, I believe, to Jim Higginbotham) that for technical problems there are technical solutions. Maybe there is, after all, some way around the apparent failures of prototypes to compose? Given all the evidence that people do have prototypes, isn't the identification of prototypes with concepts a programme that's worth persisting in? Surely, the proper response to a counter-example is to explain it away? Or simply to ignore it?

That is a methodology with which I am deeply sympathetic. But it doesn't apply in the present case since there is independent reason to doubt that the examples of failures of prototypes to compose are merely apparent. It's not just that, prima facie, the identification of contents with prototypes fails for certain cases; it's that there's a pretty convincing diagnosis of the failures which, if correct, shows why the project *can't* succeed. Here's the diagnosis.

Prototype theories of conceptual content are, as we've seen, instances of inferential role theories of conceptual content. Their only fundamental argument with the classical, definitional version of IRS is over *which* inferences are content-constitutive: classical theorists say it's the defining ones, prototype theories say that it's the statistically reliable ones. But so long as IRS is common ground for everyone concerned, this is an argument that the classical theorists are bound to win. That's because, except for definitional inferences, *inferential roles themselves don't compose*.

Compositionality says that, whatever content is, constituents must yield theirs to their hosts and hosts must derive theirs from their constituents. Roughly, the first half is required because whatever is true of cows as such or of brown things as such is *ipso facto* true of brown cows. And the second half is required because, if the content of BROWN COW is *not* fully determined by the content of BROWN and the content of COW (together with syntactic structure), then grasping BROWN and COW isn't sufficient for grasping BROWN COW, and the standard explanation of productivity is undone.

Now, complying with the first half of this constraint is easy for IRS since BROWN contributes to BROWN COW not only its *content-constitutive* inferences (whichever those may be), but *every* inference that holds of brown things in general.[16] If whatever is a cow is an animal, then brown cows are animals a fortiori. If whatever is brown is square, then, a fortiori, every brown cow is a square cow.

But the second half of the compositionality constraint is tricky for an

[16] If *all* of BROWN's inferential role is content constitutive, so be it: BROWN

IRS. If nothing can belong to the content of BROWN COW except what it inherits either from BROWN or from COW, then the content of BROWN COW *can't* be its *whole* inferential role. For, of course, all sorts of inferences can hold of brown cows (not qua brown or qua cows but) simply as such. That's because all sorts of things can be true of brown cows that aren't true either of brown things in general or of cows in general; that they are brown cows is an egregious example.

If an *X*-kind of inference is required to be such that constituents contribute all their *X*-inferences to their hosts, and hosts inherit their *X*-inferences only from their constituents, then only *defining* inferences will do as candidates for *X*: the inferential role of a complex concept is exhaustively determined by the inferential roles of its constituents *only* with respect to its defining inferences.[17] That statistical inferences fail to compose is just a special case of this general truth. The pet fish problem is therefore not a fluke. Either the classical, definitional version of IRS is right, or no version can be.

So here's the impasse: prototypes are public (i.e. they are widely shared) and they are psychologically real, so they do meet two of the non-negotiable conditions that concepts are required to meet; but they aren't compositional. Definitions would be compositional if there were any, but there aren't, so they're not. As things stand, *there is no version of the inferential role theory of conceptual content for which compositionality and psychological reality can both be claimed.* I think there must be something wrong with inferential role theories of content.

A modest proposal:

"All right, all right; but if constituent concepts don't contribute their definitions or their prototypes to their complex hosts, what *do* they contribute?"

Duck soup. *They contribute what they mean*; e.g. the properties that they express. What PET contributes to PET FISH is the property of *being a pet*; what FISH contributes to PET FISH is the property of *being a fish*. It's because PET contributes *pet* to PET FISH and FISH contributes *fish* to PET FISH that PET FISH entails PET and FISH. And it's because *pet* and *fish* exhaust the content of PET FISH that {PET, FISH} entails PET FISH. There are, to be sure, hard cases for this sort of analysis (what do RISING and TEMPERATURE contribute to THE RISING

[17] More precisely, only with respect to *conceptually necessary* inferences. (Notice that neither nomological nor metaphysical necessity will do; there might be laws about brown cows per se, and (who knows?) brown cows might have a proprietary hidden essence.) I don't know what a Classical IRS theorist should say if it turns out that conceptually

TEMPERATURE?), but they are just the cases that are hard for compositionality on *any* known view.

"Oh bother, why didn't *I* think of that!"

Presumably because the metaphysics that you had in mind says that meaning is constituted by inferential roles; in which case, the present proposal is no better off than the ones that we've just been discussing. By contrast, informational semantics contemplates the metaphysical possibility that there should be something that a concept means (e.g. a property that it expresses) even though the concept enters into *no constitutive inferential relations at all*. My advice is, therefore: if you want to say what compositionality appears to require you to—that what a concept contributes to its hosts is what it means you'd better mean by 'what it means' not its inferential role but something like *the information that it carries*, where, by assumption, RED carries information about *redness*.

Inferential role semantics is bankrupt. Because cognitive science has swallowed Inferential Role Semantics whole, its treatment of concepts is bankrupt too; it keeps writing cheques on a theory of meaning that isn't there. It is *very naughty* to write cheques that you can't cash, and it's past time for cognitive science to kick the habit. Chapters 6 and 7 will be about that.

APPENDIX 5A
Meaning Postulates

Prototypes dissociate two issues that definition theories treat together: *What is the structure of a lexical concept?* and *What modal inferences do you have to accept to have the lexical concept X?* On the definition story, both these questions get answered by reference to the relations between concepts and their parts: lexical concepts typically have constituent structure, much like phrasal concepts; and if the concept *C* is a constituent of the concept *X*, then you don't have *X* unless you believe that *X*s are *necessarily C*s. The argument between definitions and prototypes is over the second of these claims.

But it's worth noting that the question whether lexical concepts have constituent structure can be dissociated from *both* the question whether inferences constitute content and whether what makes an inference content-constitutive is something about its modality. *Inferential role*

inferences which constitute a concept's content are defined over its constituent structure.

There may be several motivations for separating the question whether (and which) inferences constitute content from the question whether typical lexical concepts are structurally complex. Some philosophers do so because they want to hold on to intuitions of analyticity in face of the mounting empirical evidence that lexical concepts generally behave like atoms by either linguistic or psychological criteria. And there's an independent, semantical argument as well; it's known in the lexical semantics literature as the 'residuum problem'.

In the most familiar cases, lexically governed inferences are supposed to follow from definitions by an analogue to simplification of conjunction. Thus, 'bachelor' entails *unmarried* because its definition is '*male and unmarried*' and the 'and' works in the usual truth-conditional way. This treatment fits naturally with the idea that concepts are bundles of semantic features, each of which express a property of the (actual or possible) things that the concept subsumes.

Now, it's natural to assume that if there is a property corresponding to the feature bundle '$F_1, F_2, \ldots, F_n$', then there should also be a property corresponding to the bundle '$F_1, F_2, \ldots, F_{n-1}$'. So, for example, what's left when you take the *unmarried* out of the definition of 'bachelor' is the definition of 'male'; and what's left when you take the *male* out of the definition of 'bachelor' is the definition of 'unmarried'. Just as the result of simplifying a conjunctive predicate is always itself a predicate, so the result of simplifying a feature bundle is always itself a feature bundle.

But there are cases of lexically governed entailment which appear not to follow this model; 'red → colour' is a paradigm. According to the definition story, this inference should be the simplification of a complex concept (the definition of 'red') which has the form: 'F_1, . . ., COLOUR, . . .'; but, on reflection, it's hard to see what could go in for the 'F_1'. A male is something that is just like a bachelor but not necessarily married; but what is just like red but not necessarily a colour? If you take the 'COLOUR' out of the definition of 'red', what you're left with *doesn't seem to be a possible meaning*; the residuum of 'red → coloured' is apparently a surd. Or, to put it the other way round, it looks like the only thing that could combine with 'COLOURED' to mean *red* is 'RED'. That, however, can't be what the lexical semanticist is proposing. To have 'RED' in the definition of 'red' would make 'COLOUR' redundant, since if 'RED' means *red*, it *thereby* entails 'COLOUR'. If the definition of 'red' includes RED, that's *all* it includes, so in effect the proposal that it does

sensory concepts. It's sometimes suggested that they illustrate the presence of an "iconic" element in concepts like RED (see the discussion above of Jackendoff 1992). Maybe 'red' means something like 'similar in respect of colour to this' where the 'this' ostensively introduces a red sample. The trouble with taking this line, however, is that the pattern RED and the like exemplify actually appears to be quite general: lots of lexical concepts for which *definitions* are very hard to find nevertheless appear to enter into the same sort of "one way" entailments that hold between 'red' and 'colour'. It's plausible that 'dog' means *animal*, but there doesn't seem to be any *F* (except DOG) such that '*F* + ANIMAL' means *dog*. 'Chair' means *furniture*, but what and FURNITURE means *chair*? Notice that it won't do to appeal to 'iconic elements' in these non-sensory cases. Maybe 'red' means '*similar in colour to this*', but 'dog' doesn't mean '*similar in X to this*' for any *X* that I can think of except *dog*hood. It appears that, contrary to traditional Empiricist doctrine, many lexical items are not independent but not definable either; 'red' entails 'colour' but can't be defined in terms of it.

A natural way to accommodate the residuum problem is to allow that some content-constitutive inferences don't arise from definitions after all. It's not that RED entails COLOUR because the definition of 'red' is COLOUR & *F*; rather, RED just entails COLOUR full stop. Following the historical usage, I'll call a principle of inference that institutes a 'one way' relation of entailment between lexical concepts a "meaning postulate". Rules of lexically governed inference that happen to be biconditional, like 'bachelor ↔ unmarried man', have no special status according to the theory that meaning postulates are what license lexically governed inferences. This version of Inferential Role Semantics is therefore *weaker* than the definitional account; the latter allows a lexical concept to enter into constitutive inferential relations *only if it is definable*.

From our perspective, the important consequence of this liberalization is that it disconnects the question whether an inference from *C* to *C* content-constitutive from the question whether C_1 is a syntactic part of Notice that it was only because definitions were required to be biconditional that they *could* be viewed as exhibiting the structural description of a concept. UNMARRIED MAN can't be the structural description BACHELOR unless 'BACHELOR' and 'UNMARRIED MAN' den the same concept. But BACHELOR and UNMARRIED MAN can't the same concept unless 'BACHELOR ↔ UNMARRIED MAN' is t

Detaching the question whether RED entails COLOUR from question whether COLOUR is a constituent of RED has its virtues, t

one to give up such claims while holding onto both "'red' means *colour* is analytic' and 'you don't have RED unless you know that red is a colour'. On the meaning postulate story, RED → COLOUR could be meaning-constitutive even if neither RED nor COLOUR have *any* internal structure; i.e. even if it's atomic.

But no free lunch, of course. We started out this chapter by remarking that one of the nicest things about the definition story was that it explains an otherwise striking and perplexing symmetry between the metaphysics of meaning and the metaphysics of concept possession: *the very inferences that are supposed to define a concept are also the ones you have to accept in order to possess the concept*. This really is striking and perplexing and not at all truistic; remember, it isn't (can't be) true of *all* necessary inferences or even of all a priori inferences that they determine the conditions for possessing the concepts involved in them. Well, the theory that concepts are definitions gets this symmetry for free; it follows from the fact that definitions relate concepts *to their constituents*. If *C* is literally a part of C_1, then *of course* you can't have C_1 unless you also have *C*. Notice that this explanation turns on precisely the idea that meaning postulates propose to abandon: viz. that the content-constitutive inferences are the ones that relate a concept to its parts.

In short, if you are *independently* convinced *both* that there are meaning-constitutive inferences *and* that most lexical concepts behave like primitives, you've got a residuum problem to which meaning postulates may indeed offer a solution. But at a price, since the solution weakens the architecture of your overall theory: it breaks the connection between the structure of a concept and its possession conditions.

Partee has tried bravely to make a virtue of this necessity:

> Meaning postulates might be a helpful tool . . . since they make the *form* [*sic*] of some kinds of lexical information no different in kind from the form of some kinds of general knowledge. That would make it possible to hypothesize that the very same 'fact' for example, whales are mammals—could be stored in either of two 'places,' a storehouse of lexical knowledge or a storehouse of empirical knowledge; whether it's part of the meaning of 'whale' or not need not be fixed once and for all. (1995: 328)

But it is inadvisable for a theory to recognize degrees of freedom that it is unable to interpret. Exactly because meaning postulates break the 'formal' relation between belonging to the structure of a concept and being among its constitutive inferences, *it's unclear why it matters* which box a given such 'fact' goes into; i.e. whether a given inference is treated as

other. Are any further differences between these minds entailed? If so, which ones? Is this wheel attached to anything at all?

It's a point Quine made against Carnap that the answer to 'When is an inference analytic?' can't be just 'Whenever I feel like saying that it is'. Definition versions of IR Semantics can hold that an inference is analytic when and only when it follows from the structure of a concept. If the meaning postulate version has an alternative proposal on offer, it's not one that I've heard of.

APPENDIX 5B
The 'Theory Theory'[18] of Concepts

The theories of concepts discussed so far all presuppose Inferential Role Semantics, so they all owe an account of which inferences determine conceptual content. The big divides are between holism (which says that all inferences do) and some sort of molecularism (which says that only some inferences do); and, within the latter, between classical theories (according to which it is modality that matters to content constitution) and prototype theories (according to which it's empirical reliability that does). In effect, the various theories of concepts we've reviewed are versions of IRS distinguished, primarily, by what they say about the problem of individuating content.

Now, a quite standard reading of the history of cognitive science has the reliability-based versions of IRS displacing the modality-based versions and in turn being displaced, very recently, by theory theories.[19] But that way of telling the story is, I think, mistaken. Though theory theories do propose a view about what concepts are (or, anyhow, about what concepts are like; or, anyhow, about what a lot of concepts are like). they don't, as far as I can tell, offer a distinct approach to the content individuation problems. Sometimes they borrow the modality story from definitional theories, sometimes they borrow the reliability story from prototype theories, sometimes they share the holist's despair of individuating concepts at all. So, for our purposes at least, it's unclear that theory theories of concepts differ substantially from the kinds of theories

[18] I'm not crazy about this terminology, if only because it invites conflation with the quite different issue whether "folk psychology" is a (tacit) theory (see, for example, Gordon 1986). But it's standard in the cognitive science literature so I'll stick with it, and from here on I'll omit the shudder-quotes.

[19] For a relatively clear example of a discussion where theory theories are viewed as *alternatives* to probabilistic accounts of concepts, see Keil 1987. See also Keil 1991, where

of concepts that we've already reviewed. Hence the relatively cursory treatment they're about to receive.

The basic idea is that concepts are like theoretical constructs in science *as the latter are often construed by post-Empiricist philosophers of science*. The caveat is important. For example, it's not unusual (see Carey 1991; Gopnik 1988) among theory theorists to postulate 'stage-like discontinuities' in conceptual development, much as Piagetians do. But, unlike Piaget, theory theorists construe the putative stage changes on the analogy of—perhaps even as special cases of—the kinds of discontinuities that 'paradigm shifts' are said to occasion in the history of science. The usual Kuhnian morals are often explicitly drawn:

> the concepts of the new and old theory and of the evidential description are incommensurab[le].(Gopnik 1988: 199)

> Asking whether or not the six-month-old has a concept of object-permanence in the same sense that the 18-month-old does is like asking whether or not the alchemist and the chemist have the same concept of gold, or whether Newton had the same concept of space as Einstein. These concepts are embedded in complex theories and there is no simple way of comparing them. Moreover, particular concepts are inextricably intertwined with other concepts in the theory. (Ibid.: 205)

It should be clear how much this account of conceptual ontogenesis relies on a Kuhnian view of science. It isn't just that if Kuhn is wrong about theory change, then Gopnik is wrong about the analogy between the history of science and conceptual development. It's also that key notions like *discontinuity* and *incommensurability* aren't explicated within the ontogenetic theory; the buck is simply passed to the philosophers. "It may not resolve our puzzlement over the phenomena of qualitative conceptual change in childhood to point out that there are exactly parallel paradoxes of incommensurability in science, but at this stage we may see the substitution of a single puzzling phenomenon for two separate puzzling phenomena as some sort of progress" (Gopnik 1988: 209). Correspondingly, however, if you find the idea that a scientific theory-change is a paradigm shift less than fully perspicuous, you will also be uncertain what exactly it is that the ontogenetic analogy asserts about stages of conceptual development. Your response will then be a sense less of illumination than of *déjà vu*.

If Gopnik finds some solace in this situation, that's because, like Kuhn, she takes IRS not to be in dispute.[20] The putative "problem of incommens-

urability" is that *if* the vocabulary of a science is implicitly defined by the theories it endorses, it's hard to see how the theories can correct or contradict each other. This state of affairs might be supposed to provide a precedent for psychologists to appeal to who hold that the minds of young children are incommensurably different from the minds of adults. Alternatively, it might be taken as a reductio of the supposition that the vocabulary of a science is implicitly defined by its theories. It's hard to say which way one ought to take it barring some respectable story about *how* scientific theories implicitly define their vocabularies; specifically, an account that makes clear which of the inferences that such a theory licenses are constitutive of the concepts it deploys. And there's no point in cognitive scientists relying on the philosophy of science for an answer to this question; the philosophy of science hasn't got one. It seems that we're back where we started.

In short, it may be that the right moral to draw from the putative analogy between scientific paradigms and developmental stages is that the ontogenesis of concepts is discontinuous, just like scientific theory-change. Or the right moral may be that, by relativizing the individuation of concepts to the individuation of theories, IRS makes a hash of *both* cognitive development *and* the history of science.

If there is any positive account of conceptual content that most theory theorists are inclined towards, I suppose that it's holism.[21] I don't, however. know of any attempt they have made seriously to confront the objections that meaning holism is prone to. Two of these are particularly relevant The first is familiar and quite general (see Chapter 1 and Fodor and Lepore 1992) and I won't go on about it here. Suffice it that if the individuation of concepts is literally relativized to whole belief systems. then no two people, and no two time slices of a given person, are ever subsumed by the same intentional generalizations, and the prospects for robust theories in intentional psychology are negligible.

probabilistic distributions of properties in the world" (1987: 196). But that's true only on the assumption that theories somehow constitute the concepts they contain. Ditto Keil's remark that "future work on the nature of concepts . . . must focus on the sorts of theories that emerge in children and how these theories come to influence the structure of the concepts that they embrace" (ibid.).

[21] There are exceptions. Susan Carey thinks that the individuation of concepts must be relativized to the theories they occur in, but that only the basic 'ontological' commitments of a theory are content constitutive. (However, see Carey 1985: 168: "I assume that there is a continuum of degrees of conceptual differences, at the extreme end of which are concepts embedded in incommensurable conceptual systems.") It's left open how basic ontological claims are to be distinguished from commitments of other kinds, and Carey is quite aware that problems about drawing this distinction are depressingly like the analytic/synthetic

But I do want to say a word or so about the second objection, which is that holism about content individuation doesn't square with key principles of the theory theory itself. Consider, in particular, the idea that new concepts get introduced, in the course of theory change, by a kind of implicit theoretical definition. In all the examples I've heard of, a theory can be used to effect the implicit definition of a new term only if at least some of its vocabulary is *isolated* from meaning changes of the sorts that holists say that concept introduction brings about. That's hardly surprising. Intuitively, implicit definition determines the meaning of a new term by determining its inferential relations to terms in the host theory that are *presumed to be previously understood*. It is, to put it mildly, hard to see how this could work if introducing a new concept into a theory *ipso facto* changes what all the old terms mean. For then the expressions by reference to which the neologism is introduced aren't 'previously understood' after all: they are just *homophones* of the previously understood expressions.[22]

Consider, for a familiar example, the introduction by implicit definition of a logical constant like '$\vee$'. The idea is that to determine that '$\vee$' has the same sense as the (truth conditional, inclusive) English 'or', it's sufficient to stipulate that:

$$P \rightarrow P \vee Q$$
$$(P \vee Q)\ \&\ {\sim}P \rightarrow Q$$

But the plausibility of claiming that these stipulations determine that '$\vee$' means 'or' depends on supposing that they preserve the standard interpretations of '&' (= conjunction), '~' (= negation), and '$\rightarrow$' (= truth-functional implication). That, however, implies that the interpretation of '&', '~', and '$\rightarrow$' must be assumed to be isolated from whatever meaning changes adding '$\vee$' to the host theory is supposed to bring about; an assumption that is contrary, apparently, to the holist thesis that the semantic effects of theory change reverberate throughout the vocabulary of the theory. (I say that it's 'apparently' contrary to the holist thesis because I know of no formulation of semantic holism that is precise enough to yield unequivocal entailments about which changes of theory effect which changes of meaning.)

[22] This point is related, but not identical, to the familiar worry about whether implicit definition can effect a 'qualitative change' in a theory's expressive power: the worry that definitions (implicit or otherwise) can only introduce concepts whose contents are already expressible by the host theory. (For discussion, see Fodor 1975.) It looks to me that implicit

This isn't just a technical problem; texts that flout it tend to defy coherent exegesis. Consider, for one example among very many, Gopnik's suggestion[23] that

> An 'object' is a theoretical entity which explains sequences of what (for lack of a better term) we might call object-appearances at the evidential level . . . At the very earliest stage infants seem to have a few rules about the relations between their own actions and object-appearances, for example, infants seem to know that objects disappear when you turn away from them and reappear when you turn back to them. (1988: 205)

(and so forth, *mutatis mutandis*, for further 'rules' that the child gets later).

How are we to interpret this passage? Notice the tell-tale aporia (where are you, Jacques Derrida, now that we need you?). The rule with which the infants are credited is said to be about "relations between their own actions and object-*appearances*" (my emphasis). But, when an instance of such a rule is offered, it turns out to be that "*objects* [my emphasis] disappear when you turn away from them". Question: what does 'objects' mean in this rule? In particular, *what does it mean to the infant* who, we're supposing, learns the concept OBJECT by a process that involves formulating and adopting the rule?[24] If it means object-*appearances*, then (quite aside from traditional worries about how an *appearance* could *reappear*) it doesn't do what Gopnik wants; since it specifies a relation *among* object-appearances, it doesn't give the infant information about the relation between *objects* and object-appearances.

So, maybe 'object' means *theoretical entity which explains sequences of what (for lack of a better term) we might call object-appearances at the evidential level*. I rush past the implausibility of claiming that infants have to have that much ontology (in particular, that much *dubious* ontology) in order to learn quotidian object-concepts like CHAIR. I'm a nativist too, after all. The more pressing problem for a theory theorist is: if *that's* what 'object' means in the infant's rule, *in what sense are there discontinuities in the development of the infant's object-concept*? On this reading of the text, it looks like what the infant has– right from the start and right to the finish – is a concept of an object that's much like Locke's: objects are unobservable kinds of things that cause experiences. Correspondingly cognitive development consists of learning more and more about things

[23] I don't particularly mean to pick on Gopnik; the cognitive science literature is full examples of the mistake that I'm trying to draw attention to. What's unusual about Gopnik treatment is just that it's clear enough for one to see what the problem is.

[24] As usual, it's essential to keep in mind that when a de dicto intentional explanati

this kind (e.g. that when you turn your back on one, it ceases to cause appearances in you . . . etc.).[25] What, then, has become of the discontinuity of the object-concept? In particular, what's become of the *incommensurability* of the infant's object-concept with grown-up Gopnik's? It turns out that Gopnik can, after all, say *exactly* what (according to her theory) the infant's earliest concept of an object is: it's the concept of *a theoretical entity which explains sequences of* . . . etc. . . . *and which ceases to cause appearances in you when you turn your back on it* . . . etc.

I suppose what Gopnik really ought to say, if she wants to be true to the implicit definition picture, is that the concept of an object is that of 'AN *X* WHICH . . .', and that cognitive development consists in adding more and more relative clauses. But it's hard to see why such a thesis would count as construing concept development as discontinuous. And, anyhow, it's hard to see how it could be swallowed by a meaning holist. Isn't meaning holism, by definition, committed to there *not* being a notion of content identity that tolerates the addition of *new* information to the same *old* concept?

The local moral, to repeat, is that maybe you can make sense of concept introduction as implicit theoretical definition, and maybe you can make sense of meaning holism. But it's very unclear that you can make sense of both at the same time. The general moral is that, if the theory theory has a distinctive and coherent answer to the 'What's a concept?' question on offer, it's a well-kept secret.

I should add, in minimal fairness, that it's not clear that theory theorists are really all that interested in what concepts are. Certainly it's often hard to tell whether they are from what they say. For example, Medin and Wattenmaker (1987; see also Murphy and Medin 1985) undertake to "review evidence that suggests concepts should be viewed as embedded in theories" (34–5), a thesis which they clearly regard as tendentious, but which, as it is stated, it's hard to imagine that anyone could disagree with. What I suppose they must have in mind is that concepts are somehow *constituted* (their identity is somehow determined) by the theories in which they are embedded. But that claim, though tendentious enough, doesn't amount to a new account of conceptual content; unless the 'somehows' are somehow cashed, it just reiterates IRS.

The situation in Medin and Wattenmaker is especially confusing because its so hard to figure out what they think that the theory theory is a theory of; they are explicit that it's supposed to provide an account of the

[25] If "object" means *thing that causes appearances* then, of course, the rule isn't that

"coherence" of concepts, but it's far from clear what they think conceptual coherence is. At one point, having suggested that the theory theory should provide "guidelines concerning which combinations of features form possible concepts and which form coherent ones" (1987: 30), they offer, as an example of an incoherent concept, "bright red, flammable, eats mealworms, found in Lapland, and used for cleaning furniture". So it sounds as though the question about conceptual coherence that the theory theory answers is: What's wrong with this and other such concepts?

But it's hard to believe that *is* the question since the answer, though perfectly obvious and entirely banal, is one that Medin and Wattenmaker don't even consider. What's wrong with the concept BRIGHT RED. FLAMMABLE, EATS MEALWORMS, . . . etc. is that, as far as anybody knows, there's nothing that is, or would be, true of things in virtue of their falling under it (except what follows trivially from their falling under it; e.g that they are, or would be, found in Lapland). In particular, there are no substantive, counterfactual-supporting generalizations about such things: so why on earth would anybody want to waste his time thinking about them? Compare such unsatisfied (but coherent) concepts as UNICORN. At least there's a *story* about unicorns. That is, there are interesting things that are *supposed to be* true about them: that their ground-up horns are antidotes to many poisons; that if there were unicorns, virgins could catch them if there were virgins, and so on. In short, such examples as Medin and Wattenmaker offer suggest that being 'coherent' isn't even a *psychological* property: the incoherence of BRIGHT RED, FLAMMABLE, . . . etc. is a defect not of the concept but of the world. It's therefore hard to see why a psychologist should care about it (though perhaps a zoologist might).

Or perhaps Medin and Wattenmaker have some other construal of conceptual coherence in mind; but search me what it is.[26]

To return to the main theme: many of the typical preoccupations of theory theorists seem to be largely neutral on the issue of concept

[26] See also Keil: "Prototypes merely represent correlated properties, they offer n· explanation of the reasons for those correlations (e.g. why the prototypical features of birds. such as beaks, feathers, and eggs tend to co-occur)" (1987: 195). The suggestion seems to be that the difference between prototype theories and theory theories is that the latter entail that having a concept involves knowing the explanation of such correlations (or knowing that there is an explanation? or knowing that some expert knows the explanation?). But, if so, it seems that theory theories set the conditions for concept possession impossibly high. I'm pretty confident that being liquid and transparent at room temperature are correlated properties of water. But I have no idea *why* they are correlated. Notice, in particular, that learning that *being water* is *being* H_2O didn't advance my epistemic situation in this respect

individuation—Is conceptual change discontinuous? What makes a concept coherent? Are children metaphysical essentialists?—and the like. There is, to be sure, much that's of interest to be said on these topics. But, thank Heaven, not here. From our point of view, the crucial question is whether, when a theory theorist says that concepts are typically embedded in theoretical inferences, he means to claim that knowing (some or all) of the theory is a necessary condition for having the concept. If yes, then the 'which inferences' question has to be faced. If no, then some positive account of concept possession/individuation is owing. The definition story and the prototype story are bona fide competing theories of concepts because they do have answers to such questions on offer. As far as I can make out, the theory theory doesn't, so it isn't.