Draft; Not for Circulation to Connectionists

.

Distributed Representations; Enough Already

Jerry Fodor

I find, as I decline into my later years that I'm increasingly irritated by what contributors to the Cog Sci literature say about distributed mental representation. It's become a sort of road rage, which, according to my doctor, isn't good for me. So I've decided to get it out of my system, Hence the present diatribe.

The rumor abroad seems to be that the notion of distributed representation is, or ought to be, central to theories of cognitive architecture. It's suggested, for example, that standard objections to neural net models of cognition apply only to the primitive versions in which distributed representation plays no role. Conversely, when the significance of distributed representation is acknowledged, a whole landscape of new options become available for empirical theorizing about concepts. So the story goes.

However, none of that is true. In fact, the discussions of distributed representation typically equivocate between ---or simply confound--- two doctrines that have been around for ages; one of which is surely true, the other of which is likely false, and neither of which is of much interest. That, anyhow, is what I'm about to argue.

I need an untendious working vocabulary; one that's neutral in respect of the issues that distributed representation raises. I propose the following ontology which will do for my purposes and which I don't think steps on anybody's toes.

For purposes of the discussion, I assume that there are:

1. *Things-in-the-world* (hereafter, `things'). These include, for example, people, cows, tables and chairs, rocks, weekends, prime numbers, and lots of other things as well. All I care about is that things are the kinds of things that concepts apply to. For example, the concept COW applies to cows (to each and every cow and to nothing else); the concept WEEKEND applies to weekends (to each and every weekend and to nothing else). The concept THIS WEEKEND applies to this weekend (and to nothing else); and so forth. All that matters is, to repeat, that things are what concepts apply to. As for the rest, have it however you like.

   Sometimes I shall speak of the `extension' of a concept, by which I will mean the set of things that it applies to. These may, for all I care, include possible things, in which case the extension of COW includes all the cows in all the possible worlds. And nothing else.

2. *Concept*s These are abstract objects that are individuated (at least inter alia) by their semantic properties.[1] Perhaps they are individuated by their extensions (or perhaps they are in individuated by the properties they express. Serious issues turn on this, but we can put them aside for the purposes at hand.)

3. Semantic properties and relations. Concepts are *satisfied* by the things they apply to; so, for example, COW is satisfied by cows (and by nothing else). So *being satisfied by* is a relation between concepts and things. It is the only semantic property we'll be much concerned with, but we may assume that it is roughly interdefinable with, for example, *being true of* and *applying to.* These too are semantic relations between concepts and things.

4. *Mental representations.* Mental representations are a species of symbols; they therefore come in both types and tokens. Token mental representations have causal powers. So, for example, maybe entertaining a token of the mental representation type `cat' causes one (by association) to entertain a token of the mental representation type `dog'. The mental representations we'll be concerned with are ones that express concepts. The mental representation `cow' expresses the concept COW and so forth.[2]

5. *Words* (of, as it might be, English). Words too are symbols, They differ from mental representations is that the former (but maybe not the latter) are expressions in public languages. Typically, words express concepts ("cow" expresses the concept COW), so we may think of a word as satisfied by things in the extension of the concept it expresses. If we do talk that way, then "cow" applies to (is satisfied by; is true of) cows.

6. *Objects.* These are whatever it is that distributed mental representations are distributed over. We'll see that the two ways of understanding distributed representation are distinguished by what they take objects to be. I suppose, in any case, that objects are a kind of thing-in-the-world. Notice, however, that objects being things-in-the-world, is compatible with their being things-in-one's-head. (Heads are in the world, and "in" is transitive).

That's all we'll need. I don't doubt that some of it is tendentious, but not, I'm pretty sure, in ways that affect the present discussion.

The thesis on the table is that lots of mental representations are distributed, and that the fact that they are matters a lot to cognitive science. Suppose, then, that the mental representation M is distributed over the objects O1, O2, O3… etc. The questions arise: what sorts of things are these objects, and in what sense is M distributed over

[1] The use of `concepts in this paper is more like the philosophers' than the cognitive scientists'. When I want to talk about concepts in the cognitive scientists' sense, I'll speak of `mental representations.' See belo0w.

[2] It's not plausible that *all* mental representations express concepts. For example, there are presumably ones that function as operators, connectives, quantifiers, singular terms, and so forth. If so, so be it.

them? I take it that the two versions of distributed representations that are floating around in the cog sci literature are distinguished by the ways that they answer these questions.[3]

First Version: Objects are neurons.

If objects are neurons, then the claim that mental representations are distributed is the claim that their tokens are typically realized by *patterns* of neurons; (eg. by patterns of interconnected neurons) . Perhaps `cow' is tokened iff a certain set of neurons fire;[4] if so, then it's got to be bigger than a unit set. Construed, this way, saying that mental representations are distributed is equivalent to denying that there is such a thing as the cow-neuron, or the grandmother neuron, etc; i.e. as denying that (in the usual case) any one concept is distributed over just one neuron.

Notice that the claim that mental representations are distributed over neurons is a claim about relations between levels of explanation. To that extent, it's exactly like the claim that tables and chairs are distributed over molecules i.e. that every chair consists of *many* molecules an so does every table; here is, in particular no such thing as the chair molecule in exactly the same sense that there is no such thing as the grandmother neuron. (The only difference is in which levels are involved. The stuff about molecules constrains the relation between macrolevel explanations and explanations that refer to (very small) parts of middle-sized objects.) Whereas, the stuff about neurons constrains the relation between intentional-level explanations (in which objects are individuated by (inter alia) their semantic properties) and neurological-level explanations (in which objects are individuated however neurons are.)

I don't, myself, consider Version One of the claim that mental representations are distributed to be very contentious. God only know what the metaphysical relation between mental things and neural things will eventually turn out to be. Maybe tokens of `cow' are, in some sense, constituted by neurons; or maybe the former merely supervene on the latter. Other possibilities are familiar from philosophical discussions of the mind/body relation. But n any case, it seems most unlikely that mental representations map one-to-one onto their neural counterparts. Nor is it news that they don't; I imagine nobody has thought otherwise for at least a century or so.

Second version: Objects are mental representations; some mental representations are distributed over others.

Notice that, unlike Version One, this account of distribution is about a within-level relation; roughly, distribution is a relation between (relatively) complex mental representations and the (relatively) simple mental representations that are their parts.

[3] Actually, there are three versions; the third of which conjoins (or conflates) the first two. Nothing interesting comes of this, however, so I won't bother with it.
[4] To ease the expositions, I'll mostly pretend that we think in English (though, of course, we don't.)

That complex concepts are somehow made out of simple ones is quite an old idea; the Anglophone tradition in philosophy and psychology is committed to it almost without exception. There is, however, a variety of ways in which it can be formulated. For example, "distributed over" might mean something like *defined by*, so the mental representation `bachelor' is, in this sense, distributed over the mental representations `unmarried' and `man'.[5] That view comports naturally with the claim that the concept BACHELOR is some sort of construction out of the concepts UNMARRIED and MAN (i.e out of the concepts which the mental representations `unmarried' and `man' express. . It likewise comports with the view that the English word "bachelor" is represented by a complex array of features[6] "at the semantic level" of linguistic description; and with the view that there are relations of `conceptual entailment' that hold between a complex concept and its constituents. Eg, it's because BACHELOR is made out of UNMARRIED MAN that 'bachelors are unmarried' is a necessary truth. To be sure, none of these options are forced; probably one might hold any of them without any of the others. But they form a familiar family, which is alive and flourishing not just in psychological theories of perception and concept acquisition, but also in such linguistic disciplines as `lexical semantics'.

There's another dimension along which formulations of the `Second Version' of the distribution thesis can differ. I've been saying that some concepts (and hence the corresponding mental representations) are complexes that are constructed out of other concepts. But I haven't said what notion of construction is supposed to be germane. In fact, I've sort of taken for granted that it's something like a *logical* construction; so the concept BACHELOR has the internal structure of a conjunction; it has the structure UNMARRIKED AND MAN. (Likewise, mutatis mutandis for the mental representation `bachelor'. But, in fact, you don't have to run the story this way if you don't want to. You could say, instead, that some concepts are *statistical* constructions out of others. This is the heart of the idea that many concepts consist of prototypes, paradigms or the like, together with a similarity metric which, together with the paradigm, determines their extension (insofar as their extensions are determinate.) So, the concept COW consists of a cluster of 'features' (viz of concepts that express properties that cows reliably exhibit (+ gives milk, + plus says moo, -flies, +has flies)), together with some way of evaluating

[5] Or one might say that the mental representation `bachelor' *is* the mental representation `married man'; hence that, appearance to the contrary notwithstanding, `bachelor' turns out to be a complex mental representations. Here too there are issues that turn on such niceities of formulation, but not such as we need to explore.

[6] A great deal of nonsense is talked about features. In fact, they are nothing more than open sentences of `Mentalese' (or whatever metalanguage is used to specify the semantic properties of mental representations. So to say that the concept COW has the feature `+ animal' is to say that COW is complex and that ANIMAL is one of its constituents. (Likewise, mutatis mutandis, for the claim that the mental represention `cow' contains the feature `+ animal'. This is to say that, once an architecture recognizes open sentences in Mentalese, nothing further is added by recognizing semantic features.

Usually semantic features are supposed to be binary; but they might, in principle, be allowed to take on any number of values you please. Here again nothing is at issue from our point of view. For example, whether or not semantic features are binary, it is OK to think of them as "vectors" in a semantic space. Doing so alters nothing (and adds nothing),

how closely a given cow approximates this prototype Mutatis mutandis, as usual, for the structure of the corresponding mental representations.

That, as far as I know, exhausts the things people have in mind when they speak of distributed representations. As previously remarked, on the First Version of the claim, it is untendentious and (not very interesting.) If concepts have neural correspondents at all, then surely the typical neural correspondent of the typical concept is complex under anatomical description.

In the Second Version, the claim that concepts are distributed is denied only by people (like me) who favor some sort of conceptual atomism. By and large, nobody in philosophy or cognitive science takes such people seriously.

In any case, both versions of the distribution thesis can be, and often have been, formulated without appealing to the notion of distribution *at all*. That is, in neither case does the notion of distribution allow us to do things that we are unable to do without it. In particular, *nothing* is gained by to saying that some mental representations (or some concepts) are `distributed over' others instead of saying what we used to: that some mental representations (/concepts) are `constructed from' others. Nothing has changed except terminology (except for the level of ambient confusion.)

So, then, the situation seems to me to be this: *Either* I have misinterpreted the thesis that concepts, mental representations and the like are typically distributed, *or* that thesis is of no interest except as an abbreviation for ideas that are commonplace in traditional and current theories of cognition that make do without the postulation of distributed representations. Correspondingly, in the best of worlds, what would happen now is *either* that somebody kindly explains to me how I've gotten the notion of distributed representation wrong (in which case, I owe an abject apology) *or* everybody stops going on about distributed representation (which would please me no end.) This is not, however, the best of worlds. So what is most likely to happen is that nobody explains to me how I've misconstrued the notion distributed representations, but everybody keeps going on about them all the same. It seems that, as I decline into my later years, I've become not just irritable, but also sort of bitter.