FRED DRETSKE

HOW BELIEFS EXPLAIN: REPLY TO BAKER

(Received 10 October, 1990)

As Lynne Baker says, I *do* think that beliefs help explain behavior in virtue of what they are beliefs about — in virtue of their content or meaning. I also think this sort of explanation is causal. When Jim, thirsty for a beer, goes to the fridge because he thinks there's a beer there, his trip to the fridge has, as (part of) its structuring cause, the fact that he has a belief with this content. Finally, I confess to believing that, in the first instance at least (things get messier when we are dealing with systems already having complex representational capacities), content or meaning derives from a learning process in which a structure acquires an indicator function (becoming, thereby a representation) by being recruited to do a certain causal job. It is here, or so Baker charges, that I run smack into my own behind. For the way something acquires an indicator function (a meaning) is by being recruited (in virtue of what it indicates) to play a certain causal role, exactly the thing the meaning is supposed to explain. So why isn't this circular?

It isn't circular because the causal roles meanings are supposed to explain aren't the causal roles from which meanings are derived. There is, in other words, and despite Baker's precautionary "without equivocation" and her later claims that these are the *same* causal roles, an equivocation in her three theses.

It is important, first, to understand that, contrary to what Baker alleges, a structuring cause *does* explain *token* behaviors. To use an example from the book, the thermostat in a home heating system is supposed to turn the furnace off and on in response to changes in room temperature. An electrician makes a mistake and wires my thermostat to the garage door opener. It then opens the garage door every time it gets cold in the room. A puzzled visitor wants to know why my thermostat behaves in this odd way? Why does it open the

Philosophical Studies **63**: 113–117, 1991.

garage door instead of turning on the furnace as other well-behaved thermostats do? In asking for an explanation of my thermostat's unusual behavior, my visitor is looking for a structuring cause (since he knows it is my thermostat that is doing this, we may suppose that my visitor knows the triggering cause: changes in room temperature). The structuring cause is the electrician's actions *yesterday* when he wired the thermostat to the garage door opener. These activities yesterday causally explain why the thermostat opens the garage door today, why it will (if we don't fix things) open it tomorrow, why it will open it *each* time it gets cold in the room. Structuring causes explain *each* and *every* tokening of the process they structure. They do so because they are the structuring cause for each such token. This is *not* to say, of course, that the structuring cause explains why the thermostat opened the door this morning, say, rather than this afternoon. *That* is the work of a triggering cause (or a more elaborate structuring cause). Though the structuring cause cannot explain why the thermostat is opening the garage door *now* (instead of later), it can explain why it is now opening the garage door (rather than turning on the furnace).

This is an important point because the fact that beliefs have a certain content is supposed to be a structuring cause of the behavior they (help) explain. Since (on my account of the matter) a belief's content derives from a *past* learning process, it is important to understand the way these past events (on which the meaning of a structure depends) can continue to function as causes, *structuring* causes, of present, ongoing, behavior. Baker suggests that questions about why someone did something — why, for example, Booth shot Lincoln — is (always?) a question concerning triggering causes. I deny that. The puzzled visitor described above, in being told about the electrician's mistake yesterday, is being given a perfectly correct explanation (a structuring cause) of why my thermostat is opening the garage door. The fact that it will *still* be the correct (structuring) explanation of the thermostat's behavior *tomorrow* (when, in response to a drop in temperature, it opens the garage door *again*) does not mean its not the correct explanation today or that structuring causes really explain something other than token behaviors.

One further preliminary point before confronting the charge of circularity. Certain forms of learning, I contend, confer on internal structures an indicator function, a function that makes the structure a representation of that condition it is supposed to indicate. Structures acquire this function by being recruited to do certain causal jobs (because of what they indicate), but, it is *not* their function to cause what they are, during learning, recruited to cause. They acquire an indicator function by being recruited to cause *something*, some bodily movement or other, but there is no bodily movement it is their function to cause. They acquire an *information-carrying* function, the job of indicating that (say) condition *F* exists, and they acquire this function (during learning) *by* causing (say) *M*, but it is not their job to cause *M*. They may, in fact, never again cause a movement of type *M*. What these structures (help) cause is, as it always is, a function of the total cognitive and motivational state of the organism, and that total state (desires and *other* beliefs) may never again be the same as it was during learning. Jerry Fodor has charged me with making this sort of mistake, the mistake of thinking that beliefs have the function, not of being true (or, as I prefer to put it, of indicating) but of causing certain behaviors (see, e.g., "Reply to Dretske's "Does Meaning Matter?"' in *Information, Semantics & Epistemology*, Enrique Villaneueva (ed.), Blackwell, 1990). I have made quite a few mistakes in my time (some of which Fodor has caught), but this is not one of them.

What this means, of course, is that *current behavior*, the causal process that the meaning of *C* is called upon to explain (as structuring cause) need not (and typically will not) be the same sort of causal process as that which was responsible (during learning) for *C*'s acquiring that meaning. *C* got the function of indicating *F* (hence, this meaning) by being recruited to cause *M*, but what its having this meaning is (typically) called on to explain is its causing *N*, quite a different movement. And even if it *is* called on to explain the production of *M* (the same type of movement that it was recruited during learning to cause), it wasn't its causing *M* that conferred an indicator function on *C*. It was its causing *something*, some movement *or other* (whatever movements were rewarded in the conditions *C* indicates). So the causal process (behavior) being explained by meaning is *never* the causal process underlying the meaning that explains it.

With these points in mind, and leaving aside the role that motivational factors (i.e., desires) play in the overall account, consider some token behavior — i.e., a particular tokening of the $C \rightarrow M$ process. Since it is important for the point at issue that we make clear distinctions between types and tokens (my thanks to Jaegwon Kim for convincing me of the need for this), let C be a structure type of which C_1, C_2, etc. are tokens. A current token of C, call it C_n ("*n*" for now), is causing a particular movement (M_n), the resulting process being a current piece of behavior — S moving his arm, say. Assume that C_n represents condition F. C_n may or may not *indicate* F; whether it does or not will depend (among other things) on whether condition F actually obtains. If condition F does not exist, then C_n *misrepresents* F to be the case. This, I claim, corresponds to a false belief that F.

With this stage-setting complete, the claim is that C_n's meaning F explains (or helps explain) why it causes M_n (why the person is moving her arm). On my account of meaning, the fact that C_n means F is a fact about earlier tokens of C, about C_l, C_2, etc.. It is, if you will, a historical fact about earlier tokens of type C, a fact about how they were (progressively) enlisted for control duties because of what they indicated about F. In the simple sort of learning conditions I describe, there must have been some sort of movement they were recruited to produce, a type of movement that was rewarded (let us say) when, but only when, it occurred in conditions F.

So, to say that C_n means F is to say that earlier tokens of this type indicated F and that this fact (that they indicated F) was responsible for a causal re-organization, a re-shaping, of control circuits so that later tokens of this type (including, of course, the current token, C_n) would have causal duties (in the determination of motor output) they did not previously have. To attribute meaning to a token internal state is, on this account of meaning (and belief), to describe the *source* of its causal efficacy. It is to say what gave it a voice in the determination of output, what led to its installation as a control structure. To say that C_n means F is to say from whence C_n got its causal powers — powers it is (presently) exhibiting by helping to produce M_n. It is to say that C_n acquired causal efficacy (the sort of efficacy relevant to shaping output) from earlier tokens of this type *indicating* (carrying the information) F. This does not make the appeal to C_n's meaning (in the explanation of why it is causing M_n) circular. For the causal job being explained is what C_n is doing (causing M_n — current behavior), and the property of C_n that is invoked to explain it, C_n's indicator function, is a property constituted by what earlier tokens of this type did.

The key to this analysis is the way two pieces interlock. These pieces fit together in a particularly intimate way. There is, first, the fact that behavior is a process wherein C_n, a token of some physical type C, causes some particular token movement. The fact that behavior is a process of this sort makes it susceptible to causal explanations of *both* triggering and structuring types. Second, there is the fact that meaning is an extrinsic, in particular an *historical*, property. Appealing to the meaning of C_n in explaining behavior (why it causes M_n) is merely a way of describing *which* indicational properties (of earlier tokens) were responsible for executive re-organization. The first fact (that behavior is a *process* having structural causes) makes behavior the sort of thing that can be explained by the sort of thing the second fact tells us meanings are (namely, historical properties, properties that do *not* supervene on the current state of the system).

The fact that meaning is an historical property reveals why meanings retain their causal efficacy even when they are false, why false beliefs are as effective as true beliefs in explaining behavior. Meanings are effective even when false, even when C_n fails to indicate what it is its function to indicate, because C_n's meaning F is, in part, constituted by past tokens of this type *doing* (i.e., indicating F) what it is (as a result) C_n's function to do.

Department of Philosophy
Stanford University
Stanford, CA 94305
USA