The Concept Concept: The Wayward Path of Cognitive Science*

FRANK C. KEIL AND ROBERT A. WILSON

1. Introduction

When one says 'dog', or 'yawl', or 'junta', there is the strong impression that discrete ideas correspond to each of those words. Cognitive science, following common sense, calls such ideas 'concepts'. Typically at least, we reflective common folk think of a word as the expression of a concept, of concepts as the constituents of larger mental units, such as thoughts, and thus of concepts as central to our mental life.

The general directions that the study of concepts has taken within cognitive science are clear, and a sweeping critique of those directions is the focus of this most recent book from Jerry Fodor. As the title suggests, Fodor thinks that not all is well in the House of Concepts. In essence, cognitive science has assumed that (a) there is a rich internal structure to concepts that can be used to predict how concepts function in our mental activities and overt behavior; and (b) there are rich external structures in which concepts are embedded, structures often called *intuitive theories*, that are needed to explain concept acquisition and use. Fodor rejects both (a) and (b), focusing largely on the various ways in which (a) has been articulated in linguistics, philosophy, and psychology (with glancing blows at artificial intelligence along the way).

At the core of Fodor's diagnosis of what has gone wrong is inferential role semantics (IRS), the view that concepts are individuated in part by the inferential relations they participate in. IRS implies that at least part of what makes something in the head, the concept DOG, say, is that it is typically/ideally/always inferred from other mental particulars (e.g. ROVER, MAN'S BEST FRIEND), and typically/ideally/always leads one to infer still other

*Review of Jerry A. Fodor, *Concepts: Where Cognitive Science Went Wrong*. Oxford: Clarendon Press, 1998. Pp. xii + 174.

Preparation of this paper was supported by NIH grant R01-HD23922 to Keil. We thank Paul Bloom for helpful comments on an earlier draft of this manuscript.

Address for correspondence: Department of Psychology, Yale University, PO Box 208205, New Haven, CT 06520-8205, USA.
Email: frank.keil@yale.edu.

Mind & Language, Vol. 15 Nos 2 and 3 April/June 2000, pp. 308–318.

concepts (e.g. PET, CHASER OF CATS). Much of the book is an attempt to show what is wrong with views of concepts in different areas of cognitive science that rest on IRS: lexical semantics in linguistics (ch. 3), appeals to analyticity in philosophy (ch. 4), and prototype theory in psychology (ch. 5).

The balance of *Concepts* is devoted to laying the ground work for this critique and articulating Fodor's alternative to IRS, a view he dubs informational atomism about concepts. Rather than being individuated by inferential relations, concepts are individuated by world-mind nomic relations (this is the informational part of the view); in addition, according to Fodor, most concepts have no internal structure at all (thus, atomism). Hence, at a first pass, what makes some chunk of my head the concept DOG, say, is that it is typically/ideally/always caused by dogs; second pass, there is a law of nature that causally relates concepts to what they are concepts of—what in Fodor's terms they lock onto—and that thus provides the individuation conditions for specific concepts.

Fodor begins (ch. 1) with an overview of the representational theory of mind (RTM), the core of which will be familiar to readers of any of his books since *The Language of Thought* (1975), and then moves (in ch. 2) to lay out five 'non-negotiable conditions on a theory of concepts'. Those conditions are that concepts be: mental particulars that function as mental causes and effects; categories that apply to things in the world; compositional; in many cases, learned; and public in that they are widely shared. Since Fodor has discussed views akin to informational atomism elsewhere (e.g. Fodor, 1987, ch. 4; 1990, chs 3–4), in this book his positive thesis about concepts focuses on the relations between that view and more or less radical forms of nativism about concepts. Chapter 6 concentrates on what Fodor calls the doorknob/DOORKNOB problem—why it is that experiences of doorknobs 'so often', as Fodor says (p. 127), 'leads one to lock to doorknobhood'—while chapter 7 discusses natural kind concepts and informational atomism. The discussion here is somewhat philosophically rarefied in a way that may make it seem irrelevant to most cognitive scientists with an interest in concepts.

The book is fast-paced and entertaining, as one has come to expect from Fodor, and many of the arguments are provocative and interesting in their own right. But regarding the big picture of concepts that Fodor paints, we are unconvinced. *Concepts* falls short of establishing its chief negative thesis—that a large muddle occupies centre stage in work in cognitive science on concepts. Moreover, it fails to offer a positive proposal for how to think about concepts that either satisfies Fodor's own conditions for a theory of concepts, or is independently plausible as a general account of concepts. Here are the contours of our dissent:

(1) Conceptual combination can be treated more adequately by referring both to the 'theory laden' parts of conceptual structure and the probabilistic parts that Fodor rejects.

(2) Rather than rejecting the concepts-in-theories view, we need to rethink what theories are in this context.
(3) Many concepts are learned that never lock onto 'things' in this world.
(4) Concepts are often organized around ideals in ways that seem incompatible with informational atomism.
(5) Conceptual change, in history and in development, follows patterns that would make little sense if concepts had no internal structures.

We will discuss (1) and (2) together as ways in which Fodor's rejection of internal and external structure is misplaced ('Why All is Not Lost'); and we will develop (3)–(5) as deep problems for Fodor's informational atomism as a constructive alternative to IRS-based views ('From the Frying Pan to the Fire'). But first we consider some problems with informational atomism and Fodor's conception of what structures underpin our cognitive performance.

2. Seeing RED?

According to Fodor, we should view most of our concepts as being, in the respects that matter, like the concept RED, and thus what it is to have concepts in general as akin to what it is to have the concept RED. RED is acquired (or triggered), and its content fixed, by red objects in the world (thus RED is an 'informational' concept). Moreover, there is no internal structure of the concept red: you can't define 'red', and you don't usually have a theory of what it is for something to be an instance of RED (thus RED is an 'atomistic' concept). As RED goes, so goes the nation of concepts. Two basic questions seem worth raising briefly before one even gets to the issue of whether one can generalize from RED to concepts more generally:

(i) Is Fodor right about RED being atomistic? While one might think that the mental state of having a red image is unstructured, unless RED is merely a mental image (and so the chance of generalizing from RED to concepts in general is nil), it doesn't follow that the concept RED is unstructured. Since RED is, after all, a colour concept, it is not implausible to think that RED involves COLOUR such that no one could have the concept RED who didn't recognize that red is a colour, and thus also have the concept COLOUR (as well as the concept of BEING A KIND OF, or something like it). Similar points could, we think, be made about the similarity that exists between RED and at least some other colours (including shades or, in philosophical jargon, determinates of red). Having the concept RED may require that one also appreciate its contrasting role with other colours. We might be reluctant to grant the concept RED to an organism that could only see red things and could not see or think about other colours.

(ii) Is Fodor right about RED being informational? There are reasons to think that the answer to this question is also 'no', at least given Fodor's sense of 'informational'. The externalist idea that RED contains information about redness because it is systematically tokened by exposure to red things in the world is intuitive, and provides a way of satisfying Fodor's publicity constraint. But the problems with this basic idea are well known, and they are problems that, if anything, are made worse by Fodor's dressing the general idea in the garb of nomological locking. RED could, seemingly, be acquired by someone without any exposure at all to red things in the world—not only by black-and-white-room-bound colour scientist Mary but also by one whose only experience of redness is produced by having a bright light shone on one's closed eyes—implying that this informational connection is not necessary for RED. Moreover, those who are visually bombarded with red things might well acquire some concept other than RED—suppose that they have an anomaly in colour vision circuitry that makes red things seem blue or that they lack any of the cognitive structures that we RED-bearing creatures have [as in (i) above]—implying that such an informational connection is not sufficient for RED.

Idealization, perhaps in some form of a competence/performance distinction, could reasonably be expected to bear some of the burden here, as well as handle the small but very real biological differences in how people perceive colour. The problem, however, is that informational atomism itself, with its general appeals to the relation between RED and red things being nomological, and the idea that RED simply locks onto red things, gives no guidance linking such hoped for idealizations and explanatory practice. And however bad things are with RED, they are surely much worse for concepts in general.

In short, even the story about RED is more complicated than Fodor implies, and it is complicated in ways that call into question the generalization of informational atomism to other sorts of concepts.

3. Why All is Not Lost I: Keeping Some Composure about Conceptual Combination

So much for RED. What about DOG? There are lots of contingent and often personal facts about dogs that we certainly don't want as part of our concepts of dogs: that dogs are the things about which Aunt Mildred has a phobia; that Dalmatians are sky rocketing in popularity ever since a certain movie came out; that Odysseus's dog waited faithfully for him for 20 years and then dropped dead on Odysseus's return. All but one extreme wing of cognitive science wants to reject the idea that everything we know, everything that is possibly associated with dogs, is actually part of our concept DOG. But Fodor

argues that once we accept that not everything is part of our dog concept, we will slowly get backed into conceding that nothing is, and that DOG is like RED with respect to internal mental structure.

The way out is not clear, but would seem to rest on the ideas that some features are central to a concept even though they aren't strictly necessary, and that there are principled ways to find out what these features are. This is not to say that concepts are merely associative tabulations of features, and that those most powerfully associated with instances of a category are constitutive conceptual structure: that account is woefully incomplete, as the last 15 years or so of cognitive science has repeatedly shown. Highly perceptually salient features that regularly occur with instances of a category can be seen by everyone, including youngish children, as irrelevant to the concept. Almost every tyre I've seen is black, yet I am convinced that blackness is irrelevant to tyrehood; children are similarly convinced. Every video cassette I've seen is black, yet I am equally convinced that blackness is irrelevant to video cassettehood; again, children are similarly convinced (Keil et al. 1998). Not every dog I have seen has four legs, yet I am convinced that four-leggedness is a central part of my concept of dog. What drives these intuitions and how might it be related to concepts?

For most sorts of things, it may have to do with the causal impact of properties—their counterfactual robustness. Redness would not undermine VIDEO CASSETTE, for example, but a weird shape might. (No easy fix here, though, since counterfactual undermining is also a matter of degree: video cassettes made out of metal might undermine their function, something not clear without more detailed knowledge of the kind of metal, etc.). The point, however, is that highlighting features through causal centrality is a very different process from highlighting them through frequencies of occurrence in members of a category. Some notion of gradedness may have to be involved and there may be uncertain cases, but it is not so clear that Fodor's concerns about the limitations of appeals to typicality and frequency extend to the notion of causal centrality. For one thing, those causal relations may help sort out the mess of conceptual combinations.

Fodor sees cognitive science as having been obsessively fixated on how we use concepts to sort the world, and as having been mostly in denial about how concepts combine to make new ones. In those few instances where conceptual combinations have been addressed, Fodor argues that cognitive science has been a complete failure, a failure he equates with the inabilities of some prototype models to explain how concepts compose. We agree with part of this assessment. Cognitive science, at least in cognitive psychology, has mostly examined concepts via categorization, and this has distorted the study of concepts themselves. More attention should be paid to concepts in conceptual combinations, induction and other forms of reasoning. But we disagree with Fodor's claim that cognitive science has made no progress in the study of conceptual combinations; in fact the progress there seems more promising than what seems to be on offer from informational atomism.

What counts as a successful account of how concepts compose? Predicting the extension of a complex concept from its constituents? Fodor uses this criterion in his criticisms of the inabilities of prototype models to predict such extensions. But he ignores the degree to which more recent work on conceptual combinations does have predictive success from analyses based on purely probabilistic representations (see Hampton, this issue, for more on such progress). Such probabilistic models are not adequate for all combinations, and the patterns of failures are interesting in their own right, but they do a good job of telling part of the story. A different putative aspect to concepts, the explanatory relations in which they are embedded, has also been able to predict a number of properties that seem to emerge when concepts combine (Murphy, 1988, 1990, in press; Gagne and Shoben, 1997; Johnson and Keil, in press). Look at the most typical features for ARCTIC and BICYCLE, and many features of ARCTIC BICYCLE (e.g. studded tyres) will not be in either constituent list. Yet if you manipulate the causal/explanatory links between apparent constituent features for ARCTIC and BICYCLE, corresponding changes in emergent features in the combination can be predicted (Johnson and Keil, in press). By contrast, we cannot see how informational atomism would help us here at all.

Conceptual combinations are complex and are not all like DEAD WHITE MALES, pointing to men that are dead and white. Intricate patterns arise by virtue of a host of properties of the constituents. Even with purely syntactic criteria for combining concepts, several options remain for the same sorts of pairings. Consider for example 'Blog Gormination', where both are nouns. It could mean gorminations made to resemble blogs (e.g. dog decoration); it could mean something that happens to blogs (dog destruction), or something made out of blogs (dog concoction). This is old news, but if we can't even predict the syntax of combinations with nonsense words and their categories, why should we point to difficulties in conceptual combinations as so telling for a theory of concepts? Moreover, how does informational atomism help? Consider new entailments arising out of composed concepts. A thing made only of PLASTIC PARTS is necessarily plastic; but a thing made only of SMALL PARTS is not necessarily small (Katz, 1972). It certainly seems that the internal structures of PLASTIC, SMALL and PART are responsible for these differences. To convince us otherwise and show how locking patterns instead can account for such compositional outcomes is the challenge not yet met in this book.

Concepts *do* compose, albeit by a complex process that is not yet well understood, but to use the difficulty in making predictions as a way of skewering some representational theories of concepts seems a bit like using the difficulty of predicting where a leaf will fall as a way of attacking classical mechanics.

4. All is Not Lost II: Removing Skeletons from the Theoretical Closet

One misleading aspect of the theory-theory of concepts is its tendency to direct one to think about concepts as parts of well-worked-out theories, and Fodor

rightly criticizes the theory-theory view on this ground. It is clearly nuts to think that anyone, for the vast majority of their concepts, if not all of them, has anything approaching such explicit, well-worked-out theories. In fact, 'theory' is an unfortunate term in that it does suggest such obviously missing propositional precision and detail, and we might explore alternative ways to think about and label this 'external' dimension to conceptual structure.

For starters, imagine that concepts are embedded in Quinean webs of belief that provide some degree of explanatory coherence to stable patterns that exist in the world. Perhaps the web is everywhere interconnected, so any belief (and thus concept) can find a route to any other belief (and thus concept). Thus, if we dig deep enough, we can find presuppositions and/or entailments for each belief that allow us to traverse the entire network without ever having to skip. It is not clear that 'explanation holism' has to be true, but even if it were, it doesn't follow that distinct clusters of explanations are not powerfully linked with different concepts. All of the moons of our solar system are influenced in their orbits by all other masses in our solar system; but each planet and its moons form a coherent system distinct from any other one, constituting a system that can be almost completely understood at that level of analysis. Explanatory beliefs are not distributed evenly in the web of understanding. They form tight, richly structured clusters that then have sparse links to other clusters. Beliefs about the mechanics of solid objects, for example, are richly structured and tightly interconnected, but their connections to the cluster of beliefs about minds are comparatively very few. (It is, admittedly, awfully hard to know how to count, but by any metric that is devised, the difference would be huge.)

These clusters of beliefs are associated with stable patternings in the world such as those governing the motions of bounded solid entities or the beliefs and actions of other minds. They are not, however, remotely close to being exhaustive or complete explanations of a domain. They are skeletal 'modes of construal' that allow one to choose among competing explanations, to constrain the building of new ones, or have effective hunches about which properties are projectible in induction; and they are often implicit. They may be analogous to constraints on natural language syntax, and few people think that those constraints aren't part of the structure of the syntax itself.

The clustering of beliefs might give concepts some immunity from holism in that clusters may cause one to discount or even deny information not closely connected to them, perhaps even when such discounting causes distortions. Moreover, it is well documented that adults can live for decades with contradictory beliefs that happily coexist until the contradiction is pointed out. Many adults believe the seasons are caused by the earth's distance from the sun, while simultaneously believing that it is winter in Australia when it is summer in Britain. They are then floored when the contradiction is pointed out. This is surely a performance limitation, but perhaps it is also a natural part of our cognition that helps avoid the pitfalls of meaning holism. We may gravitate

back towards clusters of beliefs and discard those that destabilize or somehow sequester them apart. Perhaps a carefully spelled out competence/performance distinction for concepts will show us why such limitations should not be parts of the concepts themselves, but Fodor offers little guidance on this issue.

5. From the Frying Pan to the Fire: Unlocking Fodor on Natural Kinds

Recall Fodor's publicity constraint on any theory of concepts: that concepts must be widely shared by different people. Yet, in his discussion of natural kind concepts in chapter 7, Fodor claims that very few people have had natural kind concepts in the history of the species (they are a recent invention); this also seems to imply that currently few people have such concepts, but we leave this to one side. Fodor's views here reflect his dissatisfaction with the current rave over natural kind concepts, particularly in the developmental literature, as well as his denial of IRS, since it does seem to Fodor (p. 156) that in order to have the concept NATURAL KIND you do need to have other concepts such as HIDDEN MICROSTRUCTURE. (Yet if that is true, then IRS is true, and atomism false, of NATURAL KIND.)

Viewing natural kinds in terms of the deflationary understanding of theories suggested above does a better job of adhering to the publicity constraint. Having a concept of a natural kind is having some partial appreciation of the causal patternings that jointly conspire to make it a relatively stable entity (Boyd, 1999; Wilson, 1999). It is important to stress partial explanation, since no one knows all the details about the causal mechanisms governing any natural kind; indeed, among lay people the explanatory knowledge is often little more than a sense of some causal powers, and typologies of causal patterns (such as no action at a distance with bounded solid objects). But those senses can powerfully influence categorizations, conceptual combinations, patterns of concept development, induction and almost every way we use concepts. Are all of these effects caused by elements that are not parts of the concepts themselves? An attractive alternative sees those partial understandings, and how they vary by domains, as what enable us to 'resonate' with kinds in the world. The structures that allow us to lock onto the world may arise from within the concepts themselves rather than from how they relate to the world.

Thinking about conceptual structure so construed and its role in concept acquisition also seems more constructive than simply assuming the broad-based conceptual nativism that, together with a licentiousness about psychological laws, shapes up Fodor's own account of concept acquisition. Given that Fodor eschews appeals to internal and external conceptual structure, in effect his view combines two disparate elements: the idea that the relationship between natural kind concepts to kinds-in-the-world is arbitrary (like the red belly of male stickleback triggering reproductive behaviour in a female and aggression in a

male)—which is largely responsible for the doorknob/DOORKNOB problem; and the idea that the concept–kind relationship is nomological.

6. Virtual Concepts

Fodor's second constraint is that concepts are categories. He elaborates by saying that concepts apply to things in the world (p. 24). But concepts often don't apply to things in the world, and Fodor's modified informational atomism makes it difficult to see how we have concepts in cases where we cannot or do not lock onto entities, particularly when we know that we couldn't do so because those things don't exist. It has been argued that many concepts are organized around nonexistent unattainable ideals (Barsalou, 1985). The concept CIRCLE would be one such case, but perhaps also BARGAIN (costs nothing and is infinitely valuable), or THERMOS (keeps contents warm forever in a container that is neither bulky nor fragile). Moreover, there are cases where the category might exist but we know we cannot know its nature. A biologist can tell me that he knows there must have been a mammal, which he calls a 'schmoo', that existed in a certain niche 20 million years ago because of some unique mammalian genetic fragment found in amber. The fragment is just enough to indicate that it was both a mammal and different from all other known mammals, but no one has the faintest idea what sort of mammal it was. We all have the concept of the 'schmoo' but could never identify one. Even knowing how to lock onto them in possible worlds doesn't seem possible here.

In an appendix to chapter 7, Fodor suggests that the concept of 'round squares', which cannot be locked onto, must therefore be compositional out of round and square. He might argue the same here, that the schmoo concept, or circle, or thermos, is likewise compositional and so acquired. But if those concepts are compositional, why not all others? What is the principled distinction between concepts like DOORKNOB, which Fodor considers to be primitive and thus innate, and SCHMOO, which we assume is not? (Even if NATURAL KIND is late arriving on the scene, unlike SCHMOO, at least it arrived some time antecedent to your reading of the previous paragraph.) The locking account provides a meaningful answer to this question only if there is some independent, plausible account of locking, and we doubt that there is.

7. Informational Atomism and Conceptual Change

Concepts, whatever they are, seem to have the property of being tightly connected to one another as they travel along trajectories of conceptual change. Whether it be historically, in the growing child, or in a novice-to-expert shift in local domains, the elements of what used to be semantic fields travel in packs. If a young child misunderstands 'uncle' as avuncular male adults, he likely also misunderstands, in the same way, 'aunt', 'grandmother' and other

kinship terms involving social roles. When he gains the insight of uncle as a set of kinship relations, he almost immediately gains the same insight with other kinship terms. Thus, knowing that an uncle is the male sibling of one's parents is intimately linked to knowing that an aunt is a female sibling of one's parents, and that a cousin is a child of a parent's siblings. The insight to one relation tends to occur at the same time as the others. So also for cooking terms, personality terms and a host of others. Even with terms with very different senses, you run into the same problem. One cannot have the concept NUT without having the concept BOLT, as they are 'defined' in terms of each other, yet surely they are much more likely to be atomic concepts than is DOORKNOB.

Consider conceptual change and conceptual constancy over developmental time. The infant's ability to immediately see cause when one object 'launches' another may exist just because the infant shares with us powerful constraints on the kinds of explanations they are willing to entertain about bounded physical objects, explanations that may be interconnected sets of rules or beliefs (Spelke, 1994) and where concepts such as physical object, launching and trajectory are incoherent unless seen as arising out of a tightly interconnected set of such rules/beliefs. Here we can at least see how appeals to conceptual structure guide us in what to say about developmental cases of conceptual change.

One could try to reconstrue what is changing in all these cases not as part of the concept but as part of how we lock onto instances and, when we shift locking, of how we jump to different concepts; indeed one could deny the existence of all conceptual change whatsoever. Why then should all the locking relations change at the same time across a conceptual domain? One then has to posit a theory that applies to all the concepts in the domain, and which is responsible for the coordinated change in locking, but which is in no way constitutive of the concepts. Such a view is at best inelegant. More critically, it is difficult to see how information alatomism provides *any* substantive theory of conceptual change.

Fodor's book has one key virtue worth highlighting against this background of critical dissent. By proposing such a radical alternative model for concepts, he forces all of us in cognitive science to justify more carefully the assumptions underlying our approaches, and to reflect on the progress that has been made with them relative to that which might be made assuming informational atomism. This kind of reflection on what we are actually doing is much needed and Fodor is a powerful motivator.

Department of Psychology
Yale University

Department of Philosophy
University of Illinois, Urbana-Champaign

References

Barsalou, L.W. 1985: Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 11, 629–54.

Boyd, R. 1999: Homeostasis, species and higher taxa. In R.A. Wilson. (ed.), *Species: New Interdisciplinary Essays*. Cambridge, MA: MIT Press.

Fodor, J.A. 1975: *The Language of Thought*. Cambridge, MA: Harvard University Press.

Fodor, J.A. 1987: *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

Gagne, C.L. and Shoben, E.J. 1997: The influence of thematic relations on the comprehension of modifier-noun combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 71–87.

Johnson, C. and Keil, F.C. in press: Explanatory understanding and conceptual combination. In F.C. Keil and R.A. Wilson (eds), *Explanation and Cognition*. Cambridge, MA: MIT Press.

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

Keil, F., Smith, C. et al. 1998: Two dogmas of conceptual empiricism. *Cognition* 65, 103–35.

Murphy, G.L. 1988: Comprehending complex concepts. *Cognitive Science*, 12, 529–62.

Murphy, G.L. 1990: Noun-phrase interpretation and conceptual combination. *Journal of Memory and Language*, 29, 259–88.

Murphy, G. in press: Explanatory concepts. In F.C. Keil and R.A. Wilson (eds), *Explanation and Cognition*. Cambridge, MA: MIT Press.

Spelke, E. 1994: Initial knowledge: six suggestions. *Cognition*, 50, 431–45.

Wilson, R.A. 1999: Realism, essence, and kind: resuscitating species essentialism? In R.A. Wilson (ed.), *Species: New Interdisciplinary Essays*. Cambridge, MA: MIT Press.