Dietrich, E. (2001). It Does So! Review of Jerry Fodor's *The Mind Doesn't Work That Way: The scope and limits of computational psychology*. *AI Magazine* v.22, n. 4, pp. 141-144.

It Does So.

Review of

The Mind Doesn't Work That Way:

The scope and limits of computational psychology

by Jerry Fodor, MIT Press, Cambridge, MA, 2000, 126 pages, $22.95.

ISBN: FODNH 0-262-06212-7

Eric Dietrich

Dept. of Philosophy

Virginia Tech

Blacksburg, VA 24061

SAT question: Which one of the following doesn't belong with the rest?

Astrology,

Evolutionary Biology,

Physics

Artificial Intelligence/Cognitive Science

Tarot

Answer: Physics.

Why? Because it is the only discipline in the list that is not under attack for being conceptually or methodologically confused.

Objections to AI and computational cognitive science are myriad. Accordingly, there are many different reasons for these attacks. But all of them come down to one simple observation: humans seem a lot smarter that computers -- not just smarter as in Einstein was smarter than I, or I am smarter than a chimpanzee, but more like I am smarter than a pencil sharpener. To many, computation seems like the wrong paradigm for studying the mind. (Actually, I think there are deeper and darker reasons why AI, especially, is so often the brunt of polemics, see Dietrich, 2000.) But the truth is this: AI is making exciting progress, and will one day make a robot as intelligent as a person; indeed the robot will be conscious. And all this is because of another truth: the computational paradigm is the best thing to come down the pike since the wheel.

Jerry Fodor believes this latter claim. He says:

> [The computational theory of mind] is, in my view, by far the best theory of cognition that we've got; indeed, the only one we've got that's worth the bother of a serious discussion. . . . [I]ts central idea -- that intentional processes are syntactic operations defined on mental representations -- is strikingly elegant. There is, in short, every reason to suppose that Computational Theory is part of the truth about cognition (p. 1, all references are to the book under review, unless otherwise noted).

The rub is that, whereas quite a few AI researchers and cognitive scientists think that computationalism is either *all* or *most* of the truth, Fodor thinks it is only a small part of the truth. This is what this short book is about. It is a fascinating read.

This dispute about quantity of truth is where the book gets its title. In 1997, Steven Pinker published an important book describing the current state of the art in cognitive science (see also, Plotkin 1997). Pinker's book is entitled *How the mind works*. In it, he describes how computionalism, psychological nativism (the idea that many of our concepts are innate), massive modularity (the idea that most mental processes occur within a domain-specific, encapsulated special-purpose processor), and Darwinian adaptationism combine to form a robust (but nascent) theory of mind. Fodor, however, thinks that the mind doesn't work that way, or anyhow, not very much of the mind works that way.

Fodor dubs the synthesis of computationalism, nativism, massive modularity, and adaptationism the *New Synthesis* (p. 5). He distinguishes New Synthesis nativism from his preferred Chomskian nativism (the former assumes cognitive mechanisms are innate; the latter assumes that knowledge of various types is innate), massive modularity from his preferred partial modularity, and denies that human (or anyone else's) cognitive architecture is an evolutionary adaptation. Fodor doesn't think we were created, of course, instead he thinks that our architecture is probably a saltation: a discontinuous mutational change in a phenotypic

trait (i.e., evolution by jumps, rather than by gradual, small transitions; the latter being the hallmark of classical adaptationism).

All of this is has profound consequences for theories of mind, thinking, and intelligence, which Fodor draws out, admirably. Here, I will concentrate on the most important: that much of the human mind -- the best part of it, as it turns out -- isn't computational. Fodor says: "Indeed, I am inclined to think that, sooner or later, we will *all* have to give up on the Turing story [of computation] as a general account of how the mind works, and hence, a fortiori, that we will have to give up on the generality of New Synthesis cognitive science" (p. 47).[1]

Fodor claims that the best theory of the mind that we've got is seriously inadequate; it works only for the perceptual and motor modules, and certain parts of language understanding and production, but not for higher cognition, i.e., intelligence. Fodor's Eeyoresque conclusion from this is that we have a very long way to go before we understand the mind and are able to build a replica of it, and the replica won't be a computer of any sort. He says: "So far, what our cognitive science has found out about the mind is mostly that we don't know how it works" (p. 100). (Interestingly, in the last chapter of his seminal *Language of Thought* (1975), Fodor also strongly delimits the computational theory of mind. And he delimited it to more or less the same area he does in his current book. The difference now is that his arguments for such limitations are stronger.) There is some good news, however. Fodor agrees that computationalism is correct for the perceptual parts of the mind (the perceptual modules) and language, so to explain the rest of the mind, we

will *not* have to abandon computationalism, we will simply have to add to it. But, we will have to add something radically new and not computational in the least.

Fodor's fundamental argument is that the *frame problem* is why the part of the human mind responsible for confirming beliefs (hypotheses about the world), planning, creativity, and analogy, concept learning, etc. is not computational. Fodor defines the frame problem as the problem of "[h]ow to make abductive inferences that are both reliable and feasible…" (pp. 37-38; see also Fodor, 1987). (Abductive inference is also known as inference to the best explanation.) This is a particularly robust and philosophical version of the frame problem, and widely reputed to be miles from the original version of the frame problem (see McCarthy and Hayes, 1969; and Hayes, 1987), but it is still a very serious problem and one that needs to be taken seriously by AI researchers.

But they do take it seriously. There is a lot research in AI on abductive inferencing, carried on at many different levels of generality. It goes by many names: causal reasoning, commonsense reasoning, case-based reasoning, nonmonotonic reasoning, default reasoning, qualitative physics, reasoning under uncertainty, etc. All of these research areas have in common the idea that information relevant to the problem, goal, or task at hand has to be culled, via some sort of relevance measure, from all the information available to the system. There are many kinds of relevance measures, but, following Fodor, all relevance measures can be assumed to be *heuristic guesses* as to what information is relevant to the current job and what is not. This is why the nature of heuristics and

heuristic reasoning looms large in AI research on abductive inference. Fodor, however, argues that heuristics cannot be used to implement abductive inference. Since using heuristics is, according to Fodor, the only conceivable way, in classical AI, to implement abductive reasoning, it follows that AI will never implement abductive reasoning, and so will fail in its attempt to replicate human-level intelligence[2]. Fodor says "...the computational theory of mental processes doesn't work for abductive inferences" (p. 41).

The reason for this, he says, is that abductive inferences can be, and frequently are, sensitive to information in a knowledge base antecedently deemed to be irrelevant to the inference. He calls this property of being "globally" sensitive to all information in a knowledge base, even the presumably irrelevant information, "globality" (p. 28, 30). Consider just one case from research on analogy: who would have thought before the fact that the orbits of comets around the sun would be relevant to figuring out the structure of the atom. But it was relevant. Just ask Rutherford (or rather, read his notes; and also read Gentner and Wolff, 2000). However, computation, so Fodor claims, is sensitive only to local, syntactic facts of information previously established to be relevant (he never proves this, by the way, he just asserts it; indeed, he thinks it is true by definition, following immediately from Turing's definition of computation, p. 30; however, it is certainly *not* true by definition.) Genuinely intelligent minds, unlike computation, are not so restricted, he claims (again, he never argues for this claim, or makes it clear how he knows this -- he couldn't point to the dearth of analogy machines, for there are lots of them). Fodor, doesn't even take a wild guess as to how

our minds perform what for him is the nearly magical feat of abductive inference, but this is entirely consistent with his view of the matter: it would be like Aristotle wildly guessing how the sun worked.

Fodor's rejection of heuristics is, to my mind, the central move in his argument against the computational theory of mind. He rejects heuristics as a way of making abductive inferences because they don't solve the problem of how to implement abduction, they merely move it to another location. In the crucial passage, he says:

> So the suggestion on offer is that mental processes effect local, heuristic approximations of the global determination of abductive inference. And the prima facie objection to this suggestion is that it is circular if the inferences that are required to figure out which local heuristic to employ are themselves often abductive. Which there is every reason to think that they often are. (p. 42)

In other words, deciding which heuristic to deploy in a given instance of abductive inference requires abductive inferencing. So we can't use heuristics to solve the problem of abductive inference, since abductive inferencing is required to solve the problem of which heuristics to use. Bummer.

This is a very curious objection, however. It causes an infinite regress reminiscent of the problem raised in Lewis Carroll's "What the tortoise said to Achilles" (1895), (Fodor notes this himself, p. 44).

However, this problem is as bad for Fodor as for rabid computationalists like Pinker and AI researchers. Well, actually, that's not true. It is *worse* for Fodor than for computationalists, because they can solve it, and he can't (oddly, he notes this, too, p. 45). So, we need to look at Carroll's paradox up close.

Suppose that you show the following argument to someone, say someone named T:

1). If two things are equal to a third thing, then they are equal to each other.
2). 2+2 and 2*2 are equal to a third thing, namely 4.
100). Hence, 2+2 and 2*2 are equal to each other.

T accepts that 1) is true and accepts that 2) is true, but doesn't accept that 100) is therefore true. It is tempting to dismiss T as inferentially challenged and recommend that he take to football (which is what Carroll recommends). But suppose you take it upon yourself to convince T that 100) must be accepted. What should you do? The obvious move is to get T to accept:

3). If 1) is true and 2) is true, then 100) must be true.

Then to point out to T that the following inference is logically sound:

1). If two things are equal to a third thing, then they are equal to each other.

2). 2+2 and 2*2 are equal to a third thing, namely 4.

3). If 1) is true and 2) is true, then 100) must be true.

100). Hence, 2+2 and 2*2 are equal to each other.

Suppose that T does accept 3) (i.e., he believes 3) is true), but tells you that though he now accepts 1), 2), and 3), he doesn't accept 100). Again, you point out that if someone accepts 1), 2), and 3), then they have to accept 100. T says "Well, I see that 1), 2), and 3) are true, but I don't see why I should accept the further inference you just stated, namely that if someone accepts 1), 2), and 3), then they have to accept 100)." T's position is that though he accepts 1), 2), and 3), since he fails to see the truth of the proposition you just uttered, "if someone accepts 1), 2), and 3), then they have to accept 100," then he is under no obligation to accept 100).

Approaching exasperation, you ask T to accept this new proposition. He does so willingly, by adding it to the list:

1). If two things are equal to a third thing, then they are equal to each other.

2). 2+2 and 2*2 are equal to a third thing, namely 4.

3). If 1) is true and 2) is true, then 100) must be true.

4). If 1) is true and 2) is true and 3) is true, then 100) must be true.

100). Hence, 2+2 and 2*2 are equal to each other.

Now you say to T: "Surely you accept 100) now, for anyone that accepts 1), 2), 3, and 4), must accept 100)." T replies that, yes, he does in fact now accept 1), 2), 3), and 4), but doesn't see why he should accept this last proposition you uttered, namely that if anyone that accepts 1), 2), 3, and 4), must accept 100). T points out that it is this last proposition that seems to be crucial to accepting 100). So he doesn't accept statement 100).

Sensing the despair of infinity, you add this last proposition to the list, and ask T to accept it:

5). if 1) is true and 2) is true and 3 is true and 4) is true, then 100) must be true.

T readily agrees to this proposition, but doesn't see why he should accept the next meta-proposition. . . Ad infinitum. T is under no logical obligation to accept 100) because there is always a proposition, an inference, in the meta-language used to describe the ever-lengthening chains which T can safely, i.e., logically, refuse to accept. Hence, 100) need never be accepted[3].

A shorter version of Carroll's problem is that in order to decide, one first has to decide to decide, but in order to do that one has to decide to decide to decide, etc. etc. Hence, one can never make a decision.

But of course, humans and machines can and do short-circuit this infinite regress by the process of the *immediate inference*: we can just

see that B is true given that (A implies B) and A are true. Immediate justification, like this, is a brute fact. And it has been well-known for some time now that the way to handle immediate justification is to assimilate it to perception. Standing in the Jackson Hole valley in western Wyoming, I don't need to justify that I see the Grand Tetons beyond just noting that I see them.

There are deep issues here to be sure (some involving consciousness), but there is nothing in the nature of heuristics that prevents computers and humans from using them to do abductive reasoning. It is just that the abductive reasoning will always be defeasible.

Here's a good way to put the point. Clark Glymour (personal communication) has pointed out that it is standard in machine learning to use the following rule: In any new context or domain, test a variety of heuristics on a subsample and then apply the best performing heuristics to predict new cases in the whole domain. Now, it is true that this rule is itself a heuristic, but we do not need yet another heuristic to deploy it; we just deploy it.

Fodor is aware of all this in his book. He says: "The relevant considerations are much of the sort that arose in Achilles' famous discussion with the tortoise" (p. 44). He even says:

> The reason [a computer] is able to [get B from ((A → B) & A)], the tortoise to the contrary notwithstanding, is basically this: Given a

> derivation which includes formulas of the form A and A → B, the detachment of B is effected automatically by an architectural process. . . (p. 44).

It is very puzzling, therefore, why Fodor can't see that it is the immediate inference that saves both humans and computers. But to the Fodor-phile, the answer is available. Fodor is a flaming neo-rationalist. Neo-rationalism is the view that many of our most important concepts are innate, and that reason is the primary source of knowledge, not the world. Fodor has been quite explicit that he is out to free cognitive science from the reigning empiricism (see, e.g., Fodor, 1998, and the review by Giesy and Dietrich, 2001). Empiricism is the view that almost all of our concepts are learned, and in general, it is the view that our knowledge is based on experience of, and sensorimotor interaction with, the world. Only an empiricist is going to be much impressed by the power of the immediate inference in *all* of cognition. Put another way, only an empiricist is going to see lots of cognition, especially higher cognition, as interestingly similar to perception. Rationalists draw a sharp distinction between the processes of higher cognition and those of perception. Fodor is the strongest advocate for drawing this distinction. This is, indeed, of the main themes of this book. Since Fodor hates empiricism root and branch, he cannot see that lots of cognition, especially higher cognition, is interestingly similar to perception.

So, Fodor, our latter-day Zeno, opts for a radical and pessimistic conclusion. Since we can't use heuristics (or connectionism, see footnote 2), he says the working cognitive scientist should:

> . . . concentrate one's research efforts in those areas of cognitive processing where the effects of globality [the property that the confirmation of any hypothesis might require information from a domain antecedently deemed to be irrelevant] are minimal; minimal enough, indeed, so that they can be ignored . . . . (p. 53)

These areas include only perpetual domains, those domains, such as syntactic language processing and early vision, that arguably are handled by special-purpose, domain specific processors (modules). In short, Fodor is advocating that we abandon the quest to build an intelligent robot and concentrate on building only vision machines or hearing machines or parsing machines. On his view, genuine, human-level intelligence requires some sort of processing that is not computational and hence completely mysterious to us, now and for the foreseeable future.

Well, believe it if you want. But in the next century or two (assuming our technological society survives), our descendents will be hobnobbing, dating, and otherwise communing with intelligent machines. Fodor's great great grandchildren will appreciate the irony of this . . . and so will their friends -- the machines.[4]

References

Carroll, Lewis. (1895). What the tortoise said to Achilles. *Mind*, 4, 278-280.

Dietrich, E. (2000). Cognitive Science and the Mechanistic Forces of Darkness, or Why the Computational Science of Mind Suffers the Slings and Arrows of Outrageous Fortune. *Techné*: eJournal of the *Society for Philosophy and Technology*, Winter, http://scholar.lib.vt.edu/ejournals/SPT/v5n2/dietrich.html

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

Fodor, J. (1987). Modules, frames, fridgeons, sleeping dogs, and the music of the spheres, in *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*, Z. Pylyshyn (ed.), Norwood, N.J.: Ablex.

Fodor, J. (1975). *The Language of Thought*. New York: Crowell.

Gentner, D. and Wolff, P. (2000). Metaphor and knowledge change, in *Cognitive Dynamics: Conceptual and Representational Change in Humans and Machines*, E. Dietrich and A. Markman (eds.), Mahwah, N.J. Lawrence Erlbaum, 295-342.

Giesy C. and Dietrich, E. (2001). Review of *Concepts: Where cognitive science went wrong*. *Cognitive Science Society Newsletter*. http://cognitivesciencesociety.org/newsletter

Hayes, P. (1987). What the frame problem is and isn't. In Z. Pylyshyn (ed.), *The Robot's Dilemma*. Norwood, N. J., Ablex.

McCarthy, J. and Hayes, P. (1969). Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and D. Michie (eds.), *Machine Intelligence* 4, Edinburgh, Scotland, Edinburgh Univ. Press.

Pinker, S. (1997). *How the mind works*. New York: Norton.

Plotkin, H. (1997). *Evolution in mind*. London: Alan Lane.

---

[1] Though the book is short, there is a lot in it. Fodor's discussion of evolutionary psychology, adaptationism, and modularity (in chapters 4, 5, and the appendix) is fascinating and, as always, controversial. But his arguments in this part of the book, seem much more convincing than the part I am going to focus on (chapter 1, 2, and 3). More importantly, the conclusions from the later part of the book are not the assault on AI that the conclusions are from the first part of the book. This is why I'm focussing on the first.

[2] I say "classical AI' because Fodor also considers and rejects connectionism as a route to machine abduction. But, in truth, connectionism doesn't add any genuine alternative in this context. The issues here are not about cognitive architecture, but rather about relevance measures. Connectionists have just as hard a time with this as anyone else.

[3] There is an obvious parallel between Zeno's paradoxes and Carroll's paradox (hence the name of Carroll's article). But, as pointed out to me by Chandrasekaran (personal communication), there is an important difference, too: solving Zeno's paradoxes required seeing something new about numbers and the summing of infinite series. But solving Carroll's paradox doesn't require seeing something new about logic, but rather coming to appreciate a capacity humans have: the capacity to believe our eyes.

[4] Thanks to Prof. Chandrasekaran for good comments on an earlier draft of this review.