THE RED HERRING AND THE PET FISH

WHY CONCEPTS *STILL* CAN'T BE PROTOTYPES

Jerry Fodor and Ernest Lepore

Department of Philosophy and Center for Cognitive Science, Rutgers University

Introduction.[1]

There is a Standard Objection to the idea that concepts might be prototypes (or exemplars, or stereotypes): Because they are productive, concepts must be compositional. Prototypes aren't compositional, so concepts can't be prototypes (see, e.g., Margolis, 1994).[2]

However, two recent papers (Osherson and Smith, 1988; Kamp and Partee, 1995) reconsider this consensus. They suggest that, although the Standard Objection is probably right in the long run, the cases where prototypes fail to exhibit compositionality are relatively exotic and involve phenomena which *any* account of compositionality is likely to find hard to deal with; for example, the effects of quantifiers, indexicals, contextual constraints, etc. KP are even prepared to indulge a guarded optimism: "... when a suitably rich compositional theory... is developed, prototypes will be seen ... as one property among many which only when taken altogether can support a compositional theory of combination" (p.56).

In this paper, we argue that the Standard Objection to prototype theory was right after all: The problems about compositionality are insuperable in even the most trivial sorts of examples; it is therefore as near to certain as anything in cognitive science ever gets that the structure of concepts is not statistical. Theories of categorization, concept acquisition, lexical meaning and the like, which assume the contrary simply don't work.

We commence with a general discussion of the constraints that an account of concepts must meet if their compositionality is to explain their productivity. We'll then turn to a criticism of proposals that OS2 and KP make for coping with some specific cases.

Part I: Productivity and compositionality.

Within the family of theories that identify concepts with mental representations (Representational Theories of Mind (RTMs)), there is a more or less explicit consensus that concepts are productive, and that their productivity is explained by the assumption that mental representations (MRs) are compositional.[3] We will assume, as do the authors we're discussing, that some version of this story is correct. Our first aim is to provide a minimal sketch of the explanatory architecture that it presupposes.

Here's the general idea: A compositional theory of the productivity of concepts must, at a minimum, specify two functions:

-A *Composition Function* (FC), which maps a finite basis of simple MRs onto an infinity of complex MRs together with their structural descriptions.

-An *Interpretation Function* (FI), which maps arbitrary MRs, simple or complex, into their semantic interpretations. There is, alas, no general consensus on what sorts of things semantic interpretations are; but examples of candidates for the interpretations of general concepts include properties, sets, intensions, senses and functions. In any case, it's assumed that the semantic interpretations of MRs are typically `things in the world' and not themselves mental or linguistic.

We want to make clear what justifies postulating each of these functions:

Why you need the composition function: If concepts are MRs and concepts are productive, there must be infinitely many MRs. Practically without exception, people who accept this inference conclude that infinitely many MRs must have internal structure; specifically, that infinitely many MRs must have MRs as their constituents.

Just what the argument for this is supposed to be isn't, after all, entirely obvious. What logical or metaphysical principle prohibits the existence of infinitely many *un*structured mental particulars? On the other hand, if the argument from the productivity of concepts to the internal structure of MRs isn't demonstrative, it is nevertheless extremely well evidenced. It is assumed by all the psychological and semantic accounts of productivity

(and systematicity; see Fodor and Pylyshyn, 1988) that anybody has thus far been able to imagine. And the intuitive plausibility of the idea that, for example, the concept BROWN is a constituent of the concept BROWN COW appears undeniable.

So, then, according to this line of theorizing, concepts are productive because there are infinitely many MRs. There are infinitely many MRs because new, relatively complex MRs can be constructed by using old, relatively primitive ones as their constituents. That MRs have constituent structure is thus essential to explaining the compositionality of concepts; accordingly, FC serves to specify the constituency relations that MRs enter into.

Why you need the interpretation function (FI): What RTMs offer to reconstruct the pretheoretic concept CONCEPT is *MR with its semantic interpretation.* As remarked above, semantic interpretations are assumed to be typically nonmental. Correspondingly, the interpretation function specifies a relation between MRs and things in the world (for example, between MRs and their extensions).

Though MR theorists are lost to solipsism or idealism from time to time (see, e.g., Jackendoff, 1992), that some mind/world relation is essential to concept individuation is, in fact, pretty widely understood. In particular, it is common ground that concepts are the kinds of things that *apply* to things (that is, they're the kinds of things that can function as *categories*) and that it is constitutive of the identity of a concept that it applies to the things that it does. Nothing that applies to bricks, or that fails to apply to birds, could be the concept BIRD.

Essentially all philosophers who discuss these sorts of issues take this for granted. (However, see Stich, 1983) But so too do practically all psychologists, at least implicitly. Most of the experimental data on which the discussion of prototypes turns use categorization tasks to assess concept possession. That is, they assume that whether one has the concept C is revealed, at least in part, by one's capacity to distinguish the things that C applies to from the things that it doesn't. This research strategy would be incoherent if it weren't assumed that its relation to its domain of application is among a concept's essential properties. Correspondingly, the interpretation function FI is required to specify this relation for each of the infinitely many MRs.

The argument so far is that everybody who wants to explain the productivity of concepts by reference to the compositionality of MRs has to postulate a composition function and

an interpretation function. Notice, however, that assuming the mere (as it were, Platonic) existence of FC and FI isn't good enough for the task at hand. The productivity problem is not just that there are (Platonically) infinitely many concepts; it's that (given the usual idealizations) there are infinitely many concepts *that people can entertain*. But, how could it follow from the mere Platonic existence of FC and FI that people can entertain infinitely many concepts? Compare: The facts of arithmetic don't, in and of themselves, explain how people are able to add. You also need some psychological premises about what they know and what is going on in their heads. Correspondingly, explaining the productivity of people's concepts requires postulating not only that there are the functions FI and FC, but also that people are epistemically related to these functions in appropriate ways; that people can *grasp* these functions. What explains the productivity of our concepts is that we grasp the functions FI and FC.

It's generally assumed that you can only grasp a function that is finitely specifiable.[4] Notice that, unless this *is* assumed, it's hard to see why productivity is a problem. One generates the productivity problem by asking how a finite creature could have an infinite epistemic capacity: how there could be infinitely many concepts that it can entertain. Clearly, either the existence of such infinite capacities is unproblematic, in which case the productivity problem doesn't arise; or, if there really is a productivity problem, the solution must not itself presuppose epistemic relations to infinite sets.

So, then, if the story is that we can entertain an infinity of concepts because we can grasp FI and FC, these functions must themselves be finite objects. This doesn't mean, of course, that they must have finite *extensions*. Rather, the idea is that a finite creature can get into an epistemic relation to an infinite set only by being in some epistemic relation to a finite object that specifies the set. Similar considerations suggest that each mental representation must itself be finitely specifiable, that the primitive basis from which complex MRs are constructed must be finite, and so forth. As remarked above, the consensus on this sort of point appears to be general among MR theorists who accept the productivity of concepts as a real phenomenon.

Our present concern is with a further consequence that an account of their graspability places on FI and FC. Principle P provides a rough formulation:

P: The interpretation that FI assigns to a certain MR must be computed from the structural description that FC assigns to that MR.

Let's continue to assume, for purposes of exposition, that semantic interpretations are sets (specifically, extensions). Then, presumably, FI assigns to the mental representation BROWN COW the intersection of the set of brown things with the set of cows. However, P further requires that FI does so *because* FC assigns to BROWN COW a structure which includes the constituent representations BROWN and COW (in, of course, the appropriate configuration). To put it slightly differently, the operations FI performs must be sensitive to the structural descriptions that FC enumerates, so that the structure of the interpretation that FI assigns derives from the structure that FC assigns.

To see what is at issue, consider what would happen if we had a semantical theory for MRs that failed to satisfy P. One could perhaps imagine that such a theory somehow succeeds in getting the `right' extensions assigned to each of the infinitely many MRs. So, the set of cows gets assigned to COW, the set of brown cows gets assigned to BROWN COW, etc. Technically, such a theory would succeed in representing the productivity of the MRs; it would represent them as an infinite set of interpreted objects. But it would nonetheless leave certain glaring explanatory gaps in the resulting explanation of why concepts are productive.

For one thing, we would be at a loss to explain what the constituent structure of MRs is *for*. Constituent structure would be idle in the general case in just the way that it really is idle in the case of idioms. (Compare the semantically irrelevant constituent structure of `kick the bucket' (= *die*) with the semantically relevant constituent structure of `brown cow' and BROWN COW.) Second, it would fail to explain why a given MR has the interpretation that it does (or, equivalently, why the isomorphism between the structure of MRs and the structure of their interpretations is reliable). By contrast, when principle P is enforced, we can see straight off why an MR that has COW as a constituent has as its interpretation a set all the members of which are cows (as in the case of BROWN COW) or members of the complement of the set of cows (as in the case NOT COW), etc. Failing P, a theory represents this sort of parallelism as accidental.

So much for what we take to be common assumptions; we're about to see that even MR theorists who talk as though perhaps they don't accept one or another of them actually do so in practice. We turn now to our main topic, which is how theories of the productivity of concepts fare when they take MRs to be prototypes, and hence take prototypes to constitute both the domain and the range of FC, and the domain over which the operations

of FI are defined.

We will consider two kinds of objections that an account of the compositionality of prototypes appears to face if it is to take the general form we've just been considering. According to the first objection, prototype theory can't account for certain relations of logical equivalence among concepts. According to the second, prototype theory can't predict the semantic relations between complex concepts and their constituents. We claim that both these objections are warranted.

Part II: Boolean concept.

We start by considering the sort of complex concepts that are built up from their basic constituents by the use of such Boolean operators as AND, IFTHEN, OR, and NOT. There are two sorts of problems such concepts raise for prototype theories. We'll argue that the first is a sort of red herring in that neither the problem itself, nor the sort of solution KP propose, are specific to prototype theories. The second is more serious since it appears to jeopardize principle P.

The red herring: truth value gaps.

Whether or not concepts are prototypes, it's clear that many concepts are vague; in many cases, no definite truth value attaches to the judgment that a certain concept applies. For example: It's surely just true that tables and chairs are furniture, and it's surely just false that fish are furniture; but what about wall-to-wall carpets? It's not mandatory, but it's natural, to say that there's no fact of the matter here; viz., that the judgment that wall-to-wall carpets are furniture is neither true nor false.[5]

If concepts are prototypes, then whether something falls under a concept is a matter of how similar to the prototype it is. Since SIMILARITY is itself presumably a vague concept, so too are whatever concepts are defined in terms of it. So, if vague concepts (inter alia) are prototypes, it is intelligible that judgments in which they occur will often exhibit `truth value gaps (tv-gaps). Notice that the suggested account of tv-gaps is metaphysical, not epistemic. An epistemic treatment of vagueness would claim that there *is* a fact about whether wall-to-wall carpeting is furniture, though it's a fact we don't happen to be apprised of; maybe future research will decide. We don't find it easy to take the epistemic view of vagueness seriously, and we won't discuss it in what follows.

Now, if a concept exhibits a tv-gap, then it will contribute that gap to infinitely many complex concepts of which it is a constituent. So, for example, if there is no fact of the matter about whether wall-to-wall carpets are furniture, then there is likewise no fact of the matter about whether wall-to-wall carpets are expensive furniture (viz., whether they are both furniture and expensive). And there is also no matter of fact about whether wall-to-wall carpets are furniture or not furniture, and so forth. But this sort of consequence is arguably not tolerable. Excluded middle says that *everything* is either furniture or not furniture, and excluded middle is a law of logic, and laws of logic are necessarily true. So, it looks as though, if you assume that concepts are prototypes, you will be forced to deny a necessary truth.

Whether or not you find this line of thought convincing, it bears emphasis that the problem being raised really has nothing in particular to do with prototypes as such or even with vagueness as such. It comes up wherever you suppose that tv-gaps can occur; and, prima facie, tv-gaps can have all sorts of sources: failed presuppositions, empty names, vacuous predicates, the truth paradoxes and so on. Correspondingly, we're about to see that the mechanisms that KP propose for dealing with predicates like `is wall-to-wall-carpet that either is furniture or is not furniture' do not, in any way, exploit the characteristic properties of prototypes.

We propose to scant the details. In effect, supervaluation assigns arbitrary, stipulative truth values to those base clauses of Boolean concepts which lack them (see below.) Given that supervaluations allow such tv-gaps to be filled, the logical truths can then be identified in the usual way; viz., as the sentences that remain true however tvs are assigned to their constituents. The logical forms that concepts are ascribed, according to the supervaluation treatment, are exactly what one would suppose on classical assumptions about what concepts are. (For example CARPET THAT EITHER IS FURNITURE OR IS NOT FURNITURE is assumed to have a mental representation of the form *c which is F or not-F.)* Since the classical notion of logical form is respected by supervaluation theory, all the necessary truths of classical Boolean logic can be preserved.

Here is the essential point: The supervaluation treatment allows the recovery of logical truths because it assigns classical logical forms to Boolean concepts. For exactly that reason, it is indifferent whether Boolean concepts are vague because their constituents are prototypes or whether it's something else that makes them vague. In consequence,

assuming that supervaluation works as a treatment of the vagueness of Boolean concepts if they are constructed from prototypes, it will work equally well, and in exactly the same way, if Boolean concepts are vague but *not* constructed from prototypes.

For example, they might be vague because they are definitions couched in a vague metalanguage. (BROWN COW = BROWN & COW, and BROWN is vague.) If supervaluation cures the sort of vagueness that prototypes cause, it also cures the sort of vagueness that vague definitions cause, and does so equally well and in exactly the same way (viz., by `precisifying'; see below). The moral is that vagueness is *everybody's* problem. Supervaluation may solve it, but the way it solves it doesn't favor prototype theory over any account of concepts which allows that they are often vague.

Having stressed the neutrality of the supervaluation treatment of vagueness in respect of the nature of concepts, we can't resist adding that we have pretty severe doubts whether the supervaluation treatment of vagueness actually works; in particular, whether assuming supervaluation as a semantics for tv-gaps will lead to the desired result that `John is bald or John is not bald' and the like will get appropriate truth values. The problems are technical and a full treatment would go beyond the scope of the present discussion. But a quick sketch may serve to supply the intuition.

For present purposes, the basic idea employed in applying supervaluation theory to the analysis of vagueness is that of a *precisification* of a vague language L. A precisification is an assignment of T and F to the atomic sentences of L such that: definitely false sentences are assigned F, definitely true sentences are assigned T, and vague sentences are ascribed T or F arbitrarily.[6] Then a complex sentence *s* is `definitely true' if it receives T under every precisification, and it is false if it receives F under every precisification. Just as you'd expect, a *logical* truth is a sentence that comes out true on *every* assignment of tvs to its atomic sentences, the assignments made by precisification included. So, as KP put it, supervaluation theory may provide "a sound logical framework in which prototype theory and classical logic can peacefully coexist." (ms p.22) Consider `John is bald or John is not bald,' and suppose there is no fact of the matter whether the atomic sentence `John is bald' is true. There are two ways of precisifying: assign `John is bald' T or assign it F. `John is bald or John is not bald' comes out true on either of these assignments, so the usual truth functional construal of `or' and `not' make `John is bald or John is not bald' a logical truth. Analogous treatment applies to (e.g.,) `John is bald or not bald'.

Here, however, is the problem: supervaluation theory involves stipulating assignments of truth values to sentences which, by assumption, have no truth value in the actual world (for example, to atomic sentences with vague predicates). This seems perfectly all right when it is *contingent* that a sentence has a tv-gap. For example, `John is bald' may have no truth value as a matter of fact, but we can perfectly well make sense of counterfactual hypotheticals which assume that it is true: e.g., `If John weren't bald, he would be happier than he is'. But now, there are lots of cases where if a sentence has a truth value gap at all, then it has it *necessarily*. Consider, for example, `The man who is taller than he is, is a Russian.' Since there can't be a man who is taller than he is, the presupposition of this sentence is necessarily false; so if failure of presupposition entails a tv-gap (as we may suppose), then this sentence is necessarily without truth value. But the question now arises: What could it mean to stipulate a truth value for a sentence which lacks a truth value *necessarily*? What, for example, is the force of counterfactuals with antecedents like `If the man who is taller than himself were Russian, then....'? We doubt that there's any sense to be made of such hypotheticals.

But if one can't assign a truth value to `The man who is taller than himself is Russian,' we can't assign T (or, of course, any other tv) to `Either the man who is taller than he is is Russian or the man who is taller than he is is not Russian'. This latter, however, appears to have the logical form *a is F or a is not F*, so it has to come out true if supervaluation theory is to recover the Law of Excluded Middle. Something appears to have gone wrong with the theory.

Precisely the same sort of difficulties are produced by vagueness assuming, as seems extremely plausible, that some vague sentences lack truth values necessarily. So, suppose that there's no fact of the matter about whether wall-to-wall carpeting is furniture. Then, surely, the fact that there is none is not itself contingent; if it's true in our world, then it's true in every world. If you're inclined to think that it's merely contingent that there's no fact of the matter whether wall-to-wall carpets are furniture, ask yourself what discovery about the world (or about English, for that matter) would convince you that you that, by gosh, you were wrong and that it really is (or really isn't) furniture after all. Similarly with examples like `Someone who has precisely 38 hairs is bald'; if it has a truth value gap at all, then it has its truth value gap necessarily.

But now the previous argument applies: If a sentence lacks a truth value necessarily, then it can't be precisified; if it can't be precisified, then no truth value can be assigned to

complex sentences in which it occurs, including ones which have the form of logical truths. So the program of using supervaluation theory to construct a semantics of vagueness that preserves the logical truths apparently fails.

Boolean concepts continued: The status of Principle P.

Our second point is that, whether or not prototypes generate tv-gaps in Boolean concepts, there's another, and quite different problem about assimilating the semantics of such concepts to prototype theory. Namely, that for indefinitely many Boolean concepts, there *isn't any* prototype even though:

(i) the primitive constituent concepts all have prototypes, and

(ii) the complex concept itself has definite truth conditions.

So, for example, consider the predicate `isn't a cat'; and let's suppose (probably contrary to fact) that `cat' is NOT vague; i.e., `is a cat' has either the value T or the value F for every object in the relevant universe of discourse. Then, clearly, there is a definite semantic interpretation for `is 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 `isn't a cat' is entirely well behaved on these assumptions, it pretty clearly has no stereotype; and nor do indefinitely many other Boolean complex concepts. There isn't any stereotypic nonprime number, and there isn't anything that is stereotypically pink if it's square. And so on. This is a point that KP recognize explicitly (cf. circa p. 48).

We remark, in passing, that this difficulty does not depend on a proprietary reading of `prototype'; for example, it holds whether you think that prototypes are something like feature sets, as OS2 explicitly does, or whether you think that prototypes are something like exemplars. KP aren't entirely explicit about which of these notions of prototype they have in mind, but on balance it seems to be the second.

To return to the main theme: there are indefinitely many cases in which there is no prototype corresponding to a complex Boolean concept; a fortiori, the MR corresponding to such a concept isn't a prototype. Faced with this problem, a theorist might just give up and admit that in at least indefinitely many cases, what a primitive concept transmits to its complex hosts is not its prototype; and that, in such cases, the identification of MRs with

prototypes is simply false. The obvious elaboration of this view is that the MRs corresponding to such complex concepts specify *not* their prototypes but their logical forms, and that their interpretations are computed from their logical forms in the standard Classical way. So, the function FC assigns to `isn't a cat' the logical form *not(F)*, and the rule of interpretation for an MR of that form assigns as its extension the complement of the set of Fs. This is, of course, to abandon the project of using prototype structure to account for the productivity of complex Boolean predicates. So be it.

We are seriously unclear whether, or to what extent, that is the course of action that KP endorse; their text often suggests that they have a rather different treatment in mind. Suppose you were to give up not the idea that MRs are prototypes but rather Principle P, according to which the semantic value of a concept is computed from the corresponding MR. You could then grant that the interpretation of `not(F)' isn't computed from a prototype, but argue that is compatible with prototype theory's account of MRs. That's because the interpretation of `(not)F' *isn't computed from its MR*. In fact, as far as the process of interpreting it is concerned, it needn't be assumed that `not(F*)' even has* an MR. "Consider [the concept] *red*. The concept *(is) not red* does not appear to have a prototype; for how might one resolve the choice among *white, green, black, yellow* and all the other colors that *red* excludes? Nevertheless, the degree of membership in the concept *not red* is [sic] a matter of prototypicality. Only, the relevant prototype is not some prototype for *not red* but the prototype for *red,* and the degree to which something is not red is a matter of how *little* rather than how much, it resembles that prototype" (p.48). (Similarly for conjunction, disjunction and the rest; cf. p.49.)

Notice that, on this account, if the prototype for `red' is a fire-engine in `That's red', then it's a fire-engine `That's not red' too, and `red' contributes the *same* extension in both cases; viz., it contributes the set of things that are sufficiently similar to fire engines. However, `red' does not contribute its extension to `not red' *by contributing the prototype of `red' to the prototype of `not red*'. That's the sense in which P is violated.

We remarked above that there seem to us to be decisive reasons for holding onto P. We won't repeat them here. The point we want to emphasize is that it's unclear that giving up P is even a *coherent* alternative to giving up `MRs are prototypes'. KP tell us that, in computing an interpretation for `not red', "the relevant prototype is not some prototype for *not red* but the prototype for *red,* and the degree to which something is not red is a matter of how *little* rather than how much, it resembles that prototype." But the question

arises how they know that this is so. More precisely: How does the computation that assigns an interpretation to the formula "not red" know that it's the prototype for "red" (and not, say, the prototype for "green" or "soup" or "transcendental") that it should consult when it does so? The answer surely must be that the MR it computes over is not the prototype for "red" but rather a representation of the logical form of "not red". But this means (a) that P is still in force; i.e., the interpretation of "not red" is computed from the MR of "not red," *not* from the MR of "red"; and (b) the MR of "not red" is a logical form, *not* a prototype. In effect, though their text rather suggests the contrary, what KP have really opted for is the alternative that makes all complex Boolean predicates *counterexamples* to prototype theory.

Part III: Pet Fish.

There is another kind of case, discussed in both SO2 and KP, in which it is apparently not possible to provide the correct interpretation of a complex predicate given just its structure and the prototypes of its primitive constituents. The problem here is not that the complex concept fails to have a prototype, as in many of the Boolean cases, but rather that an object's similarity to the prototype for a complex concept seems not to vary systematically as a function of its similarity to the prototypes of the constituents concepts. So, for example, a goldfish is a poorish example of a fish, and a poorish example of a pet, but it's quite a good example of a pet fish.

Now, according to prototype theory, to have a concept is to have its prototype together with a measure of the distance between the prototype and an arbitrary object in the domain of discourse; in effect, this distance measure is the form that FI takes in prototype versions of computational theories of mind. Prima facie, however, the distance of an arbitrary object from the prototypic pet fish is *not* a function of its distance from the prototypic pet and its distance from the prototypic fish. In consequence, knowing that PET and FISH have the prototypes that they do does not permit one to predict that the prototypical pet fish is more like a goldfish than like a trout or a herring, on the one hand, or a dog or a cat, on the other. But if prototypes aren't compositional, then, to put it mildly, the identification of concepts with prototypes can't explain why concepts are productive.

Both OS2 and KP offer solutions for this problem, but it seems to us that neither is even close to satisfactory. We'll review them very briefly.

As we remarked above, OS2 takes prototypes to be matrices of weighted features (rather than as exemplars).[7] So, for example, the prototype for `apple' might specify a typical shape, color, 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 that you form a derived matrix that's just like the one for apple, except that the feature *purple* replaces the feature *red* and the weight of the new feature is increased. Pet fish presumably work the same way.[8]

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 feature matrix for APPLE since, though there can be apples that aren't red, there can't be purple apples that aren't purple (any more than there can be red apples that aren't red or purple apples that aren't apples.) In effect, purple has to weigh *infinitely* much in the feature matrix for PURPLE APPLE because `purple apple ---> purple' is a *logical* truth. So the theory faces a dilemma: either treat the logical truths as (merely) extreme cases of statistically reliable truth, or admit that the *weights* assigned to the features in derived matrices aren't compositionally determined even if the features themselves are. Neither horn of this dilemma seems happy. What really sets the weight of the PURPLE in PURPLE APPLE isn't its prototype; it's its logical form.

And, even if the treatment weren't flawed, it is clearly not general. The problem is that the `features' associated with the Ns in AN constructions are not, in the general case, independent. So, suppose that the prototype for NURSE includes the feature *female*. You can't derive the prototype for MALE NURSE by just replacing *female* with *male;* all sorts of other things have to change too. Notice that this is true even though `male nurse' is (in KP's term) `intersective'; i.e., even though the set of male nurses is the overlap of the set of males with the set of nurses. Things go even worse for the OS2 proposal when one considers nonintersective concepts like STONE LION, DECOY DUCK, FAKE DIAMOND, POTENTIAL PROVOST and so forth.

KP offer a more complicated analysis of the pet fish case, but it doesn't work well either. Scanting the details once again, the basic idea is that the failure of pet fish to be good examples of pets is a kind of context effect, analogous to the failure of big ants to be good examples of big things. KP think, plausibly enough, that the meaning of `big ant' is something like *big for an ant*, so that a really good example of a big ant would be

something that's as-good-an-example-of-something-big as an ant can be. Similarly, a really good example of a male nurse would be something that's as good an example of a nurse as something male can be; a really good example of a pet fish would be something that's as good an example of a pet as a fish can be, and so on.

It seems to us, however, that the assimilation of PET FISH to BIG ANT is clearly ill advised: `pet fish' entails *pet* but `big ant' does *not* entail *big*, and KP's proposal leaves this asymmetry entirely unexplained.

We're claiming, in effect, that if AN ---> A, that's a very strong reason to suppose that the interpretation of the A in AN is *not* affected by the interpretation of the N. For, suppose the contrary; suppose that the meaning of A were some how converted from *A* to *A'* when A is a constituent of AN; then AN shouldn't entail A but A'. What better argument could there be that `male' means *male* in `male nurse' than the necessity of `Male nurses are male'. (Analogously, what better argument could there be that the content of MALE is *male* than that `Males are male' is necessary? Indeed, what other argument have we got?)

It follows from this proposal that AN ---> A is a necessary condition for an adjective being intersective, hence that `big' is not intersective in `big ant'. KP, however, use a substitutivity test to decide on intersectivity: A is intersective in AN1 only if (`a is AN1' and `a is N2') ---> `a is AN2.' So, for example, `skillful' is not intersective according to KP since (`a is a skillful violinist' and `a is a plumber') does not entail `a is a skillful plumber.

You might think that the two tests for intersectivity should be coextensive and that `big' would fail to be intersective by either one.[9] However, KP argue that `big ant' *is* intersective but that the fact that it is is obscured by the (putative) context effect of `ant' on the interpretation of `big'. We find this doctrine hard to construe. If `big ant' is in the intersection of the big things with the ants, then `big ant ---> big' must be valid, which, however, intuition denies. Indeed, on KP's own analysis, a big ant ought to be (not in the set of big things but) in the set of things that are big *for ants*. But the set of things that are big for ants isn't included in the set of things that are big *tout court*. In fact, `big ant' doesn't look to be intersective on any interpretation that we can think of.[10]

The bottom line, then, is that PET FISH is a counterexample to the compositionality of prototypes, and there is no reason at all to suppose that the problem it raises would be

solved by whatever mechanism it is that the semantics employs to cope with BIG ANT. We want to emphasize that, quite aside from the technical issues, this conclusion really does seem quite plausible. The reason you can't derive the PET FISH prototype given the PET prototype and the FISH prototype, is simply that what *kinds of fish people keep as pets is a fact about the world, not a fact about concepts or language.* It is therefore possible to be perfectly clear what `pet fish' means, and yet have no idea which pet fish are prototypical. Which pet fish are prototypical is something you *just have to go out and learn*. The *language* (/concept) assures you that the prototypical pet fish is a pet and a fish, just as the language assures you that the prototypical big ant is big for an ant. After that, you're on your own.

Conclusion: prototypes are fish out of water.

The prototype for PET FISH is, as it were, an idiom; a merely linguistic (/conceptual) inquiry will tell you that pet fish are fish, but no merely linguistic (/conceptual) inquiry will tell you which pet fish are prototypical. Putting it that way might, however, suggest that there is some hope for the prototype theorist after all. Why shouldn't he just admit that his story about the compositionality of concepts doesn't work for PET FISH, but argue that the reason it doesn't is precisely that PET FISH *is* an idiom. On anybody's story, idioms are expressions where the interpretation that the compositional semantics predicts is, as it were, over-ridden by special conventions that must simply be acquired case by case. Having a green thumb turns out not to be having a thumb that is green, compositional semantics to the contrary notwithstanding. Why, then, shouldn't the semantics say that a paradigm pet fish ought to be a paradigm pet and a paradigm fish? If that prediction is wrong, that only shows that `pet fish' isn't compositional.

So, then, here's the proposal: The prototypical ANs are the intersection of the prototypical A things with the prototypical N things *in the unmarked case*. But you default to the unmarked case only if you do not have specific information to the contrary. Just as the semantics of English supports the inference that green thumbs are green and thumbs, so the semantics of concepts supports the inference that prototypical pet fish are prototypical pets and prototypical fish. It's just that when you've learned which fish it is that people actually do keep as pets, you learn to override the inference that the semantics supports.

The idea that composition works on prototypes to deliver default values is maybe a little counterintuitive; first blush, neither `pet fish' nor `big ant' seem plausible candidates for

idioms. But it might nonetheless be a hard theory to refute. That's because it is required to predict default to the compositional prototype only in cases where no real world beliefs are over-riding. Failures to default (PET FISH, BIG ANTS, and so forth) can thus be viewed as prima facie indications that such over-riding has indeed occurred. Correspondingly, a clear test of the theory would require examining cases which there is independent reason to suppose that *only* linguistic (/conceptual) knowledge is employed in drawing an inference. It is, however, notoriously difficult to construct such cases, the vagaries of the analytic/synthetic distinction being what there are.

So we can't prove that defaulting to the compositional prototype isn't the strategy that people actually do follow. But it's easy to see that it shouldn't be because it's an irrational strategy. That it is irrational follows from two considerations, rough formulations of which go as follows:

i. All else equal, the more complexly modified a concept is, the *less* you are likely to have special knowledge about the things in its extension.

So, for example, I know a little about cows qua cows. But I know next to nothing about brown cows qua brown cows (except, of course, that they are brown and cows); and I know literally nothing about brown cows owned by people whose last names start with `W' qua brown cows owned by people whose last names start with `W' (except, of course, that they are brown, and cows and owned by people whose last names start with `W').

The upshot is that if my strategy is to default to the compositional prototype when I have no special information to the contrary, then the more heavily modified a concept is, the more likely I am to default to its compositional prototype.

Notice, however, that (ii) is also true:

ii. All else equal, the more heavily modified a concept is, the *less* likely that its prototype is predicted by the prototypes of its constituents.

So, for example, pet fish aren't good bets for satisfying the pet prototype; but still less so are pet fish that live in Armenia; and still less so are pet fish who live in Armenia and have recently swallowed their owners.... And so forth. This is actually quite close to a point we made above: the `features' that prototypical Ns exhibit are not, in the general case,

independent of one another. In consequence, the more A's you put on to modify the N, the more likely it is the prototypical ANs will *not* exhibit the features that the prototypical Ns do.

If you now put (i) and (ii) together, it's clear why defaulting to the compositional prototype is an irrational strategy: on the one hand, the more complexly modified a concept is, the more likely it is that (i) will require you to default to the compositional prototype. On the other hand, the more complexly modified a concept is, the more (ii) makes it likely that defaulting to the compositional prototype will give you the wrong interpretation. It is, however, irrational to employ a strategy if the more likely you are to use it, the more likely it is to fail.

The Bottom line.

Prototypes aren't compositional; they work like idioms. Concepts, however, *must* be compositional; nothing else could explain why they are 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 conceptual capacity. Prototypes do a not-bad job of explaining the first (though, notoriously, they're not so good at penguins being birds; see also Armstrong et al., 1983). Anyhow, they do noticeably better than definitions. But they are *hopeless* at the second job. In fact, what a constituent concept contributes to its host appears to be precisely *necessary conditions* and not *statistical correlates*; `pet' contributes *pet* to `pet fish', and not, for example, *furry and cuddly*; `bachelor' contributes *hasn't a spouse* to `elderly bachelor,' and not, for example, *has a live-in girl friend*; `big' contributes `big for an N' to `big N', and not, for example, *heavy* or *hard to kill by stepping on*. Etc. The penultimate line is: *it is not in virtue of their statistical properties that concepts are compositional.*

And the bottom line is: *nobody knows what makes concepts compositional, so nobody knows what concepts are.*

Afterthought.

While we were writing this paper, an article appeared by (Huttenlocher and Hedges, 1994) that proposes a statistical model for the formation of complex prototypes, one that relies on the assumption of independence of features. "In constructing a conjoint category.... the

listener must make some assumption about the form of the relation between values of the constituent categories. The assumption that the categories are independent is an obvious one.... If the assumption of independence holds, the formal, mechanism we have proposed is applicable" (p.163). However, as we've just been seeing, the more complex a concept is, the more likely the `listener' must depend on its internal structure (rather than his background knowledge) to decode it; and the more complex the concept is, the more the assumption that the features its constituents contribute are independent is likely *not* to be true (so long as feature assignments express statistical correlates rather than necessary conditions). The wary listener will therefore either avoid the strategy that Huttenlocher and Hedges commend, or avoid decoding complex concepts by assigning them prototypes.

Huttenlocher and Hedges refer to data that suggest that subjects do in fact reliably default to the independence assumption. We suspect that if this is so, it shows only that subjects who are required to make guesses in experimental environments don't care much whether their guesses are true. It would be interesting to know what happens in situations where the outcomes matter more. How much are *you* prepared to bet that Mongolian Grey Geese satisfy the Goose prototype? (Compare: How much are you prepared to bet that Mongolian Grey Geese are Grey, Geese, and Mongolian?)

Notes

1. A word on notation: We use caps for names of concepts (as in `the concept COW'. We use italics for names of semantic interpretations (as in `the concept COW expresses the property of *being a cow*') and, occasionally, for names of semantic features. Since we assume that words and phrases express concepts (e.g., that the English phrase `brown cow' expresses the concept BROWN COW), and that the productivity of language is parasitic on the productivity of thought. we are thus explicit about drawing the language/thought distinction only when it matters to the discussion. Often we'll go back and forth between concepts and terms as convenience of exposition dictates.

2. For further discussion of the Standard Objection, see Fodor, 1981; Osherson and Smith, 1981. For a survey of the literature on prototype effects in categorization tasks, see Smith and Medin, 1981.

3. Some Connectionists apparently hold that conceptual repertoires are intrinsically *non*productive (viz., finite). For purposes of this discussion, they are beyond the pale.

4. Clearly, this condition must be satisfied whenever grasping a function requires its explicit internal representation.

5. An alternative ("fuzzy logic") treatment assumes that there are infinitely many truth values (tv's) between T and F, and that `wall-to-wall carpets are furniture' has one of them. However, Osherson and Smith (1981) have shown that this approach leads to its own kinds of difficulties in the case of concepts that are built out of the Boolean connectives.

6. In fact, some further "penumbral" requirements on the coherence of these `arbitrary' assignments are in force, but they needn't concern us here. For details, see Fine, 1975.

7. Like every other cognitive scientist we've encountered who uses the notion, OS don't say what a feature is, or why the concept FISH has more than the one feature *+fish*. (It's true, of course, that if you ask subjects to list some typical properties of fish they will hardly ever include *being fish*. But familiar Griceian considerations explain their not doing so. Notice that they usually don't list *existing*, or *being things* either, though both are features that prototypical fish exhibit.) Since, however, we have long abandoned hope of

figuring out what cognitive scientists might mean by `feature', we'll just take the notion for granted in what follows.

8. Though it's not entirely clear what one is supposed to do if the prototype for FISH doesn't have a parameter for degree of domesticatedness. Maybe the pet fish features are just the union of the pet features and the fish features.

9. It's not possible to prove this, of course, since what rules of inference are valid for AN construction is moot. But the following arguments seem to us pretty persuasive.

For any X, if N and X are substitutable, then if 'a is AN' is true, then 'a is AX' is true; that is, 'a is A' is true *whatever else is true of a*; that is, `a is A' is true tout court. So if KP's test is satisfied, so too is ours.

Likewise in the other direction: If `a is AN1' entails `a is A', then `a is AN1 & a is N2' entails `a is AN2' (assuming that conjunction introduction holds). So if our test is satisfied, so too is KPs. So it appears that the two tests are equivalent.

10. Our own view is that `big' and the like are best analyzed as "defined in use" (see Russell 1956. In effect, the relevant semantic rule interprets `big N' for variable N, to the instances of which it assigns all and only the Ns that are big for Ns. This seems compatible with everything KP say about the case, except for their claim that `big N' is intersective; which we find puzzling, as previously remarked. In any case, KP apparently concede that `big ant' doesn't entail `big' tout court, which is all that our argument in the text requires.

References

Armstrong, S.L., Gleitman, L.R. and Gleitman, H. (1983). "What some concepts might not be." *Cognition,* 13, 263-308.

Fine, K. (1975) "Vagueness, truth and logic," *Synthese* 30, 265-300.

Fodor, J. (1981). "The present status of the innateness controversy," in REPRESENTATIONS, MIT Press, Cambridge, MA.

Fodor, J. and Pylyshyn, Z. (1988). "Connectionism and cognitive architecture: A critical analysis," *Cognition,*28, 3-71.

Huttenlocher, J. and Hedges, L. (1994) "Combining graded categories," Psych. Rev. vol 101 no 1, 157-165.

Jackendoff, R. (1992). "The problem of reality," in LANGUAGES OF THE MIND, MIT Press, Cambridge, Mass.

Kamp, H. and Partee, B. (1995) (=KP). "Prototype theory and compositionality," forthcoming, *Cognition*.

Margolis, E. (1994) "A reassessment of the shift from the Classical Theory of concepts to Prototype theory," *Cognition, 51, 73-89.*

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

Osherson, D.N. and Smith, E.E. (1988) (=OS2). "Conceptual combination with prototype concepts," in Collins A. and Smith E. (eds.) READINGS IN COGNITIVE SCIENCE, Morgan Kaufman Publishers, Inc. San Mateo, CA.

Russell, B. (1956) "On denoting," in R. Marsh (ed.) LOGIC AND KNOWLEDGE, 39-56, Unwin Hyman, London.

Smith, E.E. and Medin, D.L. (1981). CATEGORIES AND CONCEPTS, Harvard University Press, Cambridge, MA.

Stich, S. (1983) FROM FOLK PSYCHOLOGY TO COGNITGIVE SCIENCE, MIT Press, Cambridge, Mass.