-EDITORIAL-

AI, Concepts, and the Paradox of Mental Representation, with a brief discussion of psychological essentialism

Eric Dietrich
Philosophy Department
Binghamton University
Binghamton, NY 13902-6000

Mostly philosophers cause trouble. I know because on alternate Thursdays I am one -- and I live in a philosophy department where I watch all of them cause trouble. Everyone in artificial intelligence knows how much trouble philosophers can cause (and in particular, we know how much trouble one philosopher -- John Searle -- has caused). And, we know where they tend to cause it: in knowledge representation and the semantics of data structures. This essay is about a recent case of this sort of thing. One of the take-home messages will be that AI ought to redouble its efforts to understand concepts.

1. The Paradox of Mental Representation.

My colleague, Art Markman, and I have said, in print, that minds are locked inside mental representations (Dietrich and Markman, 2000; Markman and Dietrich, 2000). Minds do *not* check the veracity or the content or anything else about their mental representations with the world; that is impossible. Rather, minds check mental representations with other mental representations. What supplies truth, according to Professor Markman and me, is not *correspondence* to the world, but rather the *coherence* of the vast network of mental representations existing at different levels and in different modalities in the mind. (Actually, Professor Markman and I are really coherentists about *justification*, not truth. We really don't much care about truth. That is how you can tell that we aren't philosophers. Also, coherentism is not new: see the

major works by the great rationalists Leibniz, Spinoza; see also Hegel and Bradley. Most of AI tends to be quite coherentist; the possible exception is embodied cognition.)

This coherentist view of representation is philosophically and explanatorily robust. And we like it. Jerry Fodor, that curmudgeonly, reluctant ally of AI, however, hates it. One of his arguments against it is that it is self-refuting. On p. 78 of his excellent book, *In Critical Condition*, he says that "[Dietrich and Markman's view] looks to be self-refuting. If you really can't say anything about the world except as it is represented, then one of the things that you can't say is that you can't say anything about the world except as it is represented."[1] Bother. Self-refutation is such a faux pas and embarrassing to all concerned. Fortunately, Professor Markman and I have not committed this logical blunder. It is not a simple matter to show that our position is not self-refuting. The resolution of Fodor's objection requires some new insights into the nature of concepts. I discuss this in section 3.

To begin, let's convince ourselves that Fodor's claim does, at least prima facie, generate some logical tension. Let T be the proposition that:

You can't say anything about the world except as represented.

T says that we are locked inside a representational medium. Since any representation of anything whatsoever is from a perspective (image looking at someone's back -- you can't see their front), T says that we can't say anything about the world except from a perspective. All well and good, so far. T indeed seems quite plausible. But then how do we know T is true? T seems to be nonperspectival . . . to be perspective-free. To see this, note that T really says: "All representations are from some perspective or other." And in order to say this, it seems, T must not be said from one of those perspectives (for then, how could T even be uttered at all?). But those perspectives (referred by T) are all the perspectives that there are. Hence, T must be said from *no* perspective. T then is a bit of world-understanding -- of saying something about the world -- that is not perspectival. Yet T asserts that there are no such nonperspectival world-

understandings; there are no nonperspectival places from which to say something about the world. Hence T asserts that itself, T, is false.

So, we see that T does, in fact, seem to be self-refuting. (Note that if T is false (or incoherent), then you can say something about the world from outside of any representation. This is what Fodor would call "being objective.") There are a lot of self-refuting sentences and they have bothered logicians for centuries. A famous example is the paradox of the pop quiz: "There will be a pop quiz next week." Once uttered, this sentence entails that there cannot be a pop quiz next week. What makes self-refuting sentences paradoxical is that they seem true. It certainly does seem as if I can tell my class that there is a pop quiz next week and lo and behold give one. And it certainly does seem as if you can't say anything about the world except as represented. But apparently looks are misleading. The tension between appearance and logic results in these sentences being labeled *paradoxes*.

Self-refuting sentences are the poor cousins of a large and troubling set of sentences that are far worse than self-refuting: they alternate truth values infinitely, having, finally, no definite truth value at all -- if they are true then they are false, and if they are false then they are true. These sentences, too, are often called paradoxes, but their paradox is far deeper than that between appearance and logic -- their paradox is internal to logic itself. There is a long list of such alternating statements and logicians are well-acquainted with their insidious and diabolical nature. Examples include "This statement is false," "I am lying right now in uttering this statement" (the paradox of the liar, often just called the Liar), "S is the set that contains all and only those sets that don't contain themselves as members" (Russell's paradox that undid set theory). Everyone avoids these alternating sentences like the plague. In fact, set theorist had to completely redo set theory to avoid Russell's paradox. And dissertations are still be written on the Liar, though it was discovered by the ancient Greeks some two and half millennia ago.

Self-refuting sentences and alternating sentences result from the mysterious ability of things to refer to themselves. Without self-reference, none of these paradoxes

go through. For example, Russell's paradox results when we ask "Does S contain itself?" By far, the most famous self-referencing sentence is Godel's redoubtable incompleteness theorem: "The proposition with the Godel number G cannot be proven." I've written the sentence in English, but of course it actually has to be written in first-order logic with identity, the natural numbers, and the arithmetic operations added in. When that is done, it turns out that the sentence with Godel number G is the proposition "The proposition with the Godel number G cannot be proven." Godel's proposition is *not* self-refuting. Indeed, it is almost self-establishing: when coupled with the assumption that number theory is consistent, Godel's theorem establishes that number theory is necessarily incomplete: there are true propositions that cannot be proven, namely G itself. All of this is due to self-reference in one subtle form or another.

Proposition T is also self-referring in its own way. T, as we noted, says: "All representations are from some perspective or other." The self-referring part of T is hidden. The question is: "What perspective is T said from?" When the self-referring part is laid bare, we see that T says "All representations are from some perspective or other -- except me." But since T is a representation itself (or results from one), and since T quantifies over *all* representations, T contradicts itself and hence is incoherent. Hence, not all understanding must be from some perspective or other, hence not everything must understood from within some representation or other, hence minds are not locked inside representations, contra Dietrich and Markman. Hence there is the objective, the True, and the Good.

So says Fodor, at any rate.

2. Toward Dissolving the Paradox: Using the music of spheres to pump some intuitions.

But does T really say ". . . -- except me"? Upon closer inspection, it seems that T is not said from no perspective, but rather from a special perspective. What T really says is: "You can't say anything about the world except as represented in the following

N number of ways: R_1, R_2, . . . , R_N." T, then, is said from some other representational perspective not included in the set N.

What could this new representational perspective be? Ahhh, the perspective of concepts. But we aren't ready for this story yet. First, I need to get your intuitions going my way. So, to begin, recall that the ancients used to believe in the music of the spheres. Back when those in the know thought that Earth was the center of the universe and everything revolved around it, there was also the doctrine that the heavens revolved around the Earth in crystalline spheres. The moon was in a crystalline sphere of its own, the sun was in one of its own, and the stars were in one of their own (and the planets just wandered among the spheres somehow, hence their being called "planets"). All of these crystalline spheres made a kind of noise as they moved, as moving things are wont to do. This noise was the music of the spheres, and it was beautiful and divine . . . if you could hear it, but you couldn't. The question was: Why couldn't you hear it? And the answer the ancients gave was that since we were immersed in the music, and had been so for all of our lives, and had nothing to contrast it to, we couldn't perceive it -- we couldn't hear it. It's a little like noting that fish don't wonder what the ocean is, because since they are always in the ocean, they have nothing to contrast it with and so is completely in the background for them (nevermind that with their little fish brains, they probably don't wonder about anything at all).

Ok, so one intuition I want you to have is that perceiving something requires being able to contrast that thing with other things. The next intuition I want to generate is that perceiving such a contrast can be internal. That is, it would be possible to hear the music of the spheres if we could ever once hear it change tone, say by going up a note or down one. Here is another intuition pump to help with this intuition.

Suppose that you are looking at a white light through a yellow-colored filter. Suppose you have always seen this light through such a filter. So you will believe that this light is yellow, not that it is white. In short, you will not be aware of the filter. Why? Because you have nothing to contrast it to. (To get this example started, you can, if you like, imagine that the sun is the light and that there has always been a yellow-

colored filter between it and us, and that is why we see it as yellow -- which is fact the truth. The air of our atmosphere is the filter). Here's the question I want to consider: Is it possible to come to believe that the light is itself not colored, but that something else is coloring it? (Here's where imagining that the light is the sun is not a good idea, for we had independent ways of experiencing our atmosphere and hence coming to conclude that it was there influencing the colors of things.)

You will perhaps be unsurprised to learn that the answer is: Yes, we can figure out that the light is filtered. What will be surprising is what this says about concepts.

Here is one way you could come to figure out that the light is filtered. Suppose that you move to a different location and note that from this new location, the color of the light is blue. Suppose from yet another location, the color of the light is green. The light always appears in the same spot relative to you, but its color changes as you move around. What you conclude from this is that the light is not intrinsically yellow, or blue, or green. And from this you conclude that there is something invariant about the light, and this invariance is not its color (i.e., not its being yellow, blue, or green). What is invariant is independent of the light's color. Hence you surmise that perhaps the light is white and that something about the location changes the color of the light. Note that you were able to do this solely because of the contrast between colors: you didn't have to actually see the filter.

Ok, so now your intuitions are, at least to some extent, moving in the direction I want to them to. So let's consider a harder case: consider your five senses. It is possible for us to say that all of our knowledge of the world comes to us from our five senses. Yet, how can we acquire this bit of knowledge, which doesn't seem based on our five senses, if all of our knowledge is based on our five senses? (Technical point: Kant thought that this claim about our five senses was false and that in order to get knowledge from our five senses, we had to have innate, a priori intuitions (concepts) about space, time, and cause. Kant may have been right about this. Perhaps our knowledge of the world is based on our five senses plus our a priori intuitions. But even if Kant is right, this doesn't change things substantially because now we have the

question: "How can we acquire the knowledge, which doesn't seem based on our five senses plus our a priori intuitions, that all of our knowledge is based on our five senses plus our a priori intuitions?" Given that Kant's insight doesn't change things, I am just going to keep things simple, be an unblushing empiricist, and assume that all knowledge comes to us from our five senses.)

So, how can we come to know that all of our knowledge comes from our five senses when the knowledge that all of our knowledge comes from our five senses isn't knowledge that comes from our five senses? Answer: *concepts*. Here's how this works, I believe.

3. Concepts and world-making.

Concepts are, first and foremost, epistemic capacities. They are for recognizing and understanding the world. And, concepts don't, in general, compose, because epistemic capacities don't compose (knowing what a goose is and what a rake is doesn't tell you what a goose rake is). Concepts do however, combine. How they do this is one of the important puzzles cognitive scientists have yet to figure out (e.g., see the interesting paper by Costello and Keane, 2000). We do know that when they combine, they produce more, abstract concepts. Since even one's concept of something as basic as "dog" or "mother" is already an abstraction (indeed, one's concept of one's own mother is an abstraction), the results of combining concepts are abstractions of abstractions. (These points are completely anti-Fodorian. He believes that since concepts compose, they can't be epistemic capacities. One wonders what they're for, then. . . . A question Fodor doesn't answer, even in his book on concepts entitled *Concepts*, (1998b).)

Importantly, these abstractions are not just any, old abstractions. These abstractions tend to strip away information that is not important (relative to context) leaving the information that is, and the information that is increasingly central to the given concept. For example, the concept "dog" comes to embody a central capacity for determining whether something is a dog or not. This capacity could be a small set of

rules, or they could be something like exemplars or proto-types. What the owner of a dog concept has is something like "small, four-legged, furry, barking animal that is sometimes friendly, sometimes not." It is a deep point about concepts that the epistemic capacities they embody are very plastic, and not rigid at all. Hence concepts are not necessary and sufficient conditions. Hence no set of rules for recognizing dogs will suffice for all cases. We can still recognize a dog that can't bark because he has a sore throat. Rather, what happens is that concepts embody a small set of rules for recognition, and that when these fail, more complicated rules are activated and brought to bear. The central information concepts embody that is used for recognizing things in the world I will call *heuristically invariant information*, because such information functions like invariant information in other domains (e.g., mathematics: all circles are round; all prime numbers greater than 2 are odd) but, because of their plasticity, the recognitional information in concepts is only *heuristically* invariant, i.e., statistically invariant, invariant, all things being equal, not actually invariant. (For those keeping score, these points are also completely anti-Fodorian.)

Because they are abstractions, concepts provide their owner with a "higher" perspective -- a perspective that transcends the sensory information from which the concepts were derived. One way to put this point is to say that concepts have more information in them than the low-level sensory information from which they were derived. What's interesting about this point is that this information cannot be Shannon information because the information at the receiver is greater than the information at the source, which is impossible on standard, Shannon information theory (1948). Well, what kind of information could concepts contain then? It's a new kind of information: what I call *functional information* . (Not all functional information is heuristically invariant; such information can be more modality specific. But all heuristically invariant information is functional information.)

Analogues of functional information are really quite common. Think of a hammer. The Shannon style information it contains is information about its causal history (how it was built, etc.). But this information isn't what's important about a hammer. What's important about a hammer is how it is used. The reason this is an analogue to functional

information is that the hammer itself doesn't represent it's functional information, it merely has it. Put in the vernacular, hammers have a function (they are for pounding nails and related activities), but they don't know they have a function. Concepts are different. They have functional roles to play, too, but concepts also embody their functional information (concepts don't "know" their functional roles either, only whole cognitive agents can know things, but concepts do encapsulate their functional information).

The idea that concepts have functional roles and that they encapsulate or embody or explicitly represent a part of their functional roles is the cousin of an idea well-known to computer scientists. Abstract data types that are first-class objects look a lot like concepts. *Rational number* is an abstract data type that is a first-class object. Rational numbers (in computers) are defined by how they are constructed and what can be done to them (e.g., they can be added, subtracted, squared, etc.). They also can be passed as values of bound variables and returned as functional values. And, except for highly restricted cases of coercion, the data type *rational number* cannot be turned into or composed with other data types.

My claim is that concepts are abstract data types that are first-class objects (another way to put my claim is to say that concepts are like the objects in object-oriented programming). As such, concepts encapsulate the functional information about their construction and use. But unlike a lot ordinary abstract data types, concepts also encapsulate functional information about what other concepts they are linked to and how they are linked (actually, in object-oriented programming, this sort of linking connection is quite common). And, since they are capable of more and more abstracting, they come to represent heuristically invariant information about categories in the world.

Now the really radical claim (and the really anti-Fodorian claim) is this: because of their abstract data type nature, concepts provide a separate perspective from which the cognitive agent whose concepts they are can view its knowledge. From this "higher" perspective, the agent can "see" that all of its knowledge comes in through its five

senses and results in more and more abstract concepts. Concepts do this just like in the case of the yellow, blue, and green colored light. The information coming in is in different modalities (the analogue of the different colored light). As the information is combined to form a "picture of the world," various concepts are constructed and/or activated. Because the higher concepts have heuristically invariant information, they have information that is *not* modality specific. In the case of the colored lights, it was the different colors that led to the conclusion that there was a light that was not colored yellow or blue or green. What was invariant was the light itself. The color must be produced somehow between the light and you, the observer. The same thing happens with concepts. Since concepts are abstractions from different modalities (the five senses), and since they contain heuristically invariant information, concepts represent, the "thing itself," whatever thing one happens to be perceiving, e.g., a dog. And in representing the dog invariantly (albeit heuristically), they provide a perspective that is "above" or "higher" than the information from the senses (and from the lower-level concepts). It is this perspective that allows us to say: "You can't say anything about the world except as represented *by those modality specific and lower-level representations over there*." (That is, by those modality specific representations that are not higher-level concepts.)

We can now answer the question: "How can we come to know that all of our knowledge comes from our five senses when the knowledge that all of our knowledge comes from our five senses isn't knowledge that comes from our five senses?" First, as noted at the beginning section 2, this sentence is not said from no perspective, it is said from a special perspective. Hence, it is not that *all* of our knowledge comes from our five senses, it is that all of our knowledge of a certain type comes from our five senses. (Hence the sentence is *not* self-refuting, contra Fodor.) Higher-level concepts are a different type of knowledge. Concepts get their initial knowledge or information from the five senses, but since the higher-level concepts contain abstract, heuristically invariant functional information that is beyond any modality, these higher-level concepts give us knowledge that is beyond our five senses. From this perspective, therefore, we can "see" that all of our knowledge of a certain type comes from our five senses. Concepts provide us with a connection, seemingly, to the thing itself (e.g., the dog).

Since we seem to be connected to the dog itself, we can then see that our knowledge of the dog must be based on seeing the dog, hearing the dog, smelling the dog, and touching the dog (and tasting the dog, if you let it kiss you on the mouth). (I say "seemingly" because, of course, coherentism is true and we are locked inside our own representations.)

4. Psychological Essentialism.

I close by pointing out that this explanation of concepts and conceptual information explains why we think things in the world have essential properties. They don't, but it seems like they do. Why? Because our concepts have heuristically invariant information which they get via the process of abstraction. It is because our concepts have this property that we think there is such a thing as things-in-themselves. When any pressure is put on a proposed definition of a thing in itself -- necessary and sufficient conditions -- we come up empty-handed. But we never lose the feeling that such essential properties are there, being discerned by us. And this is because of our concepts and the fact that they contain more information than the world from which they were derived.

I know of no other theory of concepts that can explain why concepts seem to connect us to a world with essential properties. This seems to me to be serious argument in favor of it.

References.

Costello and Keane, (2000). Efficient Creativity: Constraint-guided conceptual combination, *Cognitive Science* 24, n2: 299-349.

Dietrich, E. and A. Markman (2000). Cognitive Dynamics: Computation and representation regained. In Dietrich, E. and Markman, A. (eds.) *Cognitive*

Dynamics: Conceptual change in humans and machines. Mahwah, NJ.: Lawrence Erlbaum.

Fodor, J. (1998a). *In critical condition*. Cambridge, MA.: MIT.

Fodor, J. (1998b). *Concepts Where cognitive science went wrong*. Oxford: Clarendon.

Markman, A. and E. Dietrich (2000). In defense of representations. *Cognitive Psychology* 40: 138-171.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal* 27: 379-423.

[1] What Fodor really says is this: "… transcendental idealism looks to be self-refuting. If you really can't say anything about the world except as it is represented, then one of the things that you can't say is that you can't say anything about the world except as it is represented." Transcendental idealism was the brainchild of Immanuel Kant -- who was apparently a philosopher of some repute.