A Defense of Intentional Realism

by

Erica D. Klempner

B.A. (Hons), Psychology and Philosophy

Oxford University

Submitted to the Department of Linguistics and Philosophy
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Philosophy

at the

Massachusetts Institute of Technology

June 1997

Signature of Author..
Department of Linguistics and Philosophy
May 30, 1997

Certified by..
Alex Byme
Assistant Professor of Philosophy
Thesis Supervisor

Accepted by..
Alex Byme
Chairman, Department Committee on Graduate Students

A Defense of Intentional Realism

by

Erica D. Klempner

Submitted to the Department of Linguistics and Philosophy
on May 30, 1997 in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Philosophy

ABSTRACT

A defense of realism with regard to the propositional attitudes is undertaken. It is argued that we have propositional attitudes and that propositional attitudes cause behavior.

The positions of two opposing theorists - Jerry Fodor and Daniel Dennett - are discussed in detail. Fodor's language of thought hypothesis regarding the propositional attitudes supports realism; whereas Dennett is often taken to endorse anti-realism. It is argued that the language of thought hypothesis is false, and that a Dennettian approach to the propositional attitudes is largely correct. It is further argued that, in spite of Dennett's "anti-realist" label, a Dennettian approach can support realism with regard to the propositional attitudes.

Thesis Supervisor: Alex Byrne

Title: Assistant Professor of Philosophy

I want to defend a certain kind of intentional realism; that is, the following two theses:

(1) We have propositional attitudes.

(2) They cause behavior.

I want to do this by looking at two theories of what the propositional attitudes are.

One theory is Jerry Fodor's language of thought hypothesis (LOT). Defending theses (1) and (2) is, perhaps, recognized as being his prerogative. Endorsing LOT is a way of being as strongly realist about the propositional attitudes as it is possible to be; presumably there are other ways, but I will take LOT as the paradigm example of strong Realism.

However, I think LOT is false. I want to advocate an understanding of the propositional attitudes which denies LOT but which easily lets (1) be true, and which suggests how (2) can be true too.

The understanding of propositional attitudes I want to advance is almost entirely identical to that of Daniel Dennett; his is the second theory. However, our positions differ with respect to (2), so I call myself a realist and he doesn't (or he doesn't exactly; sometimes he calls himself a "sort of realist". I will take his theory nevertheless to be a kind of realism, weaker than LOT.) His position is subtle, so he would admit (1) but deny (2) for quite complicated, rather than straightforward, reasons (which I will go into in more detail in the last section).

Here is a summary of what follows. I will begin by giving a run-down of the theoretical field regarding theses (1) and (2). This will include brief descriptions of both LOT and Dennett's position. I will then outline the principal motivations for LOT. I wish to argue that these considerations do not necessitate LOT, and that LOT is in fact implausible for other reasons. These reasons will emerge in my discussion of Dennett's position, which follows. Dennett's work is in large part a response to LOT-like theories, so my discussion inevitably reflects this. I agree with Dennett that what it is to have a propositional attitude is to be disposed to behave in certain ways under certain conditions. I then continue with a discussion of dispositional properties in general, in order to look more closely at what exactly Fodor (or any LOT-theorist) hopes to find, and why exactly he won't find it. I then offer an explanation of why he thought he could find it. I conclude by saying why, in spite of my general endorsement of Dennett's position, I think beliefs and desires are causes of actions.

Strong Realism: LOT - (1) and (2) are both true.

Fodor states, to take the famous quote, that "having a propositional attitude is being in some computational relation to an internal representation"; that is, "[a]ttitudes to propositions are ... 'reduced' to attitudes to formulae, though the formulae are couched in a proprietary inner code." (*The Language of Thought*, p. 198.) To believe or desire that p is to stand in some (complex) computational (that is, *functionally* characterized) relation to a Mentalese "sentence" in the head that means that p. To have a Mentalese sentence in your head that means that p is to have a particular bit of your brain "realize" the Mentalese sentence that means that p, in the same way that having an English sentence in a book that means that the Russians are coming is to have a particular bit of the book realize the English sentence that means that the Russians are coming: say, a sentence that reads, "The Russians are coming." This implies that if you had a person under the appropriate sort of brain scanner and you understood enough psychology, you could point and (truthfully!) say, "That's his belief that snow is white." (Just as, if the pages of the book aren't glued shut and you understand written English, you can point to the sentence that means that the Russians are coming.)

Intentional realism does not imply LOT (if it did, I would believe Fodor and there would be an end of this matter); however, Fodor has apparently claimed that "they stand or fall together."[1]

Both (1) and (2) above do seem to point in LOT's direction, even if they don't strictly imply it. The tempting image suggested by LOT of being able to open up someone's head and find particular sentences of Mentalese yields the most unproblematic way for (1) to be true. If

[1] I have this only indirectly, from Daniel Dennett's *The Intentional Stance* (hereafter *IS*), p. 232.

there is a bit of brain which *is* the belief that p, for all beliefs that p that we have (and desires, hopes, fears, intentions that p), then of course we really do have beliefs.

Similarly for (2): it is generally held to be the intrinsic, local properties of entities or events which are causally relevant to the production of other entities or the occurrence of other events. If a propositional attitude is a particular bit of brain, then we can see, straightforwardly, how it can be causally relevant to behavior, because "being a particular bit of brain" is the right sort of property to have causal relevance. It is not quite this easy, of course: "being a sentence of Mentalese in the belief box" is supposed to be a *computational* property, not a physical property; the particular bit of brain would only *realize* the propositional attitude in question. Fodor must therefore accept all the problems to do with the thesis that functional properties - in particular, computational properties - have causal relevance. But being able to locate the particular bit of brain which realizes some propositional attitude seems the best way of guaranteeing the causal efficacy of propositional attitudes, if we can. In fact, it is not at all *easy* to see how (2) could be true unless we could locate propositional attitudes in the brain in this way. I think this is one of Fodor's prime motivations for LOT.

Eliminativism - (1) and (2) are both false.

Intentional realism and LOT standing or falling together seems also to be assumed by those who opt for eliminativism with regard to intentional states; it seems to be their dislike for LOT which motivates their decision to abandon intentional realism. That is, eliminativists deny (1) (and therefore also (2)) because they think that for (1) to be true, something like LOT would have to be true; but nothing like LOT is true; therefore, (1) is false.

The most hardline eliminativists[2] think that common (folk) belief/desire psychology was always terrible anyway - a hopeless mishmash of a theory which correctly predicts

[2] See, e.g., Paul Churchland's *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*, chapter 1.

behavior so rarely that we should be glad to be rid of it. They look forward to the day when it is fully replaced by a more predictively successful psychological theory.

More forgiving eliminativists think that belief/desire psychology was always pretty good - that, in spite of its obvious shortcomings, its predictive success is such that for everyday purposes we ought to keep it as a useful way of talking, fiction though it is. This sort of eliminativist takes an instrumentalist line. (Hardline eliminativists and instrumentalists can be seen as taking the two sides of the "glass half empty or half full?" question.)

However, although it is not unheard of for people to come to regard what they thought were obviously real entities as, in actual fact, non-existent (gods and witches and souls, for example), we probably do not want to eliminate "obviously real" entities from our ontology unless we really cannot help it. The predictive and explanatory power of intentional states, our tendency to fill up our lives with talk of them - in fact, the impossibility of eliminating them from our discourse, even if we have chosen to throw them out of our ontology (note that we do not have to speak *as if* there were gods, *as if* there were witches) - would argue for continuing to let them play with the tables and chairs and everything else that litters the philosopher's world, if we can.

Dennettian (weak) realism

Dennett's brand of intentional theory has been called instrumentalism; however, it is not the kind of instrumentalism I referred to above. He thinks that people really do have beliefs and desires, but that what really having these things amounts to cannot be understood along the lines of LOT. He compares the ontological status of propositional attitudes to the ontological status of centers of gravity (if we are doing mechanics, we certainly do have to talk as if there were these). We have sympathy both with saying that objects "really" do have them and with saying that they "really" don't. It doesn't matter what we say because we understand why we are inclined to say both. Our inclination to say objects really don't have them stems from knowing that the bit of an object which exists where its center of gravity does need not differ from any other bit of the object *except* in this: that it exists at the point which can be treateá, for the purposes of mechanics, as the only part of the object which exerts, or has exerted upon it, any gravitational force. Our inclination to say objects really do have them stems from knowing that objects really do move around like this, like the gravitational forces between them act only on points. An object *really does* have a center of gr?''ity (says Dennett) if it really does move around like this - but (he might continue) you will be disappointed if you think you can crack open an object like a nut and dig out the meat which is its center of gravity.

I hope the analogy with propositional attitudes is clear. LOT implies that beliefs and desires can be dug out of the brain like nuts from their shells; Dennett denies this, but does not thereby deny that we have beliefs and desires. (It is perhaps odd to say Fodor is committed to thinking that if you dug up a belief in this way, what you would have in the jar of formaldehyde sitting on the mantelpiece would still be a belief. But functional terms often get tied to the physical realizations of their referents. "Heart" is a functional term; and if someone digs out my heart it will no longer function as a heart, and if that person waits long enough it will no longer be capable of ever functioning as a heart again; but what "my heart" refers to *will* be able to sit in a jar on the mantelpiece.)

I have alluded to the sense in which Dennett thinks objects have centers of gravity; what is the (analogous) sense in which people have propositional attitudes? Here is one of Dennett's most explicit statements of his view on this:

> ... folk psychology can best be viewed as a sort of logical behaviorism: *what it means* to say that someone believes that *p*, is that that person is disposed to behave in certain ways under certain conditions. What ways under what conditions? The ways it would be rational to behave, given the person's other beliefs and desires. The answer looks in danger of being circular, but consider: an account of what it is for an element to have a particular valence will similarly make ineliminable reference to the valences of other elements. What one is given with valence talk is a whole system of interlocking attributions, which is saved from vacuity by yielding independently testable predictions. (IS, p. 50)

I think this is right; I think that propositional attitudes are kinds of (highly complex) behavioral dispositional states. This view is what will be developed in what follows.

Motivations for LOT

I take Fodor to be motivated by two rather separate considerations. One is the productivity and systematicity involved in the having of propositional attitudes; and the other is everyday belief/desire psychological explanation: that is, the very common practice of citing beliefs and desires as causal explanations of actions.

Regarding the first: what he has noticed about the propositional attitudes is similar to what Chomsky noticed about language use: language users have the capacity to understand an infinite number of sentences in their language. (This is the productivity of linguistic capacities.) Productivity is best explained by saying that language has a combinatorial semantics - and that we therefore represent language as a system comprised of a manageably finite number of elements (words) which can be combined in a potentially infinite number of ways.

The comparison with propositional attitudes is then just this: that we have the capacity to believe and desire (etc.) an infinite number of things, and that this capacity is best explained by saying that belief-states and desire-states have a combinatorial semantics - that they are representations in a language of thought.

However, someone oddly prejudiced against a combinatorial semantics for natural language *could* claim (with nothing more than this consideration in mind) that linguistic capacities are *not* actually productive: we are finite creatures, and can understand very *many* sentences; but (in principle, anyway) a person *could* come to the point of having understood everything he is able to - of exhausting his capacity to understand.

This seems implausible, of course. But the hypothesis that there is a combinatorial semantics for sentential representations of the propositional attitudes in the language of thought is far more controversial than the hypothesis that there is a combinatorial semantics for natural language. Fodor knows this and therefore does not wish to rely on productivity arguments.

He thus adverts to a related feature of our linguistic capacities which also calls for a combinatorial semantics but which does not rely on the idealization that our capacities are infinite. This related feature is what he calls "systematicity". Our linguistic capacities to understand sentences are systematic in that "the ability to produce/understand some of the sentences is *intrinsically* connected to the ability to produce/understand many of the others." (*Psychosemantics*, p. 149.) For example (his example), native speakers of a language who understand "John loves Mary" can also understand "Mary loves John". The idea is that there are various different kinds of words in a language, and the language system is such that the different kinds combine in a systematic way. Learning your native language thus involves learning "a perfectly general procedure for determining the meaning of a sentence from a specification of its syntactic structure together with the meanings of its lexical elements." (p. 150.) A native language which is, alternatively, learnt "phrase-book fashion" (such a language would be composed of sentences which have no constituent semantic structure) is presumably conceivable - but not a plausible model for any natural language we actually know about.

The idea now is that our capacities to believe and desire are systematic in precisely the same way: anyone, for example, who is capable of believing that John loves Mary is also capable of believing that Mary loves John (given the right circumstances). And so the argument is this:

> Linguistic capacities are systematic, and that's because sentences have constituent structure. But cognitive capacities are systematic too, and that must be because *thoughts* have constituent structure. But if thoughts have constituent structure, then LOT is true.[3]...
>
> I take it that what needs defending here is the idea that cognitive capacities are systematic, *not* the idea that the systematicity of cognitive capacities implies the combinatorial structure of thoughts. I get the second claim for free for want of an alternative account. (pp. 150-151)

[3] I am supposing that Fodor is using "thought" as a blanket term covering at least the propositional attitudes and, perhaps, more.

I will deal with this argument once my general position has been made clearer. It will turn out - in spite of what Fodor says here - that, even though our capacities to believe and desire are systematic, the truth of LOT is not in fact necessary to explain this. A Dennettian approach is entirely consistent with systematicity.

The second consideration I mentioned above as motivating Fodor's advocation of LOT - everyday psychological explanation depending on beliefs and desires being causes - I discussed briefly in this context near the beginning of this paper. I said that LOT seems the most straightforward way of guaranteeing the causal efficacy of propositional attitudes; but I did not dwell much on why we should want to do this so badly.

It does seem part of the ordinary concepts of belief and desire that they are causes of action; as Davidson pointed out, for instance, there seems no other plausible rendering of the "because" in a typical explanation of action such as, "I set out west because I wanted to strike it rich and I believed there to be gold in them thar hills," than as a causal "because". Otherwise, said Davidson, there is no way to single out one reason from several reasons an agent may have had for performing a particular action as *the* reason for which the agent acted.

In fact, beliefs and desires as causes are *so* central to Fodor's conception of the human universe that he has (famously) cried, "... if it isn't literally true that my wanting is causally responsible for my reaching, ... and my believing is causally responsible for my saying, ... if none of that is literally true, then practically everything I believe about anything is false and it's the end of the world."[4] ("Making Mind Matter More", in *A Theory of Content and Other Essays*, p. 156.)

I said at the beginning that I do want to preserve the aspect of folk psychology which suggests that beliefs and desires are causes of actions. But just as I do not think that LOT is necessary to explain systematicity, I do not think it is necessary to explain this either.

[4] Incidentally, the clause I left out after "reaching" is "... and my itching is causally responsible for my scratching... ." Itching is not a propositional attitude; it is an occurrent conscious state. Any problems there might be with these kinds of mental states being causes of behavior are very different from the problems to do with the propositional attitudes.

Dennett-conceptual versus Fodor-empirical

In the quote from Dennett I gave above, the "what it means" - in, "*what it means* to say that someone believes that *p* ..." - is italicized. This emphasizes that what he is doing is giving a conceptual and not a causal-empirical answer to the question, what are propositional attitudes? (He cites Fodor-on-Ryle as illustrating the difference between these two sorts of answer like this: ""Why are Wheaties (as the ads say) the breakfast of champions?" "Because," says the dietician, "they contain vitamins, etc." "Because," says the Rylean, "they are eaten for breakfast by a non-negligible number of champions." The former is a "causal" explanation, the latter a "conceptual" explanation..." - "A Cure for the Common Code?" in *Brainstorms*, p. 94.)

Much of *The Intentional Stance* and Dennett's other work on propositional attitudes reflects this concern with the conceptual question, a question which occupies Fodor only minimally. Fodor is primarily concerned with the causal question. However, you presumably need to know - at least *roughly* - what you are talking about before you can figure out what causes it. If you thought "breakfast" referred to what people ate watching TV, and "champions" referred to small fractious children, then you might come to the (causal-empirical) conclusion that Wheaties are the breakfast of champions because they can be eaten dry straight from the box, make a really loud crunching noise good for irritating older sisters and are crumbly enough for easy sprinkling all over the carpet.

It would of course be misleading to say that Fodor is not *at all* concerned with the conceptual question; it would be rather bizarre of him to offer an empirical theory of the implementation of propositional attitudes without any idea of what propositional attitudes might be. However, I think a much higher regard for the conceptual question to be necessary for a reasonable understanding of Fodor's project: understanding the role of propositional attitudes in psychology. I think that insufficient attention to the conceptual question has made it easier for Fodor to believe in LOT than it should be.

Given *just* the fact that Dennett is concerned primarily with the conceptual question and Fodor's LOT is an empirical hypothesis, there would be no reason why their respective answers couldn't both be right. However, Dennett thinks that his conceptual answer does make Fodor's empirical one extremely implausible.

Dennett on the conceptual question

In "Three Kinds of Intentional Psychology" (in *IS*) Dennett decides to "divorce" the aspect of the folk concept of beliefs and desires which sees them as "distinguishable, causally interacting *illata* (ie "posited theoretical entities") from the rest of the folk concept: the aspect of folk psychology which he claims is "*idealized* in that it produces its predictions and explanations by calculating in a normative system; it predicts what we will believe, desire, and do, by determining what we ought to believe, desire, and do" - the normative system at work here being the assumption of *rationality*. And it is clear that the side of the divorce which he thinks keeps most of the folk notion is not the side of the *illata*, but of the "*abstracta* - calculation-bound entities or logical constructs. Beliefs and desires of folk psychology (but not all mental events and states) are *abstracta*." (pp. 52-55)

Dennett's comparatively larger interest in the conceptual question leads him to focus much more on the external affairs relating to beliefs and desires (the connections between believers and desirers and the rest of the world - that is, believers'/desirers' behavior) than does Fodor, whose focus remains mostly internal. So a large part of Dennett's propositional attitude investigations is spent looking at propositional attitude attribution: under what sorts of circumstances do people attribute beliefs and desires to things, and how do we use these attributions? What results is a very detailed, subtle portrait of what it is to take the "intentional stance". This long, hard look is what makes him ultimately characterize beliefs and desires as abstracta.

Firstly, it is a common objection to LOT (apparent to others as well as Dennett) that we have an infinity of beliefs - or, at least, an enormous number of them - and it would therefore be impossible to store them explicitly in the brain in the way that LOT requires. It certainly seems implausible to suppose that your belief, for example, that "serendipity" is not monosyllabic,

or that the Pope has never used your toothbrush, could have been explicitly represented in your brain before you read this sentence (assuming, that is, that you are like most people in this respect).

The likewise-common response to this objection, however, is that our brains don't explicitly store all of the beliefs which can be justly attributed to us; what is explicitly stored is a small subset of our beliefs, the "core" beliefs. According to one line of thought, the core beliefs are supposed to be "causally active" in a way that the others are not.[5] My getting up to open the cupboard door, for example, is partly caused by my explicitly represented core belief that there are teabags in the cupboard; and my core belief that some swine of a roommate swiped my teabags is caused by my not being able to find them. (Core beliefs help to cause actions; perceptions cause core beliefs to be formed.) However, my vast number of virtual (as opposed to core) beliefs do not cause me to do anything, and are not (obviously) the causal result of any of my perceptions.

Dennett's reaction to this idea is expressed thus:

> "... although this might turn out to be the way our brains are organized, I suspect things will be more complicated than this: there is no reason to suppose the core *elements*, the concrete, salient, separately stored representation tokens (and there must be some such elements in any complex information processing system), will explicitly represent (or *be*) a subset of our *beliefs* at all." (*IS*, p. 56)

[5] As discussed in, e.g., Block's "The Mind as the Software of the Brain", in *An Invitation to Cognitive Science*, Vol. 3: *Thinking*.

Why is it so implausible, then, that the explicit information tokens could be representations of beliefs? After all, a theory can be unmotivated without being implausible. If a sick person has some collection of widely different symptoms, it might be just as reasonable to suppose they have a common cause - some bacterial or viral infection, say - as to suppose they do not. This is the difference between a disease and a syndrome: depression, for example, is not a disease like cholera is a disease; even clinical depression is not a disease - it is caused, even in one person, by various, complex, multiply-interacting different things, not one simple thing like a bacterium. (This is what makes clinical psychiatric problems so hard to treat.)

In essence, Fodor's faith in LOT seems analogous to a misplaced expectation that a common, easily recognizable syndrome will turn out to be a disease. Having beliefs, then, is much more plausibly like being depressed than having cholera (as it were - allowing yourself to run with the analogy, please). Belief-that-*p* "symptoms" - the behaviors which justify belief-that-*p* attribution - seem very unlikely to be the effect of as straightforward a common cause as having a Mentalese sentence meaning that *p* sitting explicitly in the appropriate functional place in the brain (sitting, that is, in the belief box).

Dennett gives many detailed examples of folk psychology use which strongly suggest this. The examples range from those which suggest LOT is unmotivated to those which suggest, further, that it is implausible. Here is one example:

> In a recent conversation with the designer of a chess-playing program I heard the following criticism of a rival program: "It thinks it should get its queen out early." This ascribes a propositional attitude to the program in a very useful and predictive way, for as the designer went on to say, one can usually count on chasing that queen around the board. But for all the many levels of explicit representation to be found in that program, nowhere is anything roughly synonymous with "I should get my queen out early" explicitly tokened. The level of analysis to which the designer's remark belongs describes features of the program that are, in an entirely innocent way, emergent properties of the computational processes that have "engineering reality". I see no reason to believe that the relation between belief-talk and psychological-process talk will be any more direct. (*Brainstorms*, p. 107)

Obviously the point about this chess program is supposed to be applicable to human believers: straightforwardly attributable beliefs *need* not be explicitly represented. But such examples from artificial intelligence also suggest that explicit belief representation, in human believers, is *unlikely*: Dennett points to the awful cumbersomeness of artificial intelligence systems which employ explicit propositional attitude representation. (E.g. *IS*, p. 229: sentential models of cognition seem to lead to "hopelessly brittle, inefficient, and unversatile monstrosities of engineering that would scarcely guide an insect through life unscathed.")[6]

The obvious LOTish response to examples of beliefs like the chess program's "I should get my queen out early" is straightforwardly to deny that they *are* beliefs - in the "core" sense. But Dennett wants to stress that such examples are overwhelmingly *typical*[7]; that it is hard to believe, of *most* uncomplicated, easily attributable beliefs, that they are explicitly represented

[6] It certainly does seem to be the case that experimental cognitive psychology conducts its business - or, at least, its most profitable business - elsewhere than in the belief/desire markets. Its business seems more to do with cognitive *sub*systems: perception, language, motor control. It is, or has been, a common complaint of psychology that it only produces theories which are obviously true or obviously false. I think the kind of psychology to which this complaint is relevant is precisely that which attempts to deal with belief/desire "laws of action". The theories which such psychology produces are then obviously true or obviously false merely depending on your point of view: for example, "pain causes evasive action" is obviously true if you take it as a rough guide to animal behavior, but obviously false if you take it as an exceptionless law.

Ferreting out connections between beliefs, desires and actions seems more the provenance of the psychoanalyst or therapist; even the novelist or playwright. (Perhaps, then, it is not surprising that Freud is discussed far more these days by literary theorists than by psychologists.)

Fodor, of course, is keen to defend the idea that what psychologists are ultimately seeking is explanations involving psychological laws, including explanations involving "laws of action". This is because, he says, "it's hard to doubt that at least *some* psychological regularities are lawlike (for example: that the Moon looks largest when it's on the horizon; that the Muller-Lyer figures are seen as differing in length; that all natural languages contain nouns.)" (*The Elm and the Expert*, p. 3) I don't wish to take a stand on this for psychological laws in general; but it is significant, I think, that none of his examples of psychological regularities are examples associating beliefs, desires and actions: the first two are visual phenomena occuring at a much lower cognitive level than belief/desire descriptions of action, and the third is, of course, a linguistic phenomenon.

[7] Other examples showing the implausibility of explicit belief representation include the "Newfie joke-getter" (*IS*, pp. 76-77), the person who has a "thing about redheads" (*IS*, pp. 148-149), and the case of Jacques, Sherlock, Tom and Boris all believing that a Frenchman has committed murder in Trafalgar Square (*IS*, pp. 54-57).

in Mentalese: "... if you were to sit down and write out a list of a thousand or so of your paradigmatic beliefs, *all* of them could turn out to be virtual, only implicitly stored or represented ... ." (*IS*, p. 56)

A problem: the explicit/virtual distinction

In any case, there is a fundamental problem with the explicit/virtual distinction - at least, as it is made out in terms of causal activeness - which I would like to address. Consider this passage of Ned Block's, from a summary of commonly given pros and cons of LOT (in the same article previously cited), which I find convenient to use as a focus:

> ... you no doubt were once told that the sun is 93 million miles away from the earth. If so, perhaps you have this fact explicitly recorded in your head, available for causal action, even though until reading this paragraph, this belief hadn't been conscious for years. Such explicit beliefs have the potential for causal interaction, and thus must be distinguished from cases of belief in the ordinary sense (if they are beliefs at all), such as the belief all normal people have that trees do not light up like fireflies. (p. 404)

Block uses "ordinary belief" here in the same way that Dennett uses "virtual belief". Later on in the section (p. 406), Block says that "a first approximation to a definition of a belief in the ordinary sense" can be given as any belief which is not explicitly represented but which "naturally and easily follows from" beliefs which are explicitly represented.[8]

This is, as he says, a first approximation, but I want to see where it leads. In the first place, it is not at all obvious to me how a belief like "trees do not light up like fireflies" could follow naturally and easily from any beliefs which a normal person can be assumed to represent explicitly. (Perhaps this follows from something like, "trees just stand in place, doing nothing"? Or "trees don't radiate light"? But these beliefs seem just as bad candidates for explicit representation as the original.)

[8] Note that I have taken one or two (hopefully) excusable liberties with this definition.

But say the normal belief that trees do not light up like fireflies does follow naturally and easily from some one or more explicit beliefs. Note that this could only ever indirectly be a criterion for the attribution of the belief, "trees do not light up like fireflies." The reason I assume most normal people have this belief is that I am sure they would answer questions and generally behave as if they did, if called upon to do so. For example, if I were to ask my mother, "Do trees light up like fireflies?" I would expect an answer of the form, "Noooo - what on earth are you talking about?" - and if she were asked to embroider her answer, I assume she would be able to tell me a lot about trees (and something about fireflies) which would explain why she thinks they don't. Similarly, you could concoct a hypothetical situation (given this particular example, extremely hypothetical) in which most normal people would (non-verbally) behave in a way that showed they had this belief. Say I am conducting a scavenger hunt, and the object that gets most points is any thing that lights up like a firefly but isn't a firefly. Trees abound - even small, easily transportable trees - but no one brings one to me.

It is, *at bottom*, always such behavioral facts which are used as criteria for belief attribution. No amount of brain data - either neurophysiological or computational - would show that someone believes "trees do not light up like fireflies" if behavioral facts showed the contrary - if, for example, someone in my scavenger hunt brought me a tree and sincerely expected to win. Such behavioral facts would show in this case that the psychological theory is *false*. (The relation here between theory and criteria for attribution is similar to that regarding physical theories for color: a physical theory for red things is wrong if it implies that most things which happen to look green to us are, in fact, "really" red.)

I said above that one belief's following from another (explicit) belief could only be an "indirect" criterion for attribution of that belief. The reason for this has to do with the way we attribute beliefs to a person because of other beliefs we have attributed to him (regardless of their being explicit or not). If I have reason to think that someone believes the sun is 93 million miles away from the earth (he tells me so), then, everything being normal, I will also think he believes the earth is 93 million miles away from the sun. I called this an "indirect" criterion because, if this person sincerely tells me that the sun is 93 million miles away from the earth and also that the earth is *not* 93 million miles away from the sun, then I won't attribute either belief to him (he doesn't seem to understand what he's talking about): the behavioral facts will always override other relevant facts when it comes to belief attribution.

This is true for beliefs across the board - regardless of any putative explicit/virtual distinction. The problem with trying to draw this distinction in terms of causal activeness can now be brought out. It is clear that it is the *potential* for causal interaction, and not actual causal interaction, which is supposed to be the basis of the distinction; otherwise, the only beliefs a person would have would be those (whether conscious or not) which are causally responsible for his *current* behavior. There would be no room for explicit but dormant beliefs - like the belief many people have that the sun is 93 million miles away from the earth.

However, in what sense does the belief, "trees do not light up like fireflies" have any *less* potential for causal interaction than the belief, "the sun is 93 million miles away from the earth"? (- for normal believers who've not read this, that is). If I ask one of these people, do trees light up like fireflies?, I will probably get a "no" of some kind; if I ask, is the sun 93 million miles away from the earth?, I will probably get a "yes" - and the causal potential of both these beliefs seems comparable no matter what kind of behavioral effect - verbalization or non-verbal action - is in question.

There is a distinction which *can* be drawn between these two kinds of beliefs. No doubt your belief that the sun is 93 million miles away from the earth was learned *by rote* - the particular words used to express this belief would have had a large role to play in your formation of the belief. (We record word strings of all kinds - whether expressions of beliefs, or poems, or irritating TV commercials.) So it is probably true that "the sun is 93 million miles away from the earth" - or some word string very like it - was explicitly recorded by you, in a way that "trees do not light up like fireflies" probably wasn't (until now).[9] But surely a defender of explicit LOT belief representation would not be satisfied with such highly verbal beliefs being the only explicit ones: any belief which causally contributes to the performance

[9] See Dennett on the difference between beliefs and opinions, e.g. "How to Change Your Mind" in *Brainstorms*.

of an action is supposed to be explicit, and we certainly act on many beliefs which we have never put into words.

Incidentally, this point guards against a possible rebuttal to my claim that the sun-belief and the firefly-belief have the same causal potential (in any sense that matters). Taking the question-answering mentioned two paragraphs ago as an example of a relevant behavioral effect, it might be said that the time taken to answer a question could be a criterion for judging of the causal potential of the related belief: it is plausible to suppose it takes much less time to come up with the sun-answer than the firefly-answer - because the firefly-answer has, somehow, to be *worked out* more. But the plausibility of this surely depends on the sun-wordstring having been explicitly, verbally recorded. I can't see how any similar criterion could, in general, be used to distinguish explicit non-verbal beliefs from virtual non-verbal beliefs.

<u>Systematicity (revisited)</u>

I am now in the position of being able to say why Dennettianism is consistent with systematicity. Recall Fodor's argument:

> ... cognitive capacities are systematic ..., and that must be because *thoughts* have constituent structure. But if thoughts have constituent structure, then LOT is true.

I want to grant that our capacities to believe and desire are systematic - which is to say, that for humans, anyway, anyone who can have a belief of the form Rxy (for example) can also have a belief of the form Ryx (when that belief would be appropriate).

Fodor notes that it is possible to try to explain systematicity by saying that it is simply the effect of a conceptual constraint on any propositional attitude attribution at all: he says, for example, that someone might claim that

> ... you can't, in principle, think the thought that (*P* or *Q*) unless you are able to think the thought that *P*. (The argument might be that the ability to infer (*P* or *Q*) from *P* is *constitutive of having* the concept of disjunction.) If this claim is right, then - to that extent - you don't need LOT to explain the systematicity of thoughts which contain the concept OR; it simply *follows from* the fact that you can think that '*P* or *Q*' that you can also think that *P*. (*Psychosemantics*, p. 152)

I think this claim is plausible (at least) for this case, and for many more. (Recall that I said that if someone sincerely tells me that the sun is 93 million miles away from the earth and also that the earth is not 93 million miles away from the sun, then I wouldn't attribute either belief to him. Understanding the symmetrical nature of the distance-relation is plausibly partly constitutive of having distance-related beliefs.)

However, I agree with Fodor that it would be very hard work to explain *all* the facts about systematicity along these lines. It seems to me, too, that "there could be creatures whose mental capacities constitute a proper subset of our own; creatures whose mental lives - viewed from our perspective - appear to contain gaps." (p. 152) That is, I can at least conceive of wanting to attribute to a creature the belief (say) that dogs hate cats, while also thinking this creature to be incapable of sufficient sensitivity to the state of affairs in which cats hate dogs to be ascribed this latter belief.

So I think systematicity must be explained some other way. To this end, I want now to focus on the idea, put to work in Fodor's argument, of propositional attitudes having constituent structure.

Presumably, what makes our believing/desiring capacities systematic is our ability to respond - systematically - to a vast number of different features of the world, which can be related in a vast number of different ways. So I think Fodor is right to suppose that the best explanation for this is that we have some kind of combinatorial representation system in our heads.

But note that Dennett *agrees* with this:

> ... even if considerations of compositionality or generativity drive us to the conclusion that the brain *has* to be organized into a modest, explicit set of core elements from which "the rest" is generated somehow as needed..., no reason at all has thereby been given to suppose that any of the core elements will be beliefs rather than some as yet unnamed and unimagined neural data structures of vastly different properties. (*IS*, p. 70)

According to Dennett, although the systematicity of our beliving/desiring capacities is (most plausibly) the effect of the compositionality of *some* kinds of representational elements, these elements need not ever be structured in a way which yields good, explicit *belief* candidates. It is unnecessary for there to be sentential objects in the head that mirror the form of the *propositions* which are the objects of (correctly attributed) propositional attitudes.

To summarize, Fodor's systematicity argument is unsound because of a false premise. The false premise is: If cognitive capacities are systematic, this *must be* because thoughts have constituent structure. This is not how it must be. This is how it would most plausibly be, but *only if* LOT were not implausible for other reasons. However, LOT is implausible for other reasons.

Beliefs and desires in a nutshell

In the last few sections, I have been agreeing with Dennett about the implausibility of most of our typical beliefs being explicitly represented. Throughout this paper I have stressed the need for a thorough understanding of the "conceptual question". Obviously I think that the answer to this question is a behavioral one: beliefs, desires (and the rest) are certain (complicated) kinds of behavioral dispositions.[10]

[10] "Behavior" is not an entirely accurate term here, I suppose. I think it perfectly possible for someone to have a belief even when the only criteria for attribution of the belief are mental phenomena, not (publicly observable) behavioral phenomena. (See footnote 2.) For example, Marge can believe her husband is having an affair even though she doesn't *behave* as if she does (she doesn't spit at him, throw vases, etc.) - she is just as sunny as ever, but her internal monologue constantly harps on the theme. It would be convenient for me if such mental things were also lumped under the "behavior" label.

To fit the picture in a nutshell, I think of desire, quite straightforwardly, as a kind of a *pull* - for X to have a desire for y is (*very* metaphorically) for the possible state of affairs in question to exert some kind of attractive force upon X; or, perhaps, for X to be on a journey towards y: X is *directed towards* y.

Desire can't be any kind of pull, of course. Every living creature is on a journey towards death, but this is generally regarded as the outcome we actually most want to avoid. So an initial observation to make is that X desiring y involves X trying to bring about y. (*Roughly* - we can of course have more or less passive desires: "I want to be able to fly to the moon." I think it is interesting, incidentally, that people seem to be the only creatures on earth capable of having desires like these.) And we only attribute "trying to bring about" to something when there *seem* to be other options - when it seems possible to go a different way. Desire is a pull towards one thing *rather than another.* (I'm going to die, whether I want to or not. It is not inevitable, however, that I choose spaghetti for lunch. I choose it because that's what I want. If spaghetti is all there is, then, if I still choose it, this is because I want to eat spaghetti rather than nothing at all. If spaghetti is forced down my throat while I am tied to a chair, then it is safe to assume that I might *not* actually want it.)

Dretske, in *Explaining Behavior*, citing Armstrong citing Ramsey (!) says that beliefs (on the other hand) are "the maps by which we steer" (p. 79). I like this idea very much. It suggests that beliefs are the things about us which guide us on our journeys towards the objects of our desires. Desires as our directedness to certain destinations, and beliefs as what guide us amongst the various routes to them, seems to me to provide an image which is strong (being easy to grasp) but also *accurate.*

Dispositions

Belief and desire are thus essentially teleological notions: they are described in terms of ends. This is characteristic of dispositional notions in general. I would like now to turn, in fact, to a general discussion of dispositions, which I hope will show in more detail exactly *what* the LOT theorist hopes to find, and exactly why he won't find it.

We are often able to discover that something with a certain dispositional property has a relatively straightforward physical property which is the ground of the disposition. To take a common example (already mentioned in a Dennett quote near the beginning of this paper): the chemical elements are disposed to combine with each other in various different ways. Their different dispositions in this respect (their different "valences") were known about, however, before it was known why they are disposed to combine as they do. It turns out, of course, that elements of the same valence are all relevantly similar in atomic structure: atoms have different "layers" of electrons (their electron shells), and only a certain number of electrons can "fit" in each shell; atoms of the same valence are missing the same number of electrons needed to make up a full outer shell. (That, at least, is how it is explained in elementary high school chemistry.) This physical property is the cause of the various elements' combining as they do. (It is in this sense that the physical property is said to ground the disposition.)

Often, when this kind of ground for a disposition can be found, a theory reduction takes place: the dispositional property is redefined in terms of the physical property which grounds it, and it is the physical property which becomes the interesting and useful one to scientists. ("Valence" now refers to an electron shell property, and not, strictly speaking, to elements' combinatorial dispositions.)

It is this kind of reduction that LOT theorists want, although the desired reduction for the propositional attitudes is of course much more complicated than standard theoretical reductions in the physical sciences. For one thing, there are more levels of reduction here: the computational, as well as the dispositional and the physical (although these distinctions are rather arbitrary: computational properties are also dispositional). Also, a person has *a lot* of propositional attitudes, and each one is a disposition to do various sorts of things in various different circumstances - *and* given the person's *other* propositional attitude dispositions.

Nevertheless, LOT requires, for all of X's (causally active) beliefs/desires that p, that there be a Mentalese sentence in an appropriate functional place in X's head which means that p, which requires that there be a particular bit of X's brain (in the appropriate place) which realizes the Mentalese sentence in question. If these requirements are met, then each of X's propositional attitude-dispositional properties can be redefined in terms of the Mentalese sentence-computational property which grounds it, which in turn can be redefined in terms of the brain-physical property which grounds *it*. This brain-physical property can then be said to be *the* (potential) cause of the actions which are criterial for X's having the propositional attitude in question. (In terms of the discussion at the beginning of this paper, theses (1) and (2) - we have propositional attitudes, and they cause behavior - will have been shown to be true in the most straightforward way.)

When the relevant "background conditions" are there, this brain-physical property will kick in and *actually* cause an appropriate action. It is admitted, of course, that our current knowledge of the brain is not sophisticated enough to allow us, at present, to understand the causal relationships between these different brain properties (one for each propositional attitude), sufficiently to develop laws about them. However, the straightforward reduction from propositional attitude-dispositional properties down through to brain-physical properties envisaged by the LOT theorist *would* enable us to develop *psychological* laws relating the things we *do* know about: namely, the propositional attitudes. The brain-physical laws can come later; but the idea is that they will be exactly isomorphic to the psychological laws. (That is: because every explicit propositional attitude can be matched up with the brain-physical state which grounds it, all of the terms for propositional attitudes in the laws of the psychological theory can be replaced by terms for their corresponding brain-physical states, and the result will still be a true theory.)

Thus Fodor: "So, typical intentional generalizations might be of the form: 'If you want to _____, and you believe that you can't _____ unless you _____then, ceteris paribus, you will perform an act that is intended to be _____.' E.g.: If you want *to make an omelette*, and you believe that you can't *make an omelette* unless you *break some eggs*, then, ceteris paribus, you will perform an action that is intended *to be an egg breaking*." (*The Elm and the Expert*, p. 4.) If X desires that p, and the background conditions are right (and background conditions include other beliefs and desires which X has), then X's desire that p kicks in to cause an appropriate p-related action.

The psychological laws the LOT theorist seeks are causal laws: (sufficient) knowledge of beliefs and desires will yield correct predictions of future actions. I want now to consider an example which will illustrate the kind of discovery that a LOT theorist wants to make (but can't). However, because the case for propositional attitudes is so complicated, I want to take a much simpler example, which I hope will serve the purpose.

Say I have a piece of crystal (a wineglass) which is both fragile and - melodious (that is, it pings nicely when you hit it). So I have something which is disposed both to smash when struck and to sing when struck; but which, on a given striking, cannot both smash and sing. If the case here is similar to how the LOT theorist hopes psychology will turn out, knowledge of this wineglass's dispositions - its fragility and its melodiousness - plus knowledge of relevant background conditions, can yield knowledge of what it will do on a given striking: smash or sing.

So, firstly, it might seem that the two dispositions ought to be capable of finer tuning; for example, that this piece of crystal (or *kind* of crystal, rather) might tend to sing when it is struck softly and to smash when it is struck harder. But suppose this isn't so; suppose, instead, that something about the crystal's structure (in virtue of which it has these dispositions, of course) makes it a very close-run thing whether, on a given striking (one which is not *too* hard, anyway), it will smash or sing. Perhaps it is a very small step from the kind of resonance which is generated in the crystal when it sings to the kind of bond break-up which occurs when it smashes; the crystal-bits are all jiggling around in both cases.

It seems, from this description, that knowledge of the wineglass's dispositions (etc.) is *not* sufficient to yield knowledge of what it will do on a given striking. It seems that, in order to gain this kind of knowledge, we would have to have appropriate *physical-level* information about the wineglass. But it doesn't seem that any physical property of the wineglass can be picked out as *the* ground of its fragility, nor as *the* ground of its melodiousness (even though we can identify such grounds in other cases). Of course, it is in virtue of its physical properties that the wineglass has the dispositions it does; but what would be required for a reduction of its dispositions to the physical level cannot be found here: a sort of precise modularity of dispositions that would allow us to isolate *this* physical property as grounding the fragility, and *that* (more or less unrelated) physical property as grounding the melodiousness.

Of course, a thorough enough knowledge of the glass's physical properties, and other relevant physical details, would allow us correctly to predict what the glass will do on a given striking. But this knowledge would not in any sense be knowledge of the glass's *dispositions* to smash and sing. No doubt we would be able to explain, using our knowledge of the physical level, why the glass has the dispositions it does; but there would be no physical grounds of its dispositional properties, in the sense explicated here.

This will be the case whenever a thing's dispositional properties cut more coarsely than the physical properties in virtue of which it has its dispositions. It should not be surprising that many dispositional properties are coarse-grained in this way: saying that something *tends* to do such-and-such under such-and-such *sorts* of conditions just is a loose way of speaking. Often - as with my crystal wineglass - knowledge of a thing's dispositions will not be sufficient to yield knowledge of what it will do on a given occasion; although sometimes, of course, knowledge of dispositions will be enough. (Say I am trying to decide which horse will win a race and all but one of them have only three legs.)

My wineglass example is, of course, entirely contrived. Quite often we are able to pick out physical properties which ground the dispositional properties of ordinary material objects. Note, however, that most dispositional properties for physical objects and materials are, as Davidson put it, "pure" disposition properties: "defined in terms of a single test." ("Actions, Reasons, and Causes", p. 15 of *Essays on Actions & Events*.) Fragility, solubility and dormitivity (to name two other common examples) are defined respectively in terms of breaking, dissolving and sleep. It is perhaps not surprising, then, that physical grounds for such pure dispositions can be found in particular cases, incorporating them into the broader framework of physical theory (and law).[11]

It has long been known, however, that belief/desire psychology is particularly inapt for being incorporated into law-talk. ("Ceteris paribus" seems to be one of Fodor's favorite phrases.) Given that beliefs and desires are probably the most *im*pure dispositional properties we ever refer to, perhaps this should not be surprising either. Possessors of beliefs and desires are disposed to do all kinds of different things, such a variety of different things that "having the desire that p" certainly cannot be *defined* (as fragility, solubility and dormitivity can be defined) at all. Davidson continues (from above) as follows: "... desires cannot be defined in terms of the actions they may rationalize, even though the relation between desire and action is not simply empirical; there are other, equally essential criteria for desires - their expression in feelings and in actions that they do not rationalize, for example."[12]

Having said this, it should be noted that it certainly is not impossible to conceive of a thing for which belief/desire talk seems apt (some "intentional system") but for which a dispositional-level - physical-level reduction *can* be done. The thermostat Dennett describes in "True Believers" (in *IS*) seems just such a candidate.[13] But the thermostat is only capable of wanting and believing a small set number of things: it wants the room to be a certain

[11] I suppose valence (my earlier example) is not a pure disposition; however, the combinatorial dispositions of the elements are so precisely regular that finding a physical property to ground them should not have been surprising here either.

[12] There are, of course, other dispositional terms (besides propositional attitude terms) that resist definition: character traits, for example - being witty or charming, stupid or grumpy. This is because - like propositional attitude terms - they refer to dispositions with a huge variety of different effects. (I tried to think of some examples of such impure dispositional terms commonly used not just for people, but also for things; I could think of some - "funny", "scary", "irritating" - but I could not think of any which do not at least relate to some kind of human, or perhaps animal, activity. I think this is interesting.)

Incidentally, I will turn to the notion of rationalization, mentioned here, in the next section.

[13] I don't really care whether such a limited intentional system as this thermostat should be dubbed a genuine believer and desirer. Perhaps only more complicated intentional systems should be granted these titles. But I entirely agree with Dennett that the differences between the thermostat and the average person, with respect to believing and desiring, are merely differences of degree.

temperature, it can believe the boiler to be on or off, etc. Its belief/desire dispositions seem as precisely regular, as well-defined, as the combinatorial dispositions of the chemical elements.

People (and most animals - certainly all mammals) are different because we perceive and react to *so much more*. Our propositional attitude dispositions are enormous in number and interlock very subtly, but not in a precisely regular, well-defined way.

The relatively coarse grain of human belief and desire becomes apparent when you consider two phenomena of propositional attitude attribution in particular (both of which Dennett frequently discusses): (1) the (often unsuitable) precision of language, as compared to the propositional attitudes it is used to express; and (2) the hiccups in the smooth process of propositional attitude attribution occasioned by irrationalities and cognitive mistakes.

The first point will not be long in the making. It is just that, as Dennett says, "Language *enables* us to formulate highly specific desires, but it also *forces* us on occasion to commit ourselves to desires altogether more stringent in their conditions of satisfaction than anything we would otherwise have any reason to endeavor to satisfy. ... "I'd like some baked beans, please." "Yes sir. How many?"" As for belief, "... our linguistic environment is forever forcing us to give - or concede - precise verbal expression to convictions that lack the hard edges verbalization endows them with." (*IS*, pp. 20-21.)

As for the second point[14], let's take an ordinary example of self-deception. Joe is a smoker who seems to go out of his way to find alternative explanations for evidence which shows (to any reasonable person) that smoking is unhealthy: for example, he claims that it is not smoking itself which is unhealthy; rather, people who are genetically disposed to smoke are also genetically disposed to get lung cancer, heart disease and emphysema.[15] The way Joe does this makes his friends describe him as deceiving himself, because he exhibits the classic signs of self-deception: he seems both to believe and not to believe that smoking is unhealthy. He seems not to believe this because he does things like hotly deny it, and light up whenever he

[14] Dennett discusses the second point in far more detail in "Making Sense of Ourselves", in *IS*.

[15] Vann McGee has told me that this claim was actually made in pamphlets distributed by the tobacco lobby in the 1950's.

likes; he seems to believe it because he *avoids* the fact with such consummate skill - only someone aware of the fact could avoid it so well.

Examples of irrationality can be absolutely bewildering to anyone who believes in something like LOT. If a particular sentence is either in the belief box or not (in the desire box or not), then such examples tend to provoke ad hoc theory fix-ups which are extremely implausible. It is clear why Freud, to account for self-deception within a framework much like LOT, effectively gave the mind two belief boxes, one in its unconscious part and one in its conscious part. It is also clear why people since Plato have been denying even the *possibility* of irrational action.

However, cases like this are easy to explain within the propositional attitudes-as-dispositions framework I have been endorsing. Dispositions of any kind are straightforwardly attributable only given a certain consistency and stability of behavior on the part of the object in question. A certain material is not straightforwardly either fragile or strong if, when hit with a sledgehammer, sometimes it breaks and sometimes it breaks the hammer - even though we would understand, in this case, why we might be inclined to attribute both properties to it (or neither). It is not surprising that neither smoking-belief is straightforwardly attributable to Joe, as he shows a similar inconsistency of behavior.[16]

The kind of material I mentioned is hard to find; irrationality of human behavior may not be so entirely shy, but it is certainly rare in comparison to rational behavior. This makes sense, of course, seeing as we are evolved creatures, apparently living in a world where it does us good, for example, to react in much the same way to every set of sharp, gnashing teeth we meet.

[16] We might want to say that for these cases and others like them, whether or not a thing has a certain disposition is *indeterminate*.

Why think there are belief/desire laws of action?

The crystal wineglass I discussed was an example for which knowledge of dispositions was not sufficient to yield knowledge of what the thing would do in a particular circumstance. I did say, however, that sometimes knowledge of dispositions is enough to yield knowledge of what something will do on a given occasion. This is also true, of course, for belief/desire psychology. Time after time we successfully use the intentional stance to predict behavior: "I give you two choices: either you eat an ice cream sundae, or I immerse you in a vat of writhing leeches." Time after time, however, it is just not good enough: "So which would you like, chocolate or vanilla?"

Yet it has seemed (it certainly still does seem, to the LOT theorist) that belief/desire psychology can be developed into a theory consisting of true causal laws of action involving beliefs and desires. In this section, I would like to offer a reason for why this idea can be so compelling.

We use belief/desire talk in two significantly different ways - different ways which have apparently been the cause of some philosophical puzzlement. We use beliefs and desires to *rationalize* actions which have already been performed; we also use beliefs and desires to *predict* actions which have yet to be performed. The puzzle came from trying to square explanations in terms of reasons with explanations in terms of causes. In rationalization, the belief/desire reason which makes sense of an action seems "logically connected" to the action in a way which seems unsuitable for causal explanations.

Why "logical"? Say we have identified some action A *as* an action, and set out to explain it by giving the agent's reason. We know that, as A is an action, we will be able to attribute to the agent *some* reason for it. We also know that an agent usually (*always*?) has several different reasons for doing several mutually exclusive things; but that, because A is what the agent *did* in fact do, his reason for doing A must have been his strongest reason. (The desire which he aimed to satisfy by A'ing was his strongest - in the sense of "most motivating" -

desire. Similarly, an agent can often believe there to be several different ways of satisfying a desire; given the way he *did* do it, he must have thought that way the *best* way.)

I cut the red wire - an action of mine. I cut the red wire, say, because I wanted to defuse the bomb and thought cutting the red wire would do this. I probably also wanted to do other things - say, run away. But I did cut the red wire, so wanting to defuse the bomb must have been my strongest desire. Perhaps I thought there were other ways to defuse the bomb (maybe by covering it in molten lead), but what I did was *cut the red wire*, so I must .nave thought this the best way of defusing the bomb.

The "logical" connections involved here are suggested by all the "must have"s in the previous two paragraphs. Once we have identified a bodily movement as an action, and identified the agent's belief/desire reason for it, it falls out *by definition* that this reason was the agent's strongest reason. (It was an action, so it had to have a reason; *whatever* reason there was for it had to have been the strongest reason.)

However, this logical connection between reasons and action can suggest the following principle, which looks very much like a predictively useful, causal generalization: X will try to do what he believes to be the best thing which satisfies his strongest desire.[17] It seems as if we can determine X's strongest reason *before* he acts and use our knowledge to predict X's behavior, in the same kind of way that, if we determine the mass of an ob ect and all of the forces acting upon it, we can predict which way and how fast it will go.

However, this analogy is attractive but misleading. The difference between the two cases indicates, in fact, what is behind the (once?) common thought that reasons can't be causes of actions because they are not "logically distinct" from them, and causes must be logically distinct from their effects. It is possible to determine what forces are acting upon an object before seeing how it moves; force, mass and acceleration can be measured independently of

[17] This formulation merely gets rid of the need for the "ceteris paribus" in Fodor's general "law of action" schema. ("If you want to _____, and you believe that _____, then, ceteris paribus, you will try to _____" becomes, "If you want to _____ more strongly than anything else, and you believe that by _____ you can best satisfy this desire, then you will try to _____.")

each other, roughly suggesting (I hope) the sense in which they *are* independent of each other. (This is why "F=ma" is a physical law and not merely a tautology.)

Which is a person's strongest reason, on the other hand, is determined *entirely* by what he ends up doing. Say I was a novice bomb-defuser, and my boss was watching me tackle the bomb. Knowing I was new at the job, he might have thought it possible that I would crack under the strain and run away, but likelier, given my general strength of character, that I would stay and successfully cut the red wire. That is, he might have thought my strongest desire would be to defuse the bomb, by cutting the red wire. However, say that what I actually do, in spite of his flattering assessment, *is* run away. He would have been wrong about my strongest desire, which, in the event, *turned out to be* to run away.

Certainly, it is also possible to make a mistake about the forces acting upon an object: to be surprised by the object's movement and to infer that there must therefore have been some other (unknown) force at work; but the existence of such a force would have been ascertainable from the outset, and not just *as a result* of the object's movement.

A LOT theorist would presumably want to say that it is the same kind of mistake that my boss makes in trying to predict what I will do: he simply does not know enough about my various competing desires and dispositions to come to the right answer about what my strongest desire is. But I have been claiming, of course, that this is just not right; that in some cases we don't know what an agent's strongest desire is, and hence cannot know what he will do (not, at least, on the basis of belief/desire prediction), not because of our epistemological shortcomings, but because there simply aren't the relevant *belief/desire* facts there to be known. (*Some* relevant facts are presumably there: in a deterministic universe where I have Laplacean omniscience, I can know what the agent will do; just as, in such a universe, I can know what the winning lottery ticket will be before it is drawn from the barrel. But, before it is drawn, there is no lottery-fact which can tell me this; the lottery-facts have it, in fact, that each ticket has the same chance of winning.)

All of the foregoing discussion against beliefs and desires as illata does not imply that the principle I stated above - "(If beliefs and desires can be attributed to X), he will try to do what he believes to be the best thing which satisfies his strongest desire" - is false. In fact, I think it is true. But it does mean that it is not the kind of predictively useful, causal generalization that it looks like it is. It is a principle related to an analytic principle regarding dispositions in general: If a disposition to *k* (under circumstances *l*) can be attributed to X, then under circumstances *l*, X will *k*.

Beliefs and desires as causes (or: (1) and (2) are both true - reprise)

So in what sense is it that beliefs and desires are causes of actions? I will try to sketch an answer.[18]

I will begin this final section by looking at why Dennett thinks beliefs and desires are not causes of actions in any sense. (It is his denial of thesis (2) - that propositional attitudes cause behavior - which gives him his slight "anti-realist" flair, as he accepts thesis (1) - that we do in fact have propositional attitudes.) I think that his reason for this might have to do with his Quine-style endorsement of a kind of "indeterminacy of propositional attitude attribution", which seems to stem from the strong emphasis he is inclined to give to the importance of assuming an *idealized* rationality when taking the intentional stance.

Dennett makes this sort of comment to deny his realism:

> ... I am a sort of realist, but I am not Fodor's Realist with a capital *R* since I expect that the actual internal states that cause behavior will not be functionally individuated, even to an approximation, the way belief/desire psychology carves things up. (*IS*, p. 71)

This is said after quoting Fodor thus:

> I propose to say that someone is a *Realist* about propositional attitudes iff (a) he holds that there are mental states whose occurrences and interactions cause behaviour and do so, moreover, in ways that respect (at least to an approximation) the generalizations of common-sense belief/desire psychology; and (b) he holds that these same causally efficacious mental states are also

[18] I feel it only fair to emphasize at this point that what follows is *just* a sketch.

> semantically evaluable.("Fodor's Guide to Mental Representation", p. 78, *Mind* 1985)

Dennett continues some pages later:

> The fact about competence models [an intentional system model being one of these] that provokes my "instrumentalism" is that the decomposition of one's competence model into parts, phases, states, steps, or whatever *need* shed no light at all on the decomposition of actual mechanical parts, phases, states, or steps of the system being modeled... (*IS*, p. 76-77)

I agree with the sentiment expressed in both of these passages, and yet still wish to call myself a realist (although I suppose I am also not Fodor's Realist with a capital *R*). Dennett acknowledges this possibility for himself:

> Since the general power of the intentional stance is ... not explained by any knowledge we might happen to have about mechanisms in the objects we thereby comprehend, I continue to resist the brand of realism that concludes from the stance's everyday success that there must be belief-like and desire-like states in all such objects. Bechtel ... has suggested, constructively, that this still does not bar me from a variety of realism. I could reconstrue my instrumentalism as realism about certain abstract relational properties: properties that relate an organism (or artifact) to its environment in certain indirect ways. (*IS*, p. 81)

However, he goes on to say:

> So far as I can see, this is indeed a viable ontological option for me, but as chapter 5 [of *The Intentional Stance*] will show, the properties one would end up being a realist about are hardly worth the effort.

I have read chapter 5 ("Beyond Belief") several times; but I must confess to being unable to see exactly what he means here. Perhaps the word "abstract" is a pointer; I certainly see beliefs/desires as relational properties, but why "abstract"? Why aren't they fairly concrete relational properties of a kind? - that is, fairly concrete, but extremely numerous and complexly interlocking, dispositional properties of the kind I have described.

Dennett often stresses that when taking the intentional stance, it is necessary to idealize: to work within the normative system of ideal rationality. I think this is right, but that this way of looking at the matter (perhaps just this way of *phrasing* it) can easily help play up the instrumentalism (as opposed to the realism) of his position; it can seem that the entities of the system - beliefs and desires - are merely possessed by some ideally rational, theoretical organism, and not by the real organism which the theoretical one is designed to model. I don't mean to suggest that this is what Dennett thinks; on the contrary, I think it is clear that he takes real organisms to possess beliefs and desires. But his discussions of rationality often suggest to me that he might think something similar: that what makes it the case for an organism to possess beliefs and desires, is for it to have a particular kind of relationship to the ideally rational, theoretical organism which models it most closely (or to any of them, when there is more than one). This picture certainly shows why the term "abstract" above would be appropriate, and certainly suggests, I think, why Dennett would be wary of the claim that beliefs and desires cause actions.

I described in an earlier section (see pp. 29-31) how examples of irrationality can be easily explained within the dispositional framework I endorse. I hope it is clear from this that rationality can also be easily explained within this framework: there must (as Dennett says) be a large background of rationality for any belief/desire attributions to be made at all, but this is simply a special instance of what must be the case, in general, for dispositional attributions to be made. Once this is acknowledged, the idea that in order to attribute beliefs and desires, we must work within an idealized, normative system, loses its instrumentalist flavor. There is no stronger sense of idealization or normativity required for folk psychology than there is for attributions of fragility, or solubility, or dormitivity. (Yet, of course, the idea that in order to make attributions of these kinds of dispositional properties, we must work within an idealized, normative system, certainly does not get stressed.)

I therefore think that all of the problems to do with beliefs and desires being causes of actions are the *same* problems as those to do with dispositional properties, in general, being causally efficacious.

I realize that there are problems here. I do not intend to solve them, but I would like at least to indicate my inclinations in this matter.

I will again use an article of Ned Block's as an example. I think that the problems he discusses in it can be taken as paradigmatic. In "Can the mind change the world?" (in *Meaning and Method: Essays in Honor of Hilary Putnam*, pp. 162-163) he points out that dispositional properties can be construed in two ways. To take his example: the attribution of dormitivity to a pill (say) can either be construed as the attribution of a second-order property - "having some property or other that causes sleep" - or as a first-order property - "the property that causes sleep". ("'Dormitivity', on the former construal, always picks out the same second-order property. But on the latter construal, it picks out different first-order properties in the case of different types of sleeping potions.") Block argues that dispositional properties of the special sciences must be construed the former way; otherwise (for example) the psychological property of "believing that grass is green" would in fact pick out a physiological property and not a psychological property. Given my discussion in this paper, it is clear that I also think it necessary to construe belief/desire predicates in this way. (However, it would be misleading to say that "having the belief that p" involves "having some property or other that causes such-and-such behavior" - as if there is *one* such property. LOT assumes there is, but I don't. A less misleading, but less natural, way to put it would be: "having the belief that p" involves "being such-and-such behavior-causing". The parallel here with "dormitivity" would be: "is dormitive" = "is sleep-causing".)

Block makes a further distinction between *causal-explanatory relevance* and *causal relevance of properties*, and argues that if dispositional properties are construed the second-order way - as, in the special sciences, they must be - then they have causal-explanatory relevance, but are not causally relevant. Here is how he explains the difference:

> ... notice that we can causally explain a case of sleep via appeal to the second-order property of dormitivity. I fell asleep because I took a pill that had the property of having some property or other that causes sleep. This is causal-explanatory because it rules out alterr ıtive causal explanations of my falling asleep - for example, that I was grading papers. But ... the appeal to dormitivity

> is not an appeal to a causally relevant property. Dormitivity is causal-explanatorily relevant to sleep without being causally relevant to sleep. Why? The appeal to dormitivity *involves* an appeal to a property that is genuinely causally relevant to the production of sleep, namely, the unnamed property (presumably a first-order chemical property) whose existence is mentioned in the analysis of the second-order property, the one that the dormitivity of the pill consists in the possession of. Dormitivity has causal-explanatory relevance because it involves a causally efficacious property. ... But the causally relevant property is the first-order one, not the second-order one... .

I will not go into Block's reasons for denying that dormitivity has causal relevance. I just want to question whether causal relevance, as he has elucidated it, is something we need care about in our search of a defense of the thesis that beliefs and desires cause actions.

Say I take a sleeping pill. The dormitivity of the pill, construed as a second-order property, is causal-explanatorily relevant to my sleep. What this entails is that, if the pill had not been dormitive - if, that is, it had not been sleep-causing - then I would not have slept (unless something else had caused me to do so). Similarly: say I go to the store because I want some milk. My desire for milk is causal-explanatorily relevant to my going to the store: if I had not wanted milk, I would not have gone (unless something else - some other reason, or some non-rational cause - caused me to do so).

Block does say earlier in the paper, "... if you want to avoid epiphenomenalism, go for a counterfactual theory of causal relevance, not a nomological theory." (p. 159) I do not know enough about theories of causality to say anything more about them; but it does seem to me that a counterfactual approach accounts for every reason we could have for claiming that beliefs and desires do cause actions.

Remember Davidson's reason for claiming that beliefs and desires cause actions: we cannot account for ordinary folk psychological explanations of the form, "X did y because he had belief/desire reason z" unless we construe the "because" as a causal "because"; otherwise, if X also has another reason to do y, it would be impossible to pick out reason z as *the* reason for which he acted.

I said (see p. 12) that folk psychology does seem committed to this; folk psychology does involve making distinctions of this kind. But causal-explanatory relevance is all we need to explain the force of the causal "because". Say I wanted bread as well as milk (and actually bought both when I got to the store), but that my desire for bread was not the reason (or even part of the reason) for my going. Folk psychology has it, then, that I had two reasons for going to the store, but only one of these - the milk-reason - was the reason upon which I acted. Only the milk-reason caused me to go (folk psychology might say). So, counterfactually: if I had not wanted milk, I would not have gone; but if I had not wanted bread, I still would have.

For this example, it seems that both of these counterfactuals would be true - and their truth seems to be all that the causal commitments of folk psychology would require. I can't think of an example for which folk psychology would require anything more. I am not inclined to require anything more myself.

References

Block, Ned (1990). "Can the mind change the world?", in George Boolos (ed.), *Meaning and Method: Essays in Honor of Hilary Putnam*. Cambridge University Press.

Block, Ned (1995). "The Mind as the Software of the Brain", in Edward E. Smith and Daniel N. Osherson (eds.), *An Invitation to Cognitive Science* (Second Edition): *Thinking* (Volume 3). The MIT Press, Cambridge, Massachusetts.

Churchland, Paul M. (1989). *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. The MIT Press, Cambridge, Massachusetts.

Davidson, Donald (1963). "Actions, Reasons, and Causes", reprinted in *Essays on Actions & Events*, Oxford University Press, 1980.

Dennett, Daniel C. (1978). *Brainstorms: Philosophical Essays on Mind and Psychology*. Bradford Books, Montgomery, Vermont. (First MIT Press edition 1981. Cambridge, Massachusetts.)

Dennett, Daniel C. (1987). *The Intentional Stance*. The MIT Press, Cambridge, Massachusetts.

Dretske, Fred (1988). *Explaining Behavior: Reasons in a World of Causes*. The MIT Press, Cambridge, Massachusetts.

Fodor, Jerry A. (1975). *The Language of Thought*. Thomas Y. Crowell Company/The Language & Thought Series, New York.

Fodor, Jerry A. (1985). "Fodor's Guide to Mental Representation", *Mind*, XCIV, pp. 76-100.

Fodor, Jerry A. (1987). *Psychosemantics*. The MIT Press, Cambridge, Massachusetts.

Fodor, Jerry A. (1990). "Making Mind Matter More", *A Theory of Content and Other Essays*. The MIT Press, Cambridge, Massachusetts.

Fodor, Jerry A. (1994). *The Elm and the Expert: Mentalese and Its Semantics*. The MIT Press, Cambridge, Massachusetts.