The big idea

Can there be a science of mind?

JERRY FODOR

When I was a boy in graduate school – hardly a moment ago as geologists reckon time – the philosophy of mind was said to have two main divisions: the mind/body problem and the problem of other minds. The project was to solve these problems by doing conceptual analysis. Nobody knew what conceptual analysis was, exactly; but it seemed clear that lots that had once been considered philosophy or mind (lots of Hume and Kant, for example) didn't qualify and was not, in fact, philosophy at all. Instead, it was a misguided sort or armchair psychology, which one read by flashlight beneath the covers.

Philosophical fashions change. It's hard to believe, these days, that there is a special problem about the knowledge of other minds (as opposed to the knowledge of other anything else); and we're all materialists for much the reason that Churchill gave for being a democrat: the alternatives seem even worse. The new research project is therefore to reconcile our materialism to the psychological facts; to explain how something that is material through and through *could* have whatever properties minds actually *do* have. This seems as much a psychologist's project as a philosopher's, and might reasonably be described as the replacement of the philosophy or mind by "cognitive science". The consequent sense of temporal dislocation has been eerie. Many of the issues that are now live in the philosophy or mind (mental representation, nativism, and associationism, for three examples) are ones that Kant argued about with Hume. Whereas, topics that seemed very urgent indeed thirty years ago, now appear merely quaint. I can't remember when last I was offered a criterion.

To be a materialist is to take the view that thinking creatures are material through and through. This implies three major questions, to which a theory of mind – call it philosophy, psychology, cognitive science, or what you will – is required to find answers: 1. How could anything material be conscious? 2. How could anything material be *about* anything? 3. How could anything material be rational? These questions are now the agenda in the philosophy of mind.

I can tell you the situation in respect of the first question straight off. Nobody has the slightest idea how anything material could be conscious. Nobody even knows what it would be like to have the slightest idea about how anything material could be conscious. So much for the philosophy of consciousness.

The second question is in a slightly less parlous condition, but I won't discuss it here; it belongs more properly to semantics than to the philosophy or mind. (It's often hard to decide which of these one is working on. Not that it matters much.) Suffice it to call your attention to aboutness being a phenomenon that wants explaining, and to some implications of there being this phenomenon.

Some sorts of mental states (paradigmatically believing and desiring, rather than having an itch, or a ringing in your ears) are *about* things; so, for example, if I believe that the cat is on the mat, then what I believe is about where the cat is. Beliefs are so good at being about things that – as the philosopher Brentano famously pointed out – they can be about things that don't exist. I can believe that Father Christmas is a vegetarian, even though, or course, there isn't any Father Christmas or any matter of fact about whether he's a carnivore. Aboutness is puzzling for a materialist because it appears that – with only one kind of exception, to which we will presently return – aboutness is a thing which, in the whole material realm, only states of minds can have. Tables and chairs, for example, aren't about anything. Neither are bricks or babbling brooks, or the cat's being on the mat. But beliefs, desires, and the like do seem to be about things, and it would be nice to know how this could be so, consonant with our methodological assumption that believers are material through and through.

Because they are about things, beliefs and desires are capable of what I'll call *semantic evaluation*. Roughly, beliefs are *true or false* depending on whether what's believed actually obtains; desires are *satisfied* or *frustrated* depending on whether what's desired actually comes to pass. And so on, I suppose, for many other kinds of mental states: what semantic value they have depends on how the world turns out to be.

With that much about the second question as background, I turn now to question three: How could anything material be rational? Unlike the first question, about which nobody knows anything at all, and the second question, about which we have, at most, an occasional glimmer, somebody has actually had a deep and beautiful idea about how to answer question three. Much of the contemporary philosophy or mind, and almost all of cognitive science, is the elaboration of this deep and beautiful idea.

We commence, oddly enough, with a little Conan Doyle. Conan Doyle wasn't really much or a novelist; but he noticed a point about the mind that a lot of better writers – and a lot of philosophers and psychologists, for that matter – have failed to grasp. Here, by way of an instance, is Sherlock Holmes doing his thing at the end of "The Speckled Band".

> I instantly reconsidered my position when . . . it became clear to me that whatever danger threatened an occupant of the room couldn't come either from the window or the door. My attention was speedily drawn . . . to the ventilator, and to the bell-rope which hung down to the bed. The discovery that this was a dummy, and that the bed was clamped to the floor, instantly gave rise to the suspicion that the rope was there as a bridge for something passing through the hole, and coming to the bed. The idea of a snake instantly occurred to me, and when I coupled it with my knowledge that the Doctor was furnished with a supply of the creatures from India I felt that I was probably on the right track.

Here Watson says, "Astonishing, Holmes", or something to that effect.

Holmes offers his account as a bit of reconstructive psychology: it purports to be a capsule history or the sequence of mental events which brought him first to suspect, then to believe, that the Doctor did it with

a snake. But it bears emphasis that Holmes's story isn't supposed to be just reconstructive psychology. It also serves to assemble premises for a plausible inference. That the bed was fixed to the door, that the bell rope was in just the right place, that the doctor kept a tangle of snakes in his snuff box . . . all these are plausible reasons for Holmes to conclude that the Doctor did it. Holmes's train of thought is thus like an argument; and because it is, Holmes expects Watson to be convinced by the considerations which convinced Holmes when they occurred to him.

What connects the causal-history aspect of Holmes's story to its plausible-inference aspect is a general principle whose significance can hardly be overemphasized: the train of thought that causes one to believe that such and such provides, often enough, *reasonable grounds* for believing that such and such. Were this not the case – were there not this harmony between the semantic contents of thoughts and their causal powers – there wouldn't, after all, be much profit in thinking.

Often enough, what happens is that one starts by thinking that so-sand-so, and thinking that so-and-so causes one to think that such-and-such, and if one's thought that so-and-so was true, then so too is one's thought that such-and-such. Thinking is a process where, to quote Holmes again, "one true inference . . . suggests others". This is part of what it is for thinking to be rational, since, I suppose, rational processes should reliably take one's thoughts from truths to truths.

I'm suggesting – or rather, I'm endorsing what I take to be Conan Doyle's very acute suggestion – that, whatever else a mind may be, it is sort of rationality machine. So, now, given the methodological commitment to materialism, the question arises, how could a machine be rational? At a minimum, how could a physical system be so organized that if it starts in a state of believing something true, its causal processes will lead it to other true beliefs? Forty years or so ago, the great logician Alan Turing proposed an answer to this question. It is, I think, the most important idea about how the mind works that anybody has ever had. Sometimes I think that it is the only important idea about how the mind works that anybody has ever had.

Turing noticed that it isn't strictly true that states of minds are the only semantically evaluable material things. The other kind of material thing that is semantically evaluable is *symbols*. So, suppose I write "the cat is on the mat". On the one hand, the thing I've written is a material object in good standing; it occupies a certain amount of physical space, exhibits a certain geometrical configuration, exerts a certain (very small) gravitational attraction upon the Moon, and so forth. But, on the other hand, what I've written is about something and is therefore semantically evaluable; it's true if and only if the cat is where it says it is. So, my inscription of "the cat is on the mat" has both materiality and aboutness; as does my thought that the cat is on the mat, assuming, as we've agreed to do, that thinking is through and through material.

Having noticed this parallelism between thoughts and symbols, Turing went on to have the following perfectly stunning idea. "I'll bet", Turing (more or less) said, "that one could build a *symbol manipulating machine* whose changes of state are driven by the material properties of the symbols on which they operate (for example, by their weight, or their shape, or their electrical conductivity). And I'll bet one could so arrange things that these state changes are rational in the sense that, given a true symbol to play with, the machine will reliably convert it into other symbols that are also true."

As it turns out, Turing was right about its being possible to build such machines, "computers" as they are now called. It's a fundamental idea of current theorizing that minds too are machines of this kind, and that it is because they are that mental processes are, by and large, reliably rational.

Here's an utterly trivial example to give a quick sense of how this works. Imagine a machine that consists or two components: a "tape" (on which things can be written) and a "read/write head", about which more in a moment. We can think of the tape as indefinitely extendable (except that when a machine dies, the Fates cut its tape), and as divided into squares on which one can write, as it might be, words of English. Then here's how this machine might model the mental process in which the thought that Dinah's in the kitchen causes the thought that someone's in the kitchen. (Notice that this mental process is rational in the sense we've been discussing; it takes one from truths to truths since, if it's true that Dinah's in the kitchen, then it's also true that someone is.)

To start, one writes "Dinah is in the kitchen" on the machine's tape. The read/write head is supplied with a list of English words and is so organized that when it finds a sentence of the form " . . . Dinah . . . " on the tape, it erases the word "Dinah" and writes the word "someone" in its place. So, this machine converts the sentence "Dinah is in the kitchen" into the sentence "Someone is in the kitchen", it converts the sentence "Bill loves Dinah and his cat" into the sentence "Bill loves someone and his cat"; and so forth. Notice that, in each of these cases (and in many others), if the sentence that the machine started with is true, so too is the sentence into which the machine converts the sentence that it started with. Notice too, however, that this machine is unreliable about what follows from "Bill doesn't love Dinah"; the details of such things can be quite hard to get right.

It turns out that Turing's sort of trick can be worked with inferences that are quite a lot more complicated than the ones about Dinah. The current speculation is that perhaps *all* of the rationality of mental processes can be explained on the assumption that the mind is a symbol-driven machine of the son that Turing invented. That is, it may be that whenever a mind can proceed rationally from one true belief to another, this is because: first, it has access to symbols with which it represents what it believes; second, it is a kind of machine that converts symbols into one another in ways that are driven by physical properties of the symbols; and, third, these physically driven symbol conversions are so arranged that, on balance, they preserve semantical properties like truth. Rational mental processes are exhaustively mechanical operations on mental representations; so the speculation goes.

Whether this speculation is anywhere close to being true is, to put it mildly, an open question. Some

day we'll know, I suppose. Meanwhile, I have the following grounds for optimism to report. The pursuit of Turing's idea has led us to notice striking and pervasive features of mental processes that had not previously been remarked upon. It is always encouraging when a theory leads to unexpected insights, since that gives reason to think that maybe the theory is true. We turn to a recent example of such a result.

Turing machines work because symbols have both semantical and syntactical properties. Since the syntactical properties of a symbol are among its physical properties, there can be a symbol-transforming machine whose state transitions are driven by the syntax of the symbols it operates on. And, as we've been seeing, it is possible to arrange such a machine so that these syntactically driven state transitions preserve semantical properties of the symbols. Fine so far; but now let's look a little closer at how the relation between the syntax and the semantics of the symbols is supposed to work.

Many symbols are syntactically complex objects whose constituents are themselves symbols. Consider the sentence "John loves Mary". It's a syntactically complex object, of which the constituents are the symbols "John", "loves", and "Mary" in that order. Correspondingly, the semantic values of symbols are typically determined by complex states of affairs whose constituents are objects and their relationships. So, for example, the symbol "John loves Mary" is made true by a complex state of affairs (viz. by John's loving Mary) whose constituents are John, Mary, and the relation of loving that holds between them. This all suggests a system in which parts of sentences refer to pans of the states of affairs that determine their semantic values. "John" refers to John, "Mary" refers to Mary, "loves" refers to the relation of loving, and "John loves Mary" is made true by the state of affairs which consists of John bearing the loving relation to Mary. Splendid.

Now Turing tells us, in effect, that thinking is a kind of symbol manipulation; in effect, he says that we think in some kind of language and that thought processes boil down to mechanical operations on the symbols of that language. Let's suppose that this language that we think in shares with English the property of being what semanticists call "compositional", namely, that it has complex symbols which correspond to complex states of affairs part by part, in the sort of way we have just been discussing. Then we get the following at no extra charge: if two states of affairs are made out of the same objects and relations. then if the language contains a complex symbol corresponding to one of them, then it will also contain a complex symbol corresponding to the other.

So, for example, English has the sentence "John loves Mary" which corresponds. constituent by constituent, to a complex state of affairs made up of John, loving and Mary. But, of course, there is another state of affairs that can be made out of these same parts: viz. the one that consists of Mary's loving John. Since these two complex states of affairs are made of the same parts, and since English constructs representations of complex states of affairs from representations of their parts, we're guaranteed that if English can represent John's loving Mary, it can also represent Mary's loving John. As, of course, it can; it has the sentence "John loves Mary" to do the one, and the sentence "Mary loves John" to do the other. "So what?" you might reasonably ask. Well, if Turing is right that thinking is really the manipulation of mental symbols, and if we assume that the syntax and semantics of mental symbols are related in the ways we've just been considering, then we can make the following striking prediction: *you won't find an organism that can think that John loves Mary but that can't think that Mary loves John*. More generally, you won't find an organism that can think that the individual *a* bears the relation *R* to the individual *b*, but can't think that the individual *b* bears the relation *R* to the individual *a*, (roughly) whatever values are substituted for "a", "b" and "R".

This is, as I say, a striking prediction; one of those things that seems self-evident until you think about it, and then seems really quite surprising. Because, after all, there's no obvious reason why God couldn't have made "punctate" minds or languages, ones that have the capacity to represent *a* as bearing *R* to *b*, but not vice versa. God could have made minds and languages that way, but it appears He never exercised the option. Turing's story about the mind being a symbol-manipulation machine shows how these facts about minds and languages are connected. You don't get punctate languages because languages are compositional: representations or complex states of affairs are built up, in a systematic way, from representations of their parts. You don't get punctate minds because minds are symbol manipulating mechanisms and the symbols they manipulate constitute a compositional language.

In short, Turing's picture predicts certain symmetries in the capacity for mental and linguistic representation, and it appears that this prediction is actually true; the predicted symmetries do obtain. So. maybe Turing's story is right, then, and the mind really is some sort of Turing machine. Compare, to their disadvantage, the account of language that you get early in Wittgenstein's *Philosophical Investigations*, or the account of mind that has recently emerged from connectionist theorizing. Neither has any idea why languages and minds aren't punctate.

But again, so what? Why should we care if the mind is some sort of Turing machine? Well, there are those hard problems about consciousness, aboutness, and rationality that worry materialists. And it now appears that the third may actually be solvable. If so, then we have, for the first time, a reason to believe that a scientific theory of the mind is possible; one that adheres to the same materialist assumptions that inform the biological and physical sciences.

But why is that important? Why should it even seem surprising? Isn't it, after all, among our modern convictions that there can be a scientific theory of anything if only the taxpayers are prepared to pay what it costs to get it? I want to end by saying something about this. I put my point as a cautionary tale.

Once upon a time, there was a man who wanted to have a theory about things that happen on Tuesdays. "I will have a theory about things that happen on Tuesdays", this man said, "and it will make me rich and famous." I should be clear that he wanted not just that, for each thing that happens on Tuesdays, there should be some theory of it or other; as it might be: a

meteorological theory of Tuesday's weather, and a political theory of Tuesday's election, and a geological theory of Tuesday's earthquake. No, he had in mind something much more ambitious. He wanted a theory in which *happening on Tuesdays* explains things, in the way that *having mass* explains things in physics, and *being a mammal* explains things in biology, and *having valence two* explains things in chemistry. He wanted a theory of things that happen on Tuesdays qua things that happen on Tuesdays, as philosophers say; a theory of things that happen on Tuesdays "as such".

So, he started to construct this theory, and for a while he did fairly well. In particular, many aspects of people's behaviour seemed to depend, in an interesting way, on whether or not it was Tuesday. Thus, frequently on Tuesdays, but much less frequently on other days, people said things like "I suppose tomorrow will be Wednesday", or "It might be worse; it might still be Monday" or "Only three more days to the weekend". And so on.

But though the investigation started well, it soon petered out. Although there are some things that happen just about only on Tuesdays, and some other things that happen mostly on Tuesdays, still most things that happen on Tuesdays can just as well happen on Thursdays or Fridays and vice versa. There aren't, it turns out, any astronomical events, or geological events, or biological events, or meteorological events, or chemical events, or oceanographical events . . . , in short, there aren't even any events of most kinds that can be relied on to happen on Tuesdays and not on other days. There aren't even many kinds of behavioural events that can be relied on to do so, though, as we've seen, there are a few. The sum and substance is that, although happening on Tuesdays is a perfectly bona fide property that some events exhibit, it appears not to be a property that goes very deep into the fabric of things. Knowing whether it happened on Tuesday doesn't give you much leverage on explaining what happened, or why it happened when it did. You can't, in short, have a science of Tuesdays. Whether it's Tuesday isn't, one might say, one of the things that God cares much about.

The tale has a happy ending, however. The man who failed to get rich and famous by inventing a science of Tuesdays did get rich and famous by selling calendars. Because, although God doesn't care much whether it's Tuesdays, we do. If it's Tuesday, maybe there's a new film at the Odeon. If it's Friday, maybe there's fish for dinner. Whether it's Tuesday or Friday is humanly important even if it isn't scientifically interesting. (Meanwhile, somebody else got rich and famous by calculating the mass of protons. It turns out that, although the mass of protons is humanly boring, it is just the sort of thing that God cares about. If you want to understand how He put the world together, the mass of protons is a thing you need to know.)

The moral is that what is scientifically interesting is often, in fact, very often, humanly boring, and vice versa. The properties of things we really care about are very often not the ones that you can have deep and revealing scientific theories about. This is, I suppose, why so many people hate science. Science is forever telling us that what is most interesting to us isn't what interests God. It is disquieting to be told this; one feels unloved.

It appears, in fact, that most of the things that it is possible to have serious scientific theories about are either very small (electrons and protons and quarks and the like) or very big (stars and galaxies and the like) or very abstract and idealized (frictionless planes and the like). Whereas, most of the things that we really care about turn out to be relatively middle-sized and bristling with particularity; they are more appropriately the objects of natural history than of serious science. There are, of course, exceptions. Darwin managed to make serious science out of some middle-sized things that we care about (and thereby to irritate a lot of his contemporaries; some folks can't be satisfied). But I suppose that nobody would bother to study, for example, the Moon or the weather, except that they are *our* Moon and *our* weather, and so we are curious about them. To expect a deep scientific theory of the Moon, as such, would be like expecting a deep scientific theory of New Jersey, as such. Both are historical accidents. No doubt God cares about the Moon and New Jersey, but not, as philosophers say, under those descriptions.

My point is that it's entirely possible that our brains (/minds; I'm still assuming materialism) are historical accidents too; that they're just an example of what you get when you spend several million years piling one adventitious adaptation on top of another. If so, then our minds are another of those middle-sized things that are interesting to us (they're our minds), but quite likely inappropriate objects for serious science. There is, in particular, no obvious reason to suppose that a science of psychology, as opposed to a natural history of psychology, is possible. And, indeed, if you look at most of the attempts to make science out of the sorts of mental properties that we humanly care about (like the psychology of intelligence, or the psychology of creativity, or the psychology of mental pathology, or the psychology of affect, or the psychology of individual differences, or developmental psychology, to cite several discouraging examples), what we seem to have really is a sort of natural history. It's all humanly interesting, and some of it may even be true. But it doesn't seem to go very deep. It doesn't, by and large, suggest that you can make serious science out of the mental properties of brains. Perhaps (as, indeed, many philosophers suppose) you can only make serious science out of their neurological properties; or out of their quantum mechanical properties.

The only exception I've heard of is Turing's proposal for, in effect, making serious science out of notions like rationality and truth. Here are some properties of minds that really are humanly interesting – we want our beliefs to be true and our thoughts to be coherent – and Turing has given us some reason to suppose that these properties provide a scientific domain. It may be that, in this respect, what we care about about the mind, and what God cared about when He put the mind together, are close to being the same. The way most other things have gone, that would be a pleasant surprise and a real, if modest, consolation.