Jerry Fodor

How the mind works: what we still don't know

One could make a case that the history of cognitive science, insofar as it's been any sort of success, has consisted largely of finding more and more things about cognition that we didn't know and didn't know that we didn't. 'Throwing some light on how much dark there is,' as I've put it elsewhere. The professional cognitive scientist has a lot of perplexity to endure, but he can be pretty sure that he's gotten in on the ground floor.

For example, we don't know what makes some cognitive states conscious. (Indeed, we don't know what makes *any* mental state, cognitive or otherwise, conscious, or why any mental state, cognitive or otherwise, bothers with being conscious.) Also, we don't know much about how cognitive states and processes are implemented by neural states and processes. We don't even know *whether* they are (though many of us are prepared to assume so *faut de mieux*). And we don't know how cognition develops (if it does) or how it evolved (if it did), and so forth, very extensively.

In fact, we have every reason to expect that there are many things about cognition that we don't even know that we don't know, such is our benighted condition.

In what follows, I will describe briefly how the notions of mental process and mental representation have developed over the last fifty years or so in cognitive science (or 'cogsci' for short): where we started, where we are now, and what aspects of our current views are most likely to be in need of serious alteration. My opinions sometimes differ from the mainstream, and where they do, I will stress that fact; those are, no doubt, the parts of my sketch that are least likely to be true.

The 1950s 'paradigm shift' in theories of the cognitive mind, initiated largely by Noam Chomsky's famous review of B. F. Skinner's book *Verbal Behavior*, is usually described in terms of a conflict between 'behaviorism' and 'mentalism,' from which the latter emerged victorious. Behaviorists thought something

Jerry Fodor is State of New Jersey Professor of Philosophy at Rutgers University. He is the author of several publications in philosophy and cognitive science, including "Modularity of Mind" (1983), "A Theory of Content and Other Essays" (1990), and "The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology" (2000).

was methodologically or ontologically controversial about the claim that we (and, presumably, other advanced kinds of primates) often do the things we do because we believe and desire the things we do. Chomsky's reply was, in essence, 'Don't be silly. Our behavior is characteristically caused by our mental states; therefore, a serious psychology must be a theory about what mental states exist and what roles they play in causing our behavior. You put gas in the tank because you believe that, if you don't, the car will grind to a stop, and you don't want the car to do so. How could anyone sane believe otherwise?'

That was, to put it mildly, all to the good. Behaviorism never was a plausible view of the methodology of psychology, any more than instrumentalism was a plausible view of the methodology of physics. Unsurprisingly, the two died of much the same causes. Many of the arguments Chomsky brought against the proposed reduction of the mind to behavior recall arguments that Carl Hemple and Hilary Putnam brought against the proposed reduction of electrons (to say nothing of tables and chairs) to 'fictions' or 'logical constructions' out of sensory experience. 'Don't be silly,' they said. 'Sensations and the like are mind-dependent; tables and chairs are not. You can sit on chairs but not on sensations; a fortiori, chairs can't *be* sensations.' Chomsky's realism about the mental was thus part of a wider realist agenda in the philosophy of science. But it's important to distinguish (as many of us did not back in those days) Chomsky's objections to Skinner's *behaviorism* from the ones he raised against Skinner's *associationism*. In retrospect, the latter seem the more important.

Behaviorism was and remains an aberration in the history of psychology. In fact, the mainstream of theorizing about the mind (including both philosophical empiricists and philosophical rationalists, and the 'sensationist' tradition of psychologists like Wilhelm Wundt and Edward Tichner) wasn't behavioristic. Rather, it was a mentalistic form of associationism that took the existence of mental representations (what were then often called 'Ideas') and their causal powers entirely for granted. What associationism mainly cared about was discovering the psychological laws that Ideas fall under. And the central thesis – which, Hume said, was to psychology what gravitation was to Newtonian physics – was that Ideas succeed one another in cognitive processes according to the laws of association.

For nearly three hundred years, associationism was the consensus theory of cognition among Anglophone philosophers and psychologists. (It's still the view assumed by advocates of 'connectionism,' a movement in cognitive science that hopes to explain human intellectual abilities by reference to associations among 'nodes' in 'neural networks,' the latter corresponding, more or less, to Ideas and the former corresponding, more or less, to minds that contain them. If, in fact, you take away the loose talk about 'neurological plausibility,' the connectionist's account of cognition is practically indistinguishable from Hume's.) Associationism was widely believed to hold, not just for thought but for language and brain processes as well: thoughts are chains of associated concepts, sentences are chains of associated words, and brain processes are chains of associated neuron firings. In all three cases, transitions from one link in such a chain to the next were supposed to be probabilistic, with past experience determining the probabilities according to whatever Laws of Association happened to be in fashion.

These Anglophone theorists notwithstanding, it's been clear, at least since Kant, that the associationist picture can't be right. Thoughts aren't *mere* sequences of ideas; at a minimum, they are *structured* sequences of ideas. To think 'there's a red door' isn't to think first about red and then about a door; rather, it's to think *about* a door *that* it is red. This is, I suppose, a truism, though perhaps not one that the cogsci community has fully assimilated. Likewise, sentences aren't just lists of words. Instead, they have a kind of internal structure such that some of their parts are grouped together in ways that others of their parts are not. Intuitively, the grouping of the 'parts' of 'the red door opened' is: [(the) (red door)] (opened), not (the red) (door opened).

So sentences have not just lexical contents but also constituent structures, consisting of their semantically interpretable parts ('the red' doesn't mean anything in 'the red door opened,' but 'the red door' does). One of the main things wrong with associationism was thus its failure to distinguish between two quite different (in fact, orthogonal) relations that Ideas can enter into: *association* (a kind of causal relation) and *constituency* (a hierarchical kind of geometrical relation). Ironically, as far as anybody knows, the first isn't of much theoretical interest. But the second, the constituency relation, does a lot of the heaviest lifting in our current accounts of cognition.

For example, as Chomsky famously pointed out, sentences are 'productive.' The processes that construct sentences out of their parts must be recursive: they must be able to apply to their own outputs, thereby generating infinite sets. It turns out that these recursions are defined over the constituent structures of the expressions they apply to. Typically, they work by embedding a constituent of a certain type in another constituent of the same type, like a sentence within a sentence. (For example, the sentence 'John met the guy from Chicago' is some sort of construction out of the sentences 'John met the guy' and 'the guy is from Chicago,' with the second sentence embedded in the first.) The same sort of story goes for mental representations, since they are also productive: if there weren't boundlessly many thoughts to express, we wouldn't need boundlessly many sentences to express them.

Their potential for productivity isn't, of course, the only thing that distinguishes constituent structures from associative structures. The strength of the association between Ideas is traditionally supposed to depend largely on the frequency and spatiotemporal contiguity of their tokenings. In contrast, as Chomsky also pointed out, the structural relations among the constituents of a complex representation hold for novel representations as well as for previously tokened ones, and are typically independent of the propinquity of the relata.

In sum, by far the most important difference between the traditional theories of mind and the ones those of us who *aren't* connectionists endorse is the shift from an associationist to a constituent, or *computational*, view of cognition.

Here, then, are the two basic hypotheses on which the current computational theory of cognition rests:

First, mental representations are sentence-like rather than picture-like. This stands in sharp contrast to the traditional view in which Ideas are some kind of images. In sentences, there's a distinction between *mere* parts and constituents, of which the latter are the *semantically interpretable* parts. By contrast, every part of a picture has an interpretation: it shows part of what the picture shows.

Second, whereas associations are operations on parts of mental representations, computations are operations defined on their constituent structures.

So much for a brief (but, I think, reasonably accurate) summary of what most cognitive scientists now hold as a working hypothesis about mental representations and mental processes (except, to repeat, connectionists, who somehow never got beyond Hume). Now, the question of interest is, for how much of cognition is this hypothesis likely to be true? The available options are *none of it*, *some of it*, and *all of it*.

My guess is: at best, not very much of it. This brings me to the heart of this essay.

Here's what I'm worried about. As we've been seeing, constituent structure is a species of the part/whole relation: all constituents are parts, though not vice versa. It follows that constituency is a *local* relation: to specify the parts of a thing you don't need to mention anything that's *outside* of the thing. (To specify the part/whole relation between a cow and its left leg, you don't have to talk about anything outside the cow. Even if this cow were the only thing in the whole world, it would bear the same relation to its left leg that it bears to its left leg in this world. Only, in that world, the cow would be lonelier.) Likewise for representations – mental representations included. Since constituents are parts of representations, operations defined on constituents apply solely in virtue of the internal structure of the representations.

The question thus arises: are there mental structures, with mental processes defined on them, that aren't local in this sense? If there are, then we are in trouble because, association having perished, computation is the only notion of a mental process that we have; and, as we've just seen, computations are defined over local properties of the representations that they apply to.

Well, I think there's pretty good reason to suppose that many of the mental processes crucial to cognition are indeed *not* local. So I guess we're in trouble. It wouldn't be the first time.

There are at least two pervasive characteristics of cognitive processes that strongly suggest their nonlocality. One is their sensitivity to considerations of relevance; the other is their sensitivity to the 'global' properties of one's cognitive commitments. It is very easy to run the two together, and it's a common practice in cogsci literature to do so. For polemical purposes, perhaps nothing much is lost by that. But some differences between them are worth exploring, so I'll take them one at a time.

Consider the kind of thinking that goes on in deciding what one ought to believe or what one ought to do (the same considerations apply both to 'pure' and to 'practical' reason). In both cases, reasoning is typically *isotropic*. In other words, *any* of one's cognitive commitments (including, of course, currently available experiential data) is relevant, in principle, to accepting or rejecting the options – there is no way to determine, just by inspecting an empirical hypothesis, what will be germane to accepting or rejecting it. Relevance isn't like constituency – it's not a local property of thoughts.

So how *does* one figure out what's relevant to deciding on a new belief or plan? That question turns out to be very hard to answer. There is an infinite corpus of prior cognitive commitments that might prove germane, but one can actually visit only some relatively small, finite subset of them in the 'real time' during which problems get solved. Relevance is long, but life is short. Something, somehow,

must 'filter' what one actually thinks about when one considers what next to believe or what next to do.

Hence the infamous 'frame problem' in theories of artificial intelligence: how do I decide what I should take to be relevant when I compute the level of confidence I should invest in a hypothesis or a plan? Any substantive criterion of relevance I employ will inevitably risk omitting something that is, in fact, germane; and one of the things I want my estimate to do (all else equal) is minimize this risk. How on earth am I to arrange that?

I think the frame problem arises because we have to use intrinsically local operations (computations, as cogsci currently understands that notion) to calculate an intrinsically nonlocal relation (relevance). If that's right, the frame problem is a symptom of something deeply inadequate about our current theory of mind.

By contrast, it's a widely prevalent view among cognitive scientists that the frame problem can be circumvented by resorting to 'heuristic' cognitive strategies. This suggestion sounds interesting, but it is, in a certain sense, empty because the notion of a heuristic procedure is negatively defined – a heuristic is just a procedure that only works from time to time. Therefore, everything depends on *which* heuristic procedure is alleged to circumvent the frame problem, and about this the canonical literature tends to be, to put it mildly, pretty causal.

Here, for example, is Steven Pinker, in a recent article, explaining what heuristics investors use when they play the stock market: "Real people tend to base investment decisions on, among other things, what they hear that everyone else is doing, what their brother-in-law advises, what a cold-calling stranger with a confident tone of voice tells them, and what the slick brochures from large investing firms recommend. People, in other words, use heuristics."[1]

Pinker provides no evidence that this is, in fact, the way that investors work; it's a story he's made up out of whole cloth. At best, it's hard to see why, if it's true, some investors make lots more money than others. But never mind; what's really striking about Pinker's list is that he never considers that *thinking about the stock market* (or paying somebody else to think about it for you, if you're lazy like me) might be one of the 'heuristics' that investors employ when they try to figure out whether to buy or sell. Ironically, thinking seems largely to have dropped out of heuristic accounts of how the mind works. Skinner would have been greatly amused.

There have been, to be sure, cases when cognitive scientists have tried to tell a story about the use of heuristic strategies in cognition that amounts to more than the mere waving of hands. To my knowledge, the heuristic most often said to guide decisions about what action to perform or belief to adopt is some version of 'if things went all right with what you did last time, do the same again this time.' We owe a rather opaque formulation to the philosopher Eric Lormand: "A system should assume *by default* that a fact persists, unless there is an axiom specifying that it is changed by an occurring event .... [G]iven that an event E occurs in situation S, the system can use axioms to infer new facts existing in S+1, and then simply 'copy' the remainder of its beliefs about S over to S+1."[2] Likewise, Peter Carruthers says

1 Steven Pinker, "So How Does the Mind Work?" *Mind and Language* 20 (1) (February 2005): 1 – 24.

2 Zenon W. Pylyshyn, ed., *Robot's Dilemma: The Frame Problem in Artificial Intelligence* (Norwood, N.J.: Ablex, 1987), 66.

that "there's no reason why the choices [about what to do next] couldn't be made by higher-order heuristics, such as 'use the one which worked last time.'"[3] The idea, then, is to adopt whichever plan was successful when this situation last arose. Cogsci literature refers to this heuristic as the 'sleeping dog' strategy. Last time, I tiptoed past the sleeping dog, and I didn't get bitten. So if I tiptoe past the sleeping dog again now, I probably won't get bitten this time either. So the plan I'll adopt is *tiptoe past the sleeping dog*. What could be more reasonable? What could be less problematic?

But, on second thought, this suggestion is no help since it depends crucially on how one individuates situations, and how one individuates situations depends on what one takes as *relevant* to deciding when situations are of the same kind. Consider: What was it, precisely, that *did* happen last time? Was it that I tiptoed past a sleeping dog? Or was it that I tiptoed past a sleeping *brown* dog? Or that I tiptoed past a sleeping pet of Farmer Jones? Or that I tiptoed past *that* sleeping pet of Farmer Jones? Or that I tiptoed past a creature that Farmer Jones had thoughtfully sedated so that I could safely tiptoe past it? It could well be that these are *all* true of what I did last time. Nor, in the general case, is there any reason to suppose that I know, or have ever known, what it is about what I did last time that accounts for my success. So, when I try to apply the sleeping dog heuristic, I'm faced with figuring out which of the true descriptions of the situation *last time* is relevant to deciding what I ought to do *this* time. Keeping that in mind is crucial. If the dog I tiptoed past last time was sedated, I've got no grounds at all for thinking that tipping my toe will get me past it now. Philosophers have gotten bitten that way from time to time.[4]

So I'm back where I started: I want to figure out what action my previous experience recommends. What I need, in order to do so, is to discern what about my previous action was relevant to its success. But relevance is a nonlocal relation, and I have only local operations at hand with which to compute it. So the 'sleeping dog' strategy doesn't *solve* my relevance problem; it only begs it. You might as well say: 'Well, you decided on an action that was successful last time; so just decide on a successful action this time too.' My stockbroker tells me he has a surefire investment heuristic: 'Buy low and sell high.' It sounds all right, but somehow it keeps not making me rich. It's well-nigh useless to propose that heuristic processing is the solution to the problem of inductive relevance because deciding how to choose, and how

3 Peter Carruthers, "Keep Taking the Modules Out," *Times Literary Supplement* 5140 (October 5, 2001): 30.

4 Another way to put the same point: What counts as the last time *this* situation arose depends on how I describe this situation. Was the last time *this* happened the last time that I tiptoed past a sleeping dog? Or was it the last time that I tiptoed past a brown sleeping dog? Or was it the last time that I tiptoed past a sedated brown sleeping dog? What determines which heuristic I should use in this situation thus depends on what kind of situation I take it to be. And what kind of situation I take it to be depends on what about my successful attempts at dog-passing was *relevant* to their succeeding. We're very close here to Nelson Goodman's famous point that, in inductive inference, how you generalize your data depends crucially on what you take them to have in common – what their relevant similarity is. The frame problem is thus an instance of a perfectly general problem about the role of relevant similarity in empirical inferences. Lacking an account of that, the advice to do the same as you did last time is, quite simply, *empty*.

to apply, a heuristic itself typically involves estimating degrees of inductive relevance.

It's remarkable, and more than a bit depressing, how regularly what is taken to be a solution of the frame problem proves to be simply one of its formulations. The rule of thumb for reading the literature is: if someone thinks that he has solved the frame problem, he does not understand it; and if someone even thinks that he understands the frame problem, he doesn't understand it. But it does seem clear that, whatever the solution of the frame problem turns out to be, it isn't going to be computationally local. Its constituent structure is all of a mental representation that a mental process can 'see.' But you can't tell from just the constituent structure of a thought what tends to (dis)confirm it. Clearly, you have to look at a lot else as well. The frame problem is how you tell *what* else you have to look at. I wish I knew. If you know, I wish you'd tell me.

The frame problem concerns the *size* of a field of cognitive commitments that one has to search in order to make a successful decision. But there are also cases where the *shape* of the field is the problem. Many systems of beliefs that are germane to estimating confirmation levels have 'global' parameters; that is, they are defined over the whole system of prior cognitive commitments, so computations that are sensitive to such parameters are nonlocal on the face of them.

Suppose I have a set of beliefs that I'm considering altering in one way or another under the pressure of experience. Clearly, I would prefer that, all else equal, the alteration I settle on is the simplest of the available ways to accommodate the recalcitrant data. The globality problem, however, is that I can't evaluate the overall simplicity of a belief system by summing the intrinsic simplicities of each of the beliefs that belong to it. There is, on the face of it, *no such thing* as the 'intrinsic' simplicity of a belief (just as there is no such thing as the intrinsic relevance of a datum). Nothing *local* about a representation – nothing about the relations between the representation and its constituent parts, for example – determines how much it would complicate my current cognitive commitments if I were to endorse it.

Notice that, unlike the problems about relevance, this sort of worry about locality holds even for *very small* systems of belief. It holds even for *punctate* systems of belief (if, indeed, there can be such things). Suppose that *all* that I believe is P, but that I am now considering also adopting either belief Q or belief R. What I therefore want to evaluate, if I'm to maximize overall simplicity, is whether the belief P&Q is simpler than the belief P&R. But I can't do that by considering P, Q, and R severally – the complexity of P&Q isn't a function of the complexity of P and the complexity of Q taken separately. So it appears that the operations whereby I compute the simplicity of P&Q can't be local.

The same goes for other parameters that anyone rational would like to maximize, all else being equal. Take, for example, the relative *conservatism* of such commitments. Nobody wants to change his mind unless he has to; and if one has to, one prefers to opt for the bare minimum of change. The trouble is, once again, that conservatism is a *global* property of belief systems. On the face of it, you can't estimate how much adding P would alter the set of commitments C by considering P and C separately; on the face of it, conservatism (unlike, for example, consistency) isn't a property that beliefs have taken severally.

In short, it appears that many of the principles that control (what philoso-

phers call) the nondemonstrative fixation of beliefs have to be sensitive to parameters of whole systems of cognitive commitments.[5] Computational applications of these principles have to be nonlocal. As a result, they can't literally be computations in the sense of that term that our current cognitive science has in mind.

If you suppose (as I'm inclined to) that nondemonstrative inference is always a species of argument to the best available explanation, this sort of consideration will be seen to apply very broadly indeed: what's the best available explanation always depends on what alternative explanations are available; and, by definition, the presence or absence of alternatives to a hypothesis isn't a local property of that hypothesis.

I should, however, enter a caveat. Suppose that something you want to measure is a property of complex beliefs but not of their parts; for example, suppose that you want to assess the simplicity of P&Q relative to that of P&R. My point has been that, *prima facie*, the computations you have to perform aren't local; they must be sensitive to properties that the belief that P&Q has *as such*. One could, however, *make* the computations local by brute force. So while the complexity of P&Q isn't determined by local properties of P together with the local properties of Q, it is determined, trivially, by local properties of the representation P&Q. In effect, you can always preserve the *locality* of computations by inflating the size of the *units* of computation. The distance between Washington and Texas is a *nonlocal* property of these states, but it's a *local* property of the Northern Hemisphere.

That sort of forced reduction of global problems to local problems is, however, cheating since it offers no clue about how to solve the local problems that the global ones are reduced to. In fact, you would think that nobody sensible would even consider it. To the contrary: recent discussions of confirmation (from, say, Duhem forward) have increasingly emphasized the *holism* of nondemonstrative inferences by claiming that, in the limiting case, *whole theories* are the proper units for their evaluation. This saves the locality of the required computations by fiat, but only at the cost of making them wildly intractable. More to the point: even if we *could* somehow take whole theories as the units in computing the confirmation of our hypotheses, the patent fact is that we don't. Though we

5 It's very striking how regularly the problems cognitive psychologists have when they try to provide an explicit account of the nondemonstrative fixation of belief exactly parallel the ones that inductive logicians have when they try to understand the (dis)confirmation of empirical theories. Pinker, among many others, objects to conceptualizing individual cognition as, in effect, scientific theorizing writ small. "Granted that several millennia of Western science have given us nonobvious truths involving circuitous connections among ideas; why should theories of a single human mind be held to the same standard?" Pinker, "So How Does the Mind Work?" In fact, however, it's increasingly apparent that the philosophy of science and the psychology of cognition are beating their heads against the same wall. It is, after all, a truism that, by and large, scientists think much the same way that we do.

The similarity between the two literatures can be quite creepy given that they seem largely unaware of one another's existence. Thus, Arthur Fine raises the question: how can someone who is not a realist about the ontological commitments of scientific theories explain the convergence of the scientific community on quite a small number of explanatory options? He says it's in part because we all follow "the instrumentally justified rule 'if it worked well last time, try it again.'" David Papineau, ed., *The Philosophy of Science, Oxford Readings in Philosophy* (New York: Oxford University Press, 1996), 78. He doesn't, however, even try to explain the consensus about what 'it' is.

don't alter our cognitive commitments one by one, it's also not true that everything we believe is up for grabs all of the time, which is what the idea that the units of confirmation are whole theories claims if it is taken literally. The long and short is: in science and elsewhere, it appears that the processes by which we evaluate nondemonstrative inferences for simplicity, coherence, conservatism, and the like are *both* sensitive to global properties of our cognitive commitments *and* tractable. What cognitive science would like to understand, but doesn't, is how on earth that could be so.

There is quite possibly something deeply wrong with the cognitive psychology that we currently have available, just as there was something deeply wrong with the associative cognitive psychology that couldn't acknowledge recursion or the constituent structure of linguistic and mental representations. What, then, are we to do?

Actually, I don't know. One possibility is to continue to try for an unvacuous heuristic account of how we might compute relevance and globality. I haven't heard of any, but it's perfectly possible that there are some out there somewhere – or that there will be tomorrow, or the day after. I don't believe it, but hope is notorious for springing eternal.

Alternatively, what our cognitive psychology needs may be a new notion of computation, one that doesn't have locality built into it. That is, of course, a lot easier to say than to provide. The current computational account of mental processes is at the core of our cognitive science. The notion of a computation is what connects our theory of mental representations to our theory of mental processes; it does for our cognitive science what the laws of association promised (but failed) to do for our empiricist forebears. As things stand, we have no idea at all how to do without it. But at least we may be starting to understand its intrinsic limitations. In the long run, that could lead to revising it, or rejecting it, or, best of all, replacing it with some theory that transcends its limitations – a consummation devoutly to be wished.

So the good news is that our notions of mental representation and mental process are much better than Hume's.

The bad news is that they aren't nearly good enough.

Steven Pinker recently wrote a book called *How the Mind Works*. It is a *long* book. In fact, it is a *very* long book. For all that, my view is that he doesn't actually know how the mind works. Nor do I. Nor does anybody else. And I suspect, such is the state of the art that, if God were to tell us how it works, none of us would understand Him.