Cognition, 50 (1994) 95–113
0010-0277/94 Elsevier Science B.V.

Concepts: a potboiler

Jerry Fodor*
Graduate Center, CUNY, 33 West 42nd Street, New York, NY 10036, USA
Center for Cognitive Science, Rutgers University, Psychology Building, Busch Campus, Piscataway, NJ 08855, USA

Abstract

An informal, but revisionist, discussion of the role that the concept of a concept plays in recent theories of the cognitive mind. It is argued that the practically universal assumption that concepts are (at least partially) individuated by their roles in inferences is probably mistaken. A revival of conceptual atomism appears to be the indicated alternative.

Introduction: the centrality of concepts

What's ubiquitous goes unremarked; nobody listens to the music of the spheres (or to me, for that matter). I think a certain account of concepts is ubiquitous in recent discussions about minds; not just in philosophy but also in psychology, linguistics, artificial intelligence, and the rest of the cognitive sciences; and not just this week, but for the last fifty years or so. And I think this ubiquitous theory is quite probably untrue. This paper aims at consciousness raising; I want to get you to see that there is this ubiquitous theory and that, very likely, you yourself are among its adherents. What to do about the theory's not being true (if it's not) – what our cognitive science would be like if we were to throw the theory overboard – is a long, hard question, and one that I'll mostly leave for another time.

The nature of concepts is the pivotal theoretical issue in cognitive science; it's the one that all the others turn on. Here's why:

Cognitive science is fundamentally concerned with a certain *mind–world*

**Correspondence to:* J. Fodor, Center for Cognitive Science, Rutgers University, Psychology Building, Busch Campus, Piscataway, NJ 08855, USA.

SSDI 0010-0277(94)00618-U

relation; the goal is to understand how its mental processes can cause a creature to behave in ways which, in normal circumstances, reliably comport with its utilities. There is, at present, almost[1] universal agreement that theories of this relation must posit mental states some of whose properties are *representational*, and some of whose properties are *causal*. The representational (or, as I'll often say, *semantic*) properties of a creature's mental states are supposed to be sensitive to, and hence to carry information about, the character of its environment.[2] The causal properties of a creature's mental states are supposed to determine the course of its mental processes, and, eventually, the character of its behavior. Mental entities that exhibit both semantic and causal properties are generically called "mental representations", and theories that propose to account for the adaptivity of behavior by reference to the semantic and causal properties of mental representations are called "representational theories of the mind".

Enter concepts. Concepts are the least complex mental entities that exhibit both representational and causal properties; all the others (including, particularly, beliefs, desires and the rest of the "propositional attitudes") are assumed to be *complexes* whose constituents are concepts, and whose representational and causal properties are determined, wholly or in part, by those of the concepts they're constructed from.

This account subsumes even the connectionist tradition which is, however, often unclear, or confused, or both about whether and in what sense it is committed to *complex* mental representations. There is a substantial literature on this issue, provoked by Fodor and Pylyshyn (1988). See, for example, Smolensky (1988) and Fodor and McLaughlin (1990). Suffice it for present purpose that connectionists clearly assume that there are *elementary* mental representations (typically labeled nodes), and that these have both semantic and causal properties. Roughly, the semantic properties of a node in a network are specified by the node's label, and its causal properties are determined by the character of its connectivity. So even connectionists think there are concepts as the present discussion understands that notion.

On all hands, then, concepts serve both as the domains over which the most elementary mental processes are defined, and as the most primitive bearers of semantic properties. Hence their centrality in representational theories of mind.

[1]The caveat is because it's moot how one should understand the relation between main-line cognitive science and the Gibsonian tradition. For discussion, see Fodor and Pylyshyn (1981).

[2]There is no general agreement, either in cognitive science or in philosophy, about how the representational/semantic properties of mental states are to be analyzed; they are, in general, simply taken for granted by psychologists when empirical theories of cognitive processes are proposed. This paper will *not* be concerned, other than tangentially, with these issues in the metaphysical foundations of semantics. For recent discussion, however, see Fodor (1990) and references cited therein.

1. Ancient history: the classical background

The kind of concept-centered psychological theory I've just been sketching should seem familiar, not only from current work in cognitive science, but also from the philosophical tradition of classical British empiricism. I want to say a bit about classical versions of the representational theory of mind because, though their general architecture conforms quite closely to what I've just outlined, the account of concepts that they offered differs, in striking ways, from the ones that are now fashionable. Comparison illuminates both the classical and the current kinds of representational theories, and reveals important respects in which their story was closer to being right about the nature of concepts than ours. So, anyhow, I am going to argue.

Here's a stripped-down version of a classical representational theory of concepts. Concepts are mental images. They get their causal powers from their associative relations to one another, and they get their semantic properties from their resemblance to things in the world. So, for example, the concept DOG applies to dogs because dogs are what (tokens of) the concept looks like. Thinking about dogs often makes one think about cats because dogs and cats often turn up together in experience, and it's the patterns in one's experience, and only these, that determine the associations among one's ideas. Because association is the only causal power that ideas have, and because association is determined only by experience, any idea can, in principle, become associated to any other, depending on which experiences one happens to have. *Classical ideas cannot, therefore, be defined by their relations to one another.* Though DOG-thoughts call up CAT-thoughts, LEASH-thoughts, BONE-thoughts, BARK-thoughts and the like in most *actual* mental lives, there are *possible* mental lives in which that very same concept reliably calls up, as it might be, PRIME NUMBER-thoughts or TUESDAY AFTERNOON-thoughts or KETCHUP-thoughts. It depends entirely on how often you've come across prime numbers of dogs covered with ketchup on Tuesday afternoons.

So much by way of a reminder of what classical theorists said about concepts. I don't want to claim much for the historical accuracy of my exegesis (though it may be that Hume held a view within hailing distance of the one I've sketched; for purposes of exposition, I'll assume he did). But I do want to call your attention to a certain point about the *tactics* of this kind of theory construction – a point that's essential but easy to overlook.

Generally speaking, if you know *what an X is*, then you also know *what it is to have an X*. And ditto the other way around. No doubt, this applies to concepts. If, for example, your theory is that concepts are pumpkins, then it has to be a part of your theory that having a concept is having a pumpkin; and if your theory is that having a concept is having a pumpkin, then it has to be a part of your theory that pumpkins are what concepts are. I take it that this is just truistic.

Sometimes it's clear in which direction the explanation should go, and sometimes it isn't. So, for example, one's theory about *having* a cat ought surely to be parasitic on one's theory about *being* a cat; first you say what a cat is, and then you say that having a cat is just: *having one of those*. With jobs, pains, and siblings, however, it goes the other way round. First you say what is to *have* a job, or a pain, or a sibling, and then the story about what jobs, pains and siblings *are* is a spin-off.

These examples are, I hope, untendentious. But decisions about the proper order of explanation can be unobvious, important, and extremely difficult. To cite a notorious case: ought one first to explain what the number three is and then explain what it is for a set to have three members? Or do you first explain what sets are, and then explain what numbers are in terms of them? Or are the properties of sets and of numbers both parasitic on those of something quite else (like counting, for example). If I knew and I was rich, I would be rich and famous.

Anyhow, classical representational theories uniformly took it for granted that the explanation of *concept possession* should be parasitic on the explanation of *concept individuation*. First you say what it is for something to *be* the concept X – you give the "*identity conditions*" for the concept – and then the story about concept *possession* follows without further fuss. Well, but *how* do you identify a concept? Answer: you identify a concept by saying *what it is the concept of*. The concept DOG, for example, is the concept of *dogs*; that's to say, it's the concept that you use to think about dogs with. Correspondingly, *having* the concept DOG is just *having a concept to think about dogs with*.

Similarly, *mutatis mutandis*, for concepts of other than canine content: the concept X is the concept of *Xs*. *Having* the concept X is just *having a concept to think about Xs with*. (More precisely, having the concept X is having a concept to think about Xs "as such" with. The context "thinks about . . ." is intentional for the ". ." position. We'll return to this presently.)

So much for the explanatory tactics of classical representational theories of mind. Without exception, however, current theorizing about concepts reverses the classical direction of analysis. The substance of current theories lies in what they say about *the conditions for having the concept X*. It's the story about *being* the concept X – the story about concept *individuation* – that they treat as parasitic: the concept X is just *whatever it is that a creature has* when it has that concept. (See, for example, Peacocke, 1992, which is illuminatingly explicit on this point.) This subtle, and largely inarticulate, difference between contemporary representational theories and their classical forebears has, so I'll argue, the most profound implications for our cognitive science. To a striking extent, it determines the kinds of problems we work on and the kinds of theories that we offer as solutions to our problems. I suspect that it was a wrong turn – on balance, a catastrophe – and that we shall have to go back and do it all again.

First, however, just a little about why the classical representational view was abandoned. There were, I think, three kinds of reasons: methodological, metaphysical and epistemological. We'll need to keep them in mind when we turn to discussing current accounts of concepts.

Methodology: Suppose you're a behaviorist of the kind who thinks there are no concepts. In that case, you will feel no need for a theory about what concepts are, classical or otherwise. Behaviorist views aren't widely prevalent now, but they used to be; one of the things that killed the classical theory of concepts was simply that concepts are mental entities,[3] and mentalism went out of fashion.

Metaphysics: A classical theory individuates concepts by specifying their contents; the concept X is the concept *of Xs*. This seemed OK – it seemed not to beg any principled questions – because classical theorists thought that they had *of-ness* under control; they thought the image theory of mental representation explained it. We now know that they were wrong to think this. Even if concepts are mental images (which they aren't) and even if the concept DOG looks like a dog (which it doesn't) still, it isn't *because* it looks like a dog that it's concept *of dogs*. *Of*-ness ("content", "intentionality") does not reduce to resemblance, and it is now widely, and rightly, viewed as problematic. It doesn't follow either that classical theorists were wrong to hold that the story about concept possession should be parasitic on the story about concept identification, or that they were wrong to hold that concepts should be individuated by their contents. But it's true that if you want to defend the classical order of analysis, you need an alternative to the picture theory of meaning.

Epistemology: The third of the standard objections to the classical account of concepts, though at least as influential as the others, is distinctly harder to state. Roughly, it's that classical theories aren't adequately "ecological". Used in this connection, the term has Gibsonian ring; but I'm meaning it to pick out a much broader critical tradition. (In fact, I suspect Dewey was the chief influence; see the next footnote.) Here's a rough formulation.

What cognitive science is trying to understand is something that happens *in the world*; it's the interplay of environmental contingencies and behavioral adaptations. Viewing concepts primarily as the vehicles of *thought* puts the locus of this mind/world interaction (metaphorically and maybe literally) not in the world but in the head. Having put it in there, classical theorists are at a loss as to how to get it out again. So the ecological objection goes.

This kind of worry comes in many variants, the epistemological being, perhaps, the most familiar. If concepts are internal mental representations, and thought is conversant only with concepts, how does thought every contact the external world

[3]Terminological footnote: here and elsewhere in this paper, I follow the psychologist's usage rather than the philosopher's; for philosophers, concepts are generally *abstract* entities, hence, of course, *not* mental. The two ways of talking are compatible. The philosopher's concepts can be viewed as the types of which the psychologist's concepts are tokens.

that the mental representations are supposed to represent? If there is a "veil of ideas" between the mind and the world, how can the mind see the world through the veil? Isn't it, in fact, inevitable that the classical style of theorizing eventuates either in solipsism ("we *never do* connect with the world, only with our idea of it") or in idealism ("it's OK if we can never get outside of heads because the world is in there with us")?[4] And, surely, solipsism and idealism are both refutations of theories that entail them.

Notice that this ecological criticism of the classical story is different from the behaviorist's eschewal of intentionality as such. The present objection to "internal representations" is not that they are representations, but that they are internal. In fact, this sort of objection to the classical theory predates behaviorism by a lot. Reid used it against Hume, for example. Notice too that this objection *survives* the demise of the image theory of concepts; treating mental representation as, say, discursive rather than iconic doesn't help. What's wanted isn't either pictures of the world *or* stories about the world; what's wanted is what they call in Europe *being in* the world. (I'm told this sounds even better in German.)

This is all, as I say, hard to formulate precisely; I think, in fact, that it is extremely confused. But even if the "ecological" diagnosis of what's wrong with classical concepts is a bit obscure, it's clear enough what cure was recommended, and this brings us back to our main topic. If what we want is to get thought out of the head and into the world, we need to reverse the classical direction of analysis, precisely as discussed above; we need to take *having a concept* as the fundamental notion and define concept individuation in terms of it. This is a true Copernican revolution in the theory of mind, and we are still living among the debris.

Here, in the roughest outline, is the new theory about concept possession: *having a concept is having certain epistemic capacities*. To have the concept *of X* is to be able to recognize Xs, and/or to be able to reason about Xs in certain kinds of ways. (Compare the classical view discussed above: having the concept *of X* is just being able to have thoughts about Xs). It is a paradigmatically pragmatist idea that having a concept is being able to *do* certain things rather than being able to *think* certain things. Accordingly, in the discussion that follows, I will contrast classical theories of concepts with "pragmatic" ones. I'll try to make it plausible that all the recent and current accounts of concepts in cognitive science really are just variations on the pragmatist legacy.

[4] "Experience to them is not only something extraneous which is occasionally superimposed upon nature, but it forms a veil or screen which shuts us off from nature, unless in some way it can be 'transcended' (p. 1a)". "Other [philosophers' methods] begin with results of a reflection that has already torn in two the subject-matter and the operations and states of experiencing. The problem is then to get together again what has been sundered ..." (p. 9). Thus Dewey (1958). The remedy he recommends is resolutely to refuse to recognize the distinction between experience and its object. "[Experience] recognizes in its primary integrity no division between act and material, subject and object, but contains them both in an unanalyzed totality."

In particular, I propose to consider (briefly, you'll be pleased to hear) what I take to be five failed versions of pragmatism about concepts. Each evokes its proprietary nemesis; there is, for each, a deep fact about concepts by which it is undone. The resulting symmetry is gratifyingly Sophoclean. When we've finished with this catalogue of tragic flaws, we'll have exhausted all the versions of concept pragmatism I've heard of, or can think of, and we'll also have compiled a must-list for whatever theory of concepts pragmatism is eventually replaced by.

2.1. Behavioristic pragmatism (and the problem of intentionality)

I remarked above that behaviorism can be a reason for ruling all mentalistic notions out of psychology, concepts included. However, not all behaviorists were eliminativists; some were reductionists instead. Thus Ryle, and Hull (and even Skinner about half the time) are perfectly content to talk of concept possession, so long as the "criteria" for having a concept can be expressed in the vocabulary of behavior and/or in the vocabulary of dispositions to behave.

Do not ask what criteria are; there are some things we're not meant to know. Suffice it that criterial relations are supposed to be sort-of-semantical rather than sort-of-empirical.

So, then, *which* behaviors are supposed to be criterial for concept possession? Short answer: sorting behaviors. Au fond, according to this tradition, having the concept X is being able to discriminate Xs from non-Xs; to sort things into the ones that are X and the ones that aren't. Though behaviorist in essence, this identification of possessing a concept with being able to discriminate the things it applies to survived well into the age of computer models (see, for example, "procedural" semanticists like Woods, (1975); and lots of philosophers *still* think there must be *something* to it (see, for example, Peacocke, 1992).

This approach gets concepts into the world with a vengeance: having a concept is responding selectively, or being disposed to respond selectively, to the things in the world that the concept applies to; and paradigmatic responses are overt behaviors "under the control" of overt stimulations.

I don't want to bore your with ancient recent history, and I do want to turn to less primitive versions of pragmatism about concepts. So let me just briefly remind you of what proved to be the decisive argument against the behavioristic version: concepts can't be *just* sorting capacities, for if they were, then coextensive concepts – concepts that apply to the same things – would have to be identical. And coextensive concepts aren't, in general, identical. Even *necessarily* coextensive concepts – like TRIANGULAR and TRILATERAL, for example – may perfectly well be different concepts. To put this point another way, sorting is something that happens *under a description*; it's always relative to some or other way of conceptualizing the things that are being sorted. Though their behaviors

may *look* exactly the same, and though they may end up with the very same things in their piles, the creature that is sorting triangles is in a different mental state, and is behaving in a different way, from the creature that is sorting trilaterals; and only the first is exercising the concept TRIANGLE. (For a clear statement of this objection to behaviorism, see Dennett, 1978.)

Behaviorists had a bad case of *mauvais fois* about this; they would dearly have liked to deny the intentionality of sorting outright. In this respect, articles like Kendler (1952), according to which 'what is learned, [is] a pseudoproblem in psychology", make fascinating retrospective reading. Suppose, however, that you accept the point that sorting is always relative to a concept, but you wish, nonetheless, to cleave to some kind of pragmatist reduction of concept individuation to concept possession and of concept possession to having epistemic capacities. The question then arises: *what difference in their epistemic capacities* could distinguish the creature that is sorting triangles from the creature that is sorting trilaterals? What could the difference between them be, if it isn't in the piles that they end up with?

The universally popular answer has been that the difference between *sorting under the concept TRIANGLE* and *sorting under the concept TRILATERAL* lies in *what the sorter is disposed to infer* from the sorting he performs. To think of something as a *triangle* is to think of it *as having angles*; to think of something as a *trilateral* is to think of it as *having sides*. The guy who is collecting triangles must therefore accept *that the things in his collection have angles* (whether or not he has noticed that they have sides); and the guy who is collecting trilaterals must accept *that the things in his collection have sides* (even if he hasn't notice that they have angles).

The long and short is: having concepts is having a mixture of *abilities to sort* and *abilities to infer*.[5] Since inferring is presumably neither a behavior nor a behavioral capacity, this formulation is, of course, not one that a *behavioristic* pragmatist can swallow. So much the worse for behaviorists, as usual. But notice that pragmatists as such are still OK: even if having a concept isn't just knowing how to sort things, it still may be that having a concept is *some* kind of knowing how, and that theories of concept possession are prior to theories of concept individuation.

We are now getting very close to the current scene. All non-behaviorist

[5]The idea that concepts are (at least partially) constituted by inferential capacities receives what seems to be independent support from the success of logicist treatments of the "logical" concepts (AND, ALL, etc.). For many philosophers (though not for many psychologists) thinking of concepts as inferential capacities is a natural way of extending the logicist program from the logical vocabulary to TREE or TABLE. So, when these philosophers tell you what it's like to analyze a concept, they start with AND. (Here again, Peacocke, 1992, is paradigmatic.)

It should, however, strike you as *not obvious* that the analysis of AND is a plausible model for the analysis of TREE or TABLE.

versions of pragmatism hold that concept possession is constituted, at least in part, by inferential dispositions and capacities. They are thus all required to decide *which inferences constitute which concepts*. Contemporary theories of concepts, though without exception pragmatist, are distinguished by the ways that they approach this question. Of non-behavioristic pragmatist theories of concepts there are, by my reckoning, exactly four. Of which the first is as follows.

2.2. *Anarchic pragmatism (and the realism problem)*

Anarchic pragmatism is the doctrine that though concepts are constituted by inferential dispositions and capacities, there is no fact of the matter about which inferences constitute which concepts. California is, of course, the *locus classicus* of anarchic pragmatism; but no doubt there are those even on the East Coast who believe it in their hearts.

I'm not going to discuss the anarchist view. If there are no facts about which inferences constitute which concepts, then there are no facts about which concepts are which. And if there are no facts about which concepts are which, then there are no facts about which beliefs and desires are which (for, by assumption, beliefs and desires are complexes of which concepts are the constituents). And if there are no facts about which beliefs and desires are which, there is no intentional cognitive science, for cognitive science is just belief/desire explanation made systematic. And if there is no cognitive science, we might as well stop worrying about what concepts are and have a nice long soak in a nice hot tub.

I'm also not going to consider a doctrine that is closely related to anarchic pragmatism: namely, that while nothing systematic can be said about concept *identity*, it may be possible to provide a precise account of when, and to what degree, two concepts are *similar*. Some such thought is often voiced informally in the cognitive science literature, but there is, to my knowledge, not even a rough account of how such a similarity relation over concepts might be defined. I strongly suspect this is because a robust notion of similarity is possible only where there is a correspondingly robust notion of identity. For a discussion, see Fodor and Lepore (1992, Ch. 7).

2.3. *Definitional pragmatism (and the analyticity problem)*

Suppose the English word "bachelor" means the same as the English phrase "unmarried male". Synonymous terms presumably express the same concept (this is a main connection between theories about concepts and theories about language), so it follows that you couldn't have the concept BACHELOR and fail to have the concept UNMARRIED MALE. And from that, together with the

intentionality of sorting (see section 2.1), it follows that you couldn't be collecting bachelors *so described* unless you take yourself to be collecting unmarried males; that is, unless you accept the inference that if something belongs in your bachelor collection, then it is something that is male and unmarried.

Maybe this treatment generalizes; maybe, having the concept X *just is* being able to sort Xs and being disposed to draw the inferences that *define X-ness*.

The idea that it's *defining* inferences that count for concept possession is now almost as unfashionable as behaviorism. Still, the departed deserves a word or two of praise. The definition story offered a plausible (though partial) account of the *acquisition* of concepts. If BACHELOR *is* the concept UNMARRIED MALE, then it's not hard to imagine how a creature that has the concept UNMARRIED and has the concept MALE could put them together and thereby achieve the concept BACHELOR. (Of course the theory that complex concepts are acquired by constructing them from their elements *presupposes* the availability of the elements. About the acquisition of these, definitional pragmatism tended to be hazy.) This process of assembling concepts can be – indeed, was – studied in the laboratory; see Bruner, Goodnow, & Austin (1956) and the large experimental literature that it inspired. Other significant virtues of the definition story will suggest themselves when we discuss concepts as prototypes in section 2.4.

But alas, despite its advantages, the definition theory doesn't work. Concepts can't *be* definitions because most concepts don't *have* definitions. At a minimum, to define a concept is to provide necessary and sufficient conditions for something to be in its extension (i.e., for being among the things that concept applies to). And, if the definition is to be informative, the vocabulary in which it is couched must not include either the concept itself or any of its synonyms. As it turns out, for most concepts, this condition simply can't be met; more precisely, it can't be met unless the definition employs synonyms and near-synonyms of the concept to be defined. Maybe being male and unmarried is necessary and sufficient for being a bachelor; but try actually filling in the blanks in "x is a dog iff x is a ..." without using the words like "dog" or "canine" or the like on the right-hand side.

There is, to be sure, a way to do it; if you could make a *list* of all and only the dogs (Rover, Lassie, Spot ... etc.), then being on the list would be necessary and sufficient for being in the extension of DOG. That there is this option is, however, no comfort for the theory that concepts are definitions. Rather, what it shows is that being a necessary and sufficient condition for the application of a concept is not a sufficient condition for being a definition of the concept.

This point generalizes beyond the case of lists. *Being a creature with a backbone* is necessary and sufficient for *being a creature with a heart* (so they tell me). But it isn't the case that "creature with a backbone" defines "creature with a heart" or vice versa. Quite generally, it seems that Y doesn't define X unless Y applies to all and only the *possible* Xs (as well, of course, as all and only the

actual Xs). It is, then, the modal notion – possibility – that's at the heart of the idea that concepts are definitions. Correspondingly, what killed the definition theory of concepts, first in philosophy and then in cognitive psychology, is that nobody was able to explicate the relevant sense of "possible".

It seems clear enough that even if Rover, Lassie and Spot are all the dogs that there actually are, it is *possible*, compatible with the concept of DOG, that there should be others; that's why you can't define DOG by just listing the dogs. But is it, in the same sense, possible, compatible with the concept DOG that some of these non-actual dogs are ten feet long? How about twenty feet long? How about twenty miles long? How about a light-year long? To be sure, it's not *biologically* possible that there should be a dog as big as a light-year; but presumably biology rules out a lot of options that the *concept* DOG, as such, allows. Probably biology rules out zebra-striped dogs; surely it rules out dogs that are striped red, white and blue. But I suppose that red, white and blue striped dogs are *conceptually* possible; somebody who thought that there might be such dogs wouldn't thereby show himself not to have the concept DOG – would he?

So, again, are light-year-long dogs possible, compatible with the concept DOG? Suppose somebody thought that maybe there could be a dachshund a light-year long. Would that show that he has failed to master the concept DOG? Or the concept LIGHT-YEAR? Or both?

To put the point in the standard philosophical jargon: even if light-year-long dogs aren't really possible, "shorter than a light-year" is part of the *definition* of DOG only if "some dogs are longer than a light-year" is *analytically* impossible; mere biological or physical (or even metaphysical) impossibility won't do. Well, is it analytically impossible that there should be such dogs? If you doubt that this kind of question has an answer, or that it matters a lot for any serious purpose what the answer is, you are thereby doubting that the notion of definition has an important role to play in the theory of concept possession. So much for definitions.

2.4. Stereotypes and prototypes (and the problem of compositionality)

Because it was pragmatist, the definition story treated having a concept as having a bundle of inferential capacities, and was faced with the usual problem about which inferences belong to which bundles. The notion of an *analytic* inference was supposed to bear the burden of answering this question, and the project foundered because nobody knows what makes an inference analytic, and nobody has any idea how to find out. "Well", an exasperated pragmatist might nonetheless reply, "even if I don't know what makes an inference analytic, I do know what makes an inference statistically reliable. So why couldn't the theory of concept possession be statistical rather than definitional? Why couldn't I exploit

the notion of a *reliable* inference to do what definitional pragmatism tried and failed to do with the notion of an *analytic* inference?"

We arrive, at last, at modern times. For lots of kinds of Xs, people are in striking agreement about what properties an arbitrarily chosen X is likely to have. (An arbitrarily chosen bird is likely to be able to fly; an arbitrarily chosen conservative is likely to be a Republican; an arbitrarily chosen dog is likely to be less than a light-year long.) Moreover, for lots of kinds of Xs, people are in striking agreement about which Xs are prototypic of the kind (diamonds for jewels; red for colors; not dachshunds for dogs). And, sure enough, the Xs that are judged to be prototypical are generally ones that have lots of the properties that an arbitrary X is judged likely to have; and the Xs that are judged to have lots of the properties that an arbitrary X is likely to have are generally the ones that are judged to be prototypical.

Notice, in passing, that stereotypes share one of the most agreeable features of definitions: they make the learning of (complex) concepts intelligible. If the concept of an X is the concept of something that is reliably Y and Z, then you can learn the concept X if you have the concepts Y and Z together with enough statistics to recognize reliability when you see it. It would be OK, for this purpose, if the available statistical procedures were analogically (rather than explicitly) represented in the learner. Qua learning models, "neural networks" are analog computers of statistical dependencies, so it's hardly surprising that prototype theories of concepts are popular among connectionists. (See, for example, McClelland & Rummelhart, 1986.)

So, then, why shouldn't having the concept of an X be having the ability to sort by X-ness, together with a disposition to infer from something's being X to its having the typical properties of Xs? I think, in fact, that this is probably the view of concepts that the prototypical cognitive scientist holds these days.

To see why it doesn't work, let's return one last time to the defunct idea that concepts are definitions. It was a virtue of that idea that it provides for the *compositionality* of concepts, and hence for the productivity and systematicity of thought. This, we're about to see, is no small matter.

In the first instance, productivity and systematicity are best illustrated by reference to features (not of minds but) of natural languages. To say that languages are productive is to say that there is no upper bound to the number of well-formed formulas that they contain. To say that they are systematic is to say that if a language can express the proposition that P, then it will be able to express a variety of other propositions that are, in one way or another, semantically related to P. (So, if a language can say that P and that $-Q$, it will also be able to say that Q and that $-P$; if it can say that John loves Mary, it will be able to say that Mary loves John . . . and so forth.) As far as anybody knows, productivity and systematicity are universal features of human languages.

Productivity and systematicity are also universal features of human *thought*

(and, for all I know, of the thoughts of many infra-human creatures). There is no upper bound to the number of thoughts that a person can think. (I am assuming the usual distinctions between cognitive "competence" and cognitive "performance"). And also, if a mind can entertain the thought that P and any negative thoughts, it can also entertain the thought that −P; if it can entertain the thought that Mary loves John, it can entertain the thought that John loves Mary . . . and so on.

It is extremely plausible that the productivity and the systematicity of language and thought are both to be explained by appeal to the systematicity and productivity of mental representations, and that mental representations are systematic and productive because they are compositional. The idea is that mental representations are constructed by the application of a finite number of combinatorial principles to a finite basis of (relatively or absolutely) primitive concepts. (So, the very same process that gets you from the concept MISSILE to the concept ANTIMISSILE, also gets you from the concept ANTIMISSILE to the concept ANTIANTIMISSLE, and so on ad infinitum.) Productivity follows because the application of these constructive principles can iterate without bound. Systematicity follows because the concepts and principles you need to construct the thoughts that P and −Q are the very same ones that you need to construct the thoughts that Q and −P; and the concepts and principles you need to construct the thought that John loves Mary are the very same ones that you need to construct the thought that Mary loves John.

This sort of treatment of compositionality is familiar, and I will assume that it is essentially correct. I want to emphasize that it places a heavy constraint on both theories of concept possession and theories of concept individuation. If you accept compositionality, then you are required to say that whatever the concept DOG is that occurs in the thought that *Rover is a dog*, that *very same* concept DOG also occurs in the thought that *Rover is a brown dog*; and, whatever the concept BROWN is that occurs in the thought that *Rover is brown*, the very same concept BROWN also occurs in the thought that *Rover is a brown dog*. It's on these assumptions that compositionality explains how being able to think that Rover is brown and that Rover is a dog is linked to being able to think that Rover is a brown dog. Compositionality requires, in effect, that constituent concepts must be *insensitive* to their host; a constituent concept contributes the same content to all the complex representations it occurs in.

And compositionality further requires that the content of a complex representation is *exhausted* by the contributions that its constituents make. Whatever the content of the concept of BROWN DOG may be, it must be completely determined by the content of the constituent concepts BROWN and DOG, together with the combinatorial apparatus that sticks these constituents together; if this were not the case, your grasp of the concepts BROWN and DOG wouldn't explain your grasp of the concept BROWN DOG.

In short, when complex concepts are compositional, the whole must *not* be more than the sum of its parts, otherwise compositionality won't explain productivity and systematicity. And if compositionality doesn't, nothing will. If this account of compositionality strikes you as a bit austere, it may be some comfort that the systematicity and productivity of thought is compatible with compositionality failing in any finite number of cases. It allows, for example, that finitely many thoughts (hence *a fortiori*, finitely many linguistic expressions) are idiomatic or metaphoric, so long as there are infinitely many that are neither.

We can now see why, though concepts might have turned out to be definitions, they couldn't possibly turn out to be stereotypes or prototypes. Concepts do contribute their *defining* properties to the complexes of which they are constituents, and the *defining* properties of complex concepts are exhaustively determined by the defining properties that the constituents contribute. Since bachelors are, by definition, unmarried men, tall bachelors are, by the same definition, tall unmarried men; and very tall bachelors are very tall unmarried men, and very tall bachelors from Hoboken are very tall unmarried men from Hoboken ... and so on. Correspondingly, there is nothing more to the *definition* of "very tall bachelor from Hoboken" than *very tall unmarried man from Hoboken*; that is, there is nothing more to the definition of the phrase than what the definitions of its constituents contribute.

So, then, if concepts were definitions, we could see how thought could be compositional, and hence productive and systematic. Concepts aren't definitions, of course. It's just that, from the present perspective, it's rather a pity that they're not.

For stereotypes, alas, don't work the way that definitions do. Stereotypes aren't compositional. Thus, "ADJECTIVE X" can be a perfectly good concept even if there is no *adjective X* stereotype. And even if there are stereotypic *adjective* Xs, they don't have to be stereotypic *adjectives* or stereotypic *Xs*. I doubt, for example, that there is a stereotype of very tall men from Hoboken; but, even if there were, there is no reason to suppose that it would be either a stereotype for tall men, or a stereotype for men from Hoboken, or a stereotype for men. On the contrary: often enough, the adjective in "ADJECTIVE X" is there precisely to mark a way that adjective Xs *depart* from stereotypic Xs. Fitzgerald made this point about stereotypes to Hemingway when he said, "The rich are different from the rest of us." Hemingway replied by making the corresponding point about definitions: "Yes", he said, "they have more money".

In fact, this observation about the uncompositionality of stereotypes generalizes in a way that seems to me badly to undermine the whole pragmatist program of identifying concept possession with inferential dispositions. I've claimed that knowing what is typical of *adjective* and what is typical of *X* doesn't, in the general case, tell you what is typical of *adjective Xs*. The reason it doesn't is perfectly clear; though some of your beliefs about *adjective Xs* are compositional-

ly inherited from your beliefs about *adjectives*, and some are compositionally inherited from your beliefs about *Xs*, *some are beliefs that you have acquired about adjective Xs as such*, and these aren't compositional at all.

The same applies, of course, to the inferences that your beliefs about *adjective Xs* dispose you to draw. Some of the inferences you are prepared to make about green apples follow just from their being green and from their being apples. That is to say: they derive entirely from the constituency and structure of your GREEN APPLE concept. But others depend on information (or misinformation) that you have picked up about green apples as such: that green apples go well in apple pie; that they are likely to taste sour; that there are kinds of green apples that you'd best not eat uncooked, and so forth. Patently, these inferences are not definitional and not compositional; they are *not* ones that GREEN APPLE inherits from its constituents. They belong to what you know about green applies, not to what you know about the corresponding words or concepts. You learned that "green apple" means *green and apple* when you learned English at your mother's knee. But probably you learned that green apples mean apple pies from the likes of Julia Child.

The moral is this: the content of complex concepts has to be compositionally determined, so whatever about concepts is *not* compositionally determined is therefore not their content. But, as we've just been seeing, the inferential role of a concept is not, in general, determined by its structure together with the inferential roles of its constituents. That is, the inferential roles of concepts are *not*, in general, compositional; only defining inferences are.

This puts your paradigmatic cognitive scientist in something of a pickle. On the one hand, he has (rightly, I think) rejected the idea that concepts are definitions. On the other hand, he cleaves (wrongly, I think) to the idea that having concepts is having certain inferential dispositions. But, on the third hand (as it were), *only defining inferences are compositional* so if there are no definitions, then having concepts *can't* be having inferential capacities.

I think that is very close to being a proof that the pragmatist notion of what it is to have a concept must be false. This line of argument was first set out in Fodor and Lepore (1992). Philosophical reaction has been mostly that if the price of the pragmatist account of concepts is reviving the notion that there are analytic/definitional inferences, then there must indeed be analytic/definitional inferences. My own view is that cognitive science is right about concepts not being definitions, and that it's the analysis of having concepts in terms of drawing inferences that is mistaken. Either way, it seems clear that the current situation is unstable. Something's gotta give.

I return briefly to my enumeration of the varieties of pragmatist theories of concept possession. It should now seem unsurprising that none of them work. In light of the issues about compositionality that we've just discussed, it appears there are principled reasons why none of them could.

2.5. *The "theory theory" of concepts (and the problem of holism)*

Pragmatists think that having a concept is having certain epistemic capacities; centrally it's having the capacity to draw certain inferences. We've had trouble figuring out *which* inferences constitute which concepts; well, maybe that's because we haven't been taken the *epistemic* bit sufficiently seriously.

Concepts are typically parts of beliefs; but they are also, in a different sense of "part", typically parts of theories. This is clearly true of sophisticated concepts like ELECTRON, but perhaps it's *always* true. Even every-day concepts like HAND or TREE or TOOTHBRUSH figure in complex, largely inarticulate knowledge structures. To know about hands is to know, *inter alia*, about arms and fingers; to know about toothbrushes is, *inter alia*, to know about teeth and the brushing of them. Perhaps, then, concepts are just *abstractions from* such formal and informal knowledge structures. On this view, to have the concept ELECTRON is to know what physics has to say about electrons; and to have the concept TOOTHBRUSH is to know what dental folklore has to say about teeth.

Here are some passages in which the developmental cognitive psychologist Susan Carey (1985) discusses the approach to concepts that she favors: "... [young] children represent only a few theory-like cognitive structures, in which their notions of causality are embedded and in terms of which their deep ontological commitments are explicated. Cognitive development consists, in part, in the emergence of new theories out of these older ones, with the concomitant reconstructing of the ontologically important concepts and emergence of new explanatory notions" (p. 14); "... successive theories differ in three related ways: in the domain of phenomena accounted for, the nature of explanations deemed acceptable, and even in the individual concepts at the center of each system ... Change of one kind cannot be understood without reference to the changes of the other kinds" (pp. 4–5). The last two sentences are quoted from Carey's discussion of theory shifts in the history of science; her proposal is, in effect, that these are paradigms for conceptual changes in ontogeny.

Cognitive science is where philosophy goes when it dies. The version of pragmatism according to which concepts are abstractions from knowledge structures corresponds exactly to the version of positivism according to which terms like "electron" are defined implicitly by reference to the theories they occur in. Both fail, and for the same reasons.

Suppose you have a theory about electrons (viz. that they are X) and I have a different theory about electrons (viz. that they are Y). And suppose, in both cases, that our use of the term "electron" is implicitly defined by the theories we espouse. Well, the "theory theory" says that you have an essentially different *concept* of electrons from mine if (and only if?) you have an essentially different *theory* of electrons from mine. The problem of how to individuate concepts thus reduces to the problem of how to individuate theories, according to this view.

But, of course, nobody knows how to individuate theories. Roughly speaking, theories are bundles of inferences, just as concepts are according to the pragmatist treatment. The problem about which inferences constitute which concepts has therefore an exact analagon in the problem which inferences constitute which theories. Unsurprisingly, these problems are equally intractable. Indeed, according to the pragmatist view, they are interdefined. Theories are essentially different if they exploit essentially different concepts; concepts are essentially different if they are exploited by essentially different theories. It's hard to believe it matters much which of these shells you keep the pea under.

One thing does seem clear: if your way out of the shell game is to say that a concept is constituted by the *whole* of the theory it belongs to, you will pay the price of extravagant paradox. For example: it turns out that you and I can't disagree about dogs, or electrons, or toothbrushes since we have no common conceptual apparatus in which to couch the disagreement. You utter "Some dogs have tails." "No dogs have tails" I reply. We seem to be contradicting one another, but in fact we're not. Since tailessness is part of my *theory* of dogs, it is also part of my *concept* DOG according to the present, holist account of concept individuation. Since you and I have different concepts of dogs, we mean different things when we say "dog". So the disagreement between us is, as comfortable muddleheads like to put it, "just semantic". You might have thought that our disagreement was about the facts and that you could refute what I said by producing a dog with a tail. But it wasn't and you can't, so don't bother trying; you have you idea of dogs and I have mine. (What, one wonders, makes them both ideas *of dogs*?) First the pragmatist theory of concepts, then the theory theory of concepts, then holism, then relativism. So it goes. Or so, at least, it's often gone.

I want to emphasize two caveats. The first is that I'm *not* accusing Carey of concept holism, still less of the slide from concept holism to relativism. Carey thinks that only the "central" principles of a theory individuate its concepts. The trouble is that she has no account of centrality, and the question "which of the inferences a theory licenses are *central*?" sounds suspiciously similar to the question "which of the inferences that a concept licenses are *constitutive*?" Carey cites with approval Kuhn's famous distinction between theory changes that amount to paradigm shifts and those that don't (Kuhn, 1962). If you have caught onto how this game is played, you won't be surprised to hear that nobody knows how to individuate paradigms either. *Where is this buck going to stop*?

My second caveat is that holism about *the acquisition of beliefs* and about *the confirmation of theories* might well both be true even if holism about the *individuation of concepts* is, as I believe, hopeless. There is no contradiction between Quine's famous dictum that it's only as a totality that our beliefs "face the tribunal of experience", and Hume's refusal to construe the content of one's concepts as being determined by the character of one's theoretical commitments.

There is, to be sure, a deep, deep problem about how to get a theory of confirmation and belief fixation if you are an atomist about concepts. But there is also a deep, deep problem about how to get a theory of confirmation and belief fixation if you are *not* an atomist about concepts. So far as I know, there's no reason to suppose that the first of these problems is worse than the second.

So much for caveats. It's worth noticing that the holistic account of concepts at which we've now dead-ended is diametrically opposite to the classical view that we started with. We saw that, for the likes of Hume, any concept could become associated to any other. This was a way of saying that the identity of a concept is independent of the theories one holds about the things that fall under it; it's independent, to put it contemporary terms, of the concept's *inferential role*. In classical accounts, concepts are individuated by what they are concepts of, and not by what theories they belong to. Hume was thus a radical atomist just where contemporary cognitive scientists are tempted to be radically holist. In this respect, I think that Hume was closer to the truth than we are.

Here's how the discussion has gone so far: modern representational theories of mind are devoted to the pragmatist idea that having concepts is having epistemic capacities. But not just sorting capacities since sorting is itself relativized to concepts. Maybe, then, inferential capacities as well? So be it, but *which* inferential capacities? Well, at a minimum, inferential capacities that respect the compositionality of mental representations. *Defining* inferences are candidates since they do respect the compositionality of mental representations. Or, rather, they would if there were any definitions, but there aren't any definitions to speak of. Statistical inferences aren't candidates because they aren't compositional. It follows that concepts can't be stereotypes. The "theory theory" merely begs the problem it is meant to solve since the individuation of theories *presupposes* the individuation of the concepts they contain. Holism would be a godsend and the perfect way out except that it's preposterous on the face of it. What's left, then, for a pragmatist to turn to?

I suspect, in fact, that there is nothing left for a pragmatist to turn to and that our cognitive science is in deep trouble. Not that there aren't mental representations, or that mental representations aren't made of concepts. The problem is, rather, that Hume was right: concepts aren't individuated by the roles that they play in inferences, or, indeed, by their roles in any other mental processes. If, by stipulation, semantics is about what constitutes concepts and psychology is about the nature of mental processes, then the view I'm recommending is that *semantics isn't part of psychology*.

If semantics isn't part of psychology, you don't need to have a sophisticated theory of mental processes in order to get it right about what concepts are. Hume, for example, did get it right about what concepts are, even though his theory of mental processes was associationistic and hence hopelessly primitive. Concepts are the constituents of thoughts; as such, they're the most elementary mental

objects that have both causal and representational properties. Since, however, concepts are individuated by their representational *and not* by their casual properties, all that has to specified in order to identify a concept is what it is the concept of. The whole story about the individuation of the concept DOG is that it's the concept that represents dogs, as previously remarked.

But if "What individuates concepts?" is easy, that's because its the wrong question, according to the present view. The right questions are: "How do mental representations represent?" and "How are we to reconcile atomism about the individuation of concepts with the holism of such key cognitive processes as inductive inference and the fixation of belief?" Pretty much all we know about the first question is that here Hume was, for once, wrong; mental representation doesn't reduce to mental imaging. What we know about the second question is, as far as I can tell, pretty nearly nothing at all. The project of constructing a representational theory of the mind is among the most interesting that empirical science has ever proposed. But I'm afraid we've gone about it all wrong.

At the very end of *Portnoy's Complaint*, the client's two hundred pages of tortured, non-directive self-analysis comes to an end. In the last sentence of the book, the psychiatrist finally speaks: "So [said the doctor]. Now vee may perhaps to begin. Yes?"

References

Bruner, J., Goodnow, J., & Austin, G. (1956). *A Study of Thinking*. New York: Wiley.

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

Dennett, D. (1978). Skinner Skinned. In *Brainstorms*. Cambridge, MA: MIT Press.

Dewey, J. (1958). *Experience and Nature*. New York: Dover Publications.

Fodor, J. (1990). *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

Fodor, J., & Lepore, E. (1991). Why meaning (probably) isn't conceptual role. *Mind and Language*, *6*, 328–343.

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

Fodor, J., & McLaughlin, B. (1990). Connectionism and the problem of systematicity: why Smolensky's solution doesn't work. *Cognition*, *35*, 183–204.

Fodor, J., & Pylyshyn, Z. (1981). How direct is visual perception? Some reflection on Gibson's "ecological approach". *Cognition*, *9*, 139–196.

Fodor, J., & Pylyshyn, Z. (1988). Correctionism and cognitive architecture. *Cognition*, *28*, 3–71.

Kendler, H. (1952). "What is learned?" A theoretical blind alley. *Psychological Review*, *59*, 269–277.

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

McClelland, J., & Rummelhart, D. (1986). A distributed model of human learning and memory. In J. McClelland & D. Rummelhart (Eds.), *Parallel Distributed Processing* (Vol. 2). Cambridge, MA: MIT Press.

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

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*, 1–23.

Woods, W. (1975). What's in a link? In D. Bobrow & A. Collins (Eds.), *Representation and Understanding*. New York: Academic Press.