LYNNE RUDDER BAKER

DRETSKE ON THE EXPLANATORY ROLE OF BELIEF

(Received 29 June, 1990)

Two or three decades ago, the status of explanations by reasons was uncertain. Then, with the assimilation of the view that reasons are causes, philosophers stopped worrying about the fate of reasons-explanations: To explain a person's behavior in terms of her reasons was just another way to explain behavior in terms of its causes. Recently, however, a new threat to reasons has come to light. Even if reasons are taken to be internal events that cause behavior, does the fact that the cause is a reason make any difference to the production of behavior?

The problem is this: Not every aspect of a cause is relevant to its producing an effect. To borrow a telling example from Fred Dretske, a soprano's high-*C* may shatter glass and it may mean, e.g., "Help!" But its meaning "Help!" had nothing to do with its effect on the glass, which is explained wholly in terms of the acoustic properties of the sound. If the high-*C* had meant something else or nothing at all, the glass would still have shattered. The worry is that the meaning of reasons may be as explanatorily irrelevant to behavior as the meaning of the soprano's high-*C* is to the shattering of the glass.

Dretske has developed an influential account of "how ordinary explanations, explanations couched in terms of an agent's *reasons*, explain."[1] In order to show how reasons explain, Dretske tightens the issue in the following way: Supposing that reasons are attitudes like beliefs, desires and intentions, and that beliefs, etc., are identified by their meaning, the task is to show how meaning (e.g., *what* one believes) can help explain behavior. The goal is to find an explanatory role for meaning in a world of causes.

Although Dretske's account is rightly regarded as ingenious, I believe that it falls to circularity. For, as I shall try to show, Dretske is

Philosophical Studies **63**: 99–111, 1991.

committed to the following three theses, construed without equivocation:

(1) *X*'s explanatory role is *X*'s causal role.

(2) A state *C* has an explanatory role in virtue of having meaning.

(3) A state *C* has meaning in virtue of having a causal role.

From (1)—(3), the following conjunction is deducible: A state has a causal role in virtue of having meaning, and has meaning in virtue of having a causal role. My aim here is to show that Dretske is committed to (1)—(3), and then to suggest what he might do about it.

First, I shall explain Dretske's key distinction between triggering causes and structuring causes — a distinction that allows Dretske to understand 'causal role' in such a way that semantic features (meaning or representing *F*) may be relevant to causal role, and vice versa. Then, I shall consider (1)—(3) *seriatim*: (1) is a standard "realist" view of explanation; (2) is the central desideratum for Dretske, and (3) follows from Dretske's account of natural representation.

STRUCTURING CAUSES

Since Dretske agrees with other physicalists that meanings per se do not cause motor output, he looks to another kind of causal role for meanings. He finds this in the idea of a structuring cause. We usually think of a cause as a token event that produces another token event; such causes are triggering causes. (E.g., the furnace ignited because a switch in the thermostat closed the cricuit.) But there is another kind of cause, says Dretske: the events that structured the process in the first place; such causes are structuring causes. (E.g., the switch in the thermostat was hooked up to the furnace because it is activated by temperatures at which the designer wants the furnace to come on.) Now apply this general distinction to the case of behavior.

Schematically, in the case of behavior, if *C* is an internal state (event type) that causes a bodily motion *M*, the behavior is *C*'s causing *M*.[2] Then, on some occasion on which the bodily motion is produced, the structuring cause is the background condition that

established the connection between C and M in the first place; the triggering cause is a token of C that produces a token of M.

Now Dretske's strategy is to find a role for meanings as structuring causes of behavior, where behavior is understood as a process of an internal event's causing a bodily motion. So, the question that Dretske seeks to answer is this: How can meaning or content, without the intervention of any designer with intentional states, have a causal role in structuring the causal connections between internal events and motor outputs? And his answer, in brief, is that C is an indicator that detects the presence of perceptually salient properties of the environment. Then, the claim is that C is "recruited" as a cause of M because of what C indicates. In this way, C's indicating F (for some F) is a structuring cause of C's causing M.

Note, however, that what can be explained by structuring causes are types: $C \rightarrow M$ as a type of process. A structuring cause explains why the $C \rightarrow M$ process is hooked up the way it is; it can not explain any actual tokening of a $C \rightarrow M$ process.[3] But the problem of explanations by reasons at least partly concerns tokens: We want to know how reasons identified by meaning explain Booth's shooting Lincoln. This question (about triggering causes) is not answered by showing how reasons explain Booth's being "structured" in such a way that when a certain internal event occurred, it caused a shooting.

So, even if Dretske were to accomplish his aims, we still may not be satisfied that the problem of mental causation has been laid to rest. Since I have addressed this issue in detail elsewhere,[4] I shall leave it aside here and turn to showing that Dretske is committed to (1)—(3), discussion of which will be facilitated by a more precise definition of 'structuring cause.' So, define 'structuring cause' as follows:

(D0) A state C's having extrinsic property P_e is a *structuring cause* of some behavior E if (i) C has been recruited to produce E and (ii) C would not have been so recruited if it had not had P_e.

If C's having P_e is a structuring cause of behavior E, then P_e has a causal role in the behavior. Dretske's project is to show that meaning or representation has a causal role in this sense. If successful, Dretske

will have shown how an internal state *C* has a causal role in virtue of having meaning.

EXPLANATORY ROLE AS CAUSAL ROLE

It is uncontroversial that Dretske is committed to (1). Like most physicalists, Dretske focuses exclusively on causal explanations, explanations that mention a cause of the event to be explained.[5] And since Dretske defines 'behavior' as a causal process and takes beliefs and other attitudes to be inner causes, his identification of explanatory role and causal role seems inevitable: "The project is to see how reasons — our beliefs, desires, purposes and plans — operate in a world of causes, and to exhibit the role of reasons in the *causal* explanation of human behavior." (EB, x; emphasis his.)

The novelty of Dretske's approach lies in his broadening the notion of causal explanation to include structuring, as well as triggering, causes. Identifying explanatory and causal roles is standard practice (whether it ought to be or not). Thus, Dretske endorses (1). Turn now to (2).

THE EXPLANATORY ROLE OF MEANING

The major purpose of Dretske's book is to find an explanatory (i.e., causal) role for belief (and other attitudes that figure as reasons) in terms of their contents: in terms of what they mean or represent. Thus, to achieve that goal, Dretske must find "a causal role for meaning" (EB, 80). That is, he must establish (2).

"The project," Dretske says, "is to understand how something's having meaning could itself have a physical effect — the kind of effect (e.g., muscular contraction) required for most forms of behavior . . ." (EB, 83) Or again: assuming that thoughts are physical states or structures in the brain, Dretske asks:

> [W]hat about the *meanings* of these physical structures? Are they, like the mass, charge, and velocity of objects, properties whose possession could make a difference, a *causal* difference, to the way these neural structures interact? If meaning, or something's *having* meaning, is to do the kind of work expected of it — if it is to help explain *why* we do what we do — it must, it seems, influence the operation of those

electrical and chemical mechanisms that control muscles and glands." (EB, 80; emphases his.)

Although there is some looseness in terminology, to find an explanatory role for belief is to find an explanatory role for the semantic properties of structures: "We are, remember, looking for an *explanatory* role for belief and, hence, an explanatory role for the semantic properties of a structure." (EB, 81; emphasis his.) Dretske suggests that to find an explanatory role for belief is to find "a genuine case where an element's semantic character helps to determine its causal role in the production of output." (EB, 95) 'Semantic character' here must be taken to be representation or meaning or content if it is to provide an explanatory role for belief.[6]

Thus, to find an explanatory role for belief is precisely to find a structure or state that has an explanatory role in virtue of its having meaning. That is, belief has an explanatory role only if (2) is true. Therefore, Dretske is committed to (2).

REPRESENTATIONAL SYSTEMS

To show that (3) is a central tenet of Dretske's, I shall formulate a series of interlocking definitions, culminating in a definition of a natural representational system. The basic idea underlying Dretske's account of representation is that of indication. As Dretske introduces the idea of indication, indication is a relation between a *token* and an external condition: 'This doorbell's ringing indicates that somebody is at the door' is rendered roughly as '(i) There is a dependency of doorbell ringings on people at the door, which results in nonfortuitous correlation between ringings and people's being at the door, and (ii) that correlation is holding now (i.e., somebody is at the door).' (EB, 57) Although he introduces indication as a property of tokens, Dretske deploys it as a property, not of tokens, but of *types* of internal states. For example, when he says, "*C* will normally indicate a great many things other than *F*" (EB, 84), he obviously is speaking of *C* as a type, not of a token of *C*, which can only indicate what actually caused it. So, let me begin by defining 'indication' as a relation between types:[7]

(D1) A state of *type C indicates* an external condition *F* iff there is a dependency of tokens of *C* on occurrences of *F* such that, in normal conditions, all and only occurrences of *F* cause tokens of *C*.[8]

Then, as a property of tokens, indication may be defined as follows:

(D2) A *token t indicates* an external condition *F* iff (i) *t* is of a type that indicates *F* and (ii) the conditions in which *t* occurs are normal.

According to (D2), a token either indicates what causes it or it indicates nothing at all. There is no such thing as *mis*indication.[9] However, in order for there to be representation, there must be the possibility of misrepresentation:

> What we are after is the power of a system to say, mean, or represent (or, indeed, *take*) things as *P whether or not P is the case*. . . . Whatever word we use to describe the relation of interest (representation? meaning?), it is the power to misrepresent, the capacity to get things wrong, to say things that are not true, that helps *define* the relation of interest." (EB, 65)

Thus, the idea of bare indication does not suffice for an account of meaning or representing anything. We get from bare indication to a property more suitable as a basis for representation by determining what it is the *function* of an internal state to indicate. What converts a state from being an indicator of *F* to having the function of indicating *F* is its *being recruited* to perform a certain task because it indicates *F*. What a mental state is recruited to do is to produce behavioral output. So, we may define 'having the function of indicating' as follows:

(D3) A state of type *C* has the *function of indicating external condition F* iff (i) state *C* is of a type that indicates *F*, (ii) state *C* has been recruited to play a role in producing some behavior *E*, and (iii) state *C* would not have been so recruited if it had not indicated *F*.

The last clause allows that state *C* may have the function of indicating *F* even if some tokens of *C* are not caused by *F*. The above distinction between indication for types and indication for tokens

makes for a tidy statement of "wildness:" a token *t* of type *C* is "wild" iff *t* does not indicate *F*, but type *C* does indicate *F*.

Since (D3) allows artifacts (e.g., the gas gauge in my car) to have the function of indicating a condition, and since such functions are assigned by designers and users of the artifacts, the account is not yet suitably naturalistic. What we want are natural functions, functions acquired without intervention of any intentional agent. This suggests:

(D4) A state of type *C* has the *natural function of indicating external condtion F* iff (i) state *C* has the function of indicating *F*, and (ii) state *C*'s recruitment was brought about without intervention of any intentional agent.

Although (D4) is an intuitive suggestion, it may not be Dretske's. In order to eliminate the possibility of an homunculus or "understander-or-meaning" in the conferral of function, Dretske looks to discriminative learning. Although he countenances biological examples as illustrations of "intrinsic functions," he says that "the case for representational systems of Type III [natural representational systems; see below.] will rest on quite different sorts of functions: those that are derived, not from the evolution of the species, but from the development of the individual." (EB, 64) Although I am not sure how to interpret this, it at least suggests:

(D4′) A state of type *C* has the *natural function of indicating external condition F* iff (i) state *C* has the function of indicating *F*, and (ii) state *C*'s recruitment was brought about entirely by discriminative learning.[10]

For purposes here, it does not matter whether Dretske endorses (D4) or (D4′). Each immediately leads to a naturalistic account of representing *F*.

(D5) A state of type *C* represents *F* (in the first instance) iff it has the natural function of indicating external condition *F*.

Note that (D5) defines 'representing *F* (in the first instance),' as opposed to 'representing *F*' simpliciter. The qualification, "in the first instance," is intended to restrict application of the definition in two ways. First, the definition is to apply only to naturalistic systems —

those that represent without the intervention of any intentional agent or of any homunculus in the head. Second, it is to apply only at the "ground level" of representation and not to representation of things for which we have no internal indicators. (This leaves room to add further definitions to accommodate representation of, say, unicorns and bachelors.) Now, the definitions of 'natural representation' and 'meaning' follow immediately.

(D6) A state of type *C* is a natural representation iff it is a state that represents *F* (in the first instance).

(D7) A state of type *C* has meaning (in the relevant sense) in virtue of its being a natural representation.

For completeness, add:

(D8) A system is a natural representational system iff it has a state that is a natural representation.

Natural representational systems are also called by Dretske 'Type III representational systems' to contrast them with Type II systems, like thermostats, which have assigned rather than natural functions of indicating *F*, and with Type I systems, like representations of basketball games with pieces of popcorn, all of whose representational powers derive from their creators.

(D1)—(D8), I believe, give Dretske's naturalistic account of representation. Since they are definitions, the biconditionals are not intended to be material biconditionals, but rather they must be interpreted in a substantially stronger sense — part of which is captured by 'in virtue of'.

Assuming that the "in virtue of" relation is transitive, we get the following chain: A state *C* has meaning in virtue of being a natural representation by (D8); a state *C* is a natural representation in virtue of having the natural function of indicating *F* by (D6) and (D5); a state *C* has the natural function of indicating *F* in virtue of having the function of indicating *F*, where the function of indicating *F* has a certain etiology by (D4). So, a state *C* has meaning in virtue of having the function of indicating *F*, where the function of indicating *F* has a certain etiology.

Now (D3) explicates the idea of having the function of indicating *F* in part in terms of (ii) [state *C* has been recruited to play a role in producing behavior] and (iii) [state *C* would not have been so recruited if it had not indicated *F*]. But (ii) and (iii) of (D3) are just the definition of a structuring cause in (D0); let P_e be the property of indicating *F*, and look at the definiens of (D0). From (D3) and (D0), we can derive (D3′), which makes explicit the dependence of having the function of indicating *F* on being a structuring case of behavior:

(D3′) A state of type *C* *has the function of indicating F* iff the following: (i) State *C* is of a type that indicates *F*, and (ii) *C*'s indicating *F* *is a structuring cause* of behavior some behavior E.[11]

So, a state *C* has meaning in virtue of the fact that *C*'s indicating *F* is a structuring cause of behavior, where it became a structuring cause by natural means. Therefore, the state *C* has a property (viz., indicating *F*) that is a structuring cause. Assuming that a state has a causal role if it has a property that is a structuring cause, state *C* has meaning in virtue of having a causal role, where its causal role came about by entirely natural means. This is to say, the definitions that embody Dretske's naturalistic account of representation entail

(3) A state *C* has meaning in virtue of having a causal role.

We can also see that the causal role referred to in (3) and the explanatory role referred to in (2) are one and the same. The relevant role for both is the role conferred by the fact that *C*'s indicating *F* is a *structuring cause* of some behavior *E*. Let '*S*' stand for 'the fact that *C*'s indicating *F* is a structuring cause of some behavior *E*.'[12] Then, putting it awkwardly, by (2), *C* is such that *S* in virtue of having meaning; but by (3), *C* has meaning in virtue of *S*.

This completes my argument that Dretske's theory commits him to (1)—(3) and hence to the circularity that they embody.

WHAT'S TO BE DONE?

If we can discover what generates the circle, we can see how Dretske can get out of it — though at the cost of denying an explanatory role to

belief. The source of the circle, I believe, lies in lack of appreciation of the following: since C represents F in virtue of the fact that C's indicating F plays a causal role in structuring the process $C \rightarrow M$, C does not mean or represent anything unless the $C \rightarrow M$ process is structured. Therefore, the fact that C has meaning can not help structure the $C \rightarrow M$ process. In short, on the one hand, the only causal role that Dretske sees for meaning is in structuring the $C \rightarrow M$ process; but on the other hand (as (D3′) shows), there is no meaning or represenation unless the $C \rightarrow M$ process is already structured. So, meaning can not have a causal role in behavior, on pain of circularity. And (2) should be rejected.[13]

This point has gone unnoticed, I believe, because of a subtle slippage between 'indicating F' and 'having the function of indicating F.' That Dretske does not always observe the difference is revealed in remarks like this: The "causal relationship between C and M, if it is going to be explained by something like the meaning of C, will have to be explained by the fact that *C indicates, or has the function of indicating*, how things stand elsewhere in the world." (EB, 84; emphasis mine.) However, the difference between indicating and having the function of indicating is crucial in the context of explicating representation or meaning.[14]

On the one hand, in order for something to represent F, it must have the function of indicating F. Mere indication is not enough. For there is no such thing as misindication, but representation is partly defined, as Dretske says, by the possibility of misrepresentation. But on the other hand, we see from the definitions (and from the diagram on p. 84) that the causal role belongs only to indication: the structuring cause of $C \rightarrow M$ is C's indicating F, not C's having the function of indicating F. If the structuring cause is C's indicating F, then the property that has the causal role is that of indicating F, not that of having the function of indicating F. So, from the claim that indicating F has a causal role, it is a non sequitur to conclude that the property of having the function of indicating F has a causal role, and a fortiori, a non sequitur to conclude that representation or meaning has a causal role.[15]

The slippage conceals the difference between these two theses:

(A) Representation (or meaning) has a causal role because the $C \rightarrow M$ connection would not have been structured in the way that it is if C had not represented (or meant) what it did.

(B) Indication has a causal role because the $C \rightarrow M$ connection would not have been structured in the way that it is if C had not indicated what it did.

As we saw in the subsection, "The Explanatory Role of Meaning," Dretske sometimes sounds as if (A) were his thesis. Indeed, he clearly implies that meaning or content (not just indication) accounts for causal role when he characterizes (defines?) beliefs as "those representations whose causal role in the production of output is determined by their meaning or content — by the way they represent what they represent." (EB, 52)

At other times, however, Dretske sounds as if (B) were his thesis. For example, when Dretske refers to "the relation between C's indicating F and C's causing M" as "the explanatory relationship" (EB, 84), he is assigning to indication (not to representation or meaning) the causal role. And as we saw in the subsection, "Representational Systems," his naturalistic reduction of representation requires assigning to indication (not representation) the structuring causal role.

Although we have evidence that each of (A) and (B) is Dretske's thesis, neither will serve his purposes. (A) underwrites (2) and thus leads straight to vicious circularity; (B) avoids circularity but gives no causal or explanatory role to meaning or representation, and hence no causal or explanatory role to belief.

The best that Dretske can get without circularity is an explanatory role for indication: There is a single underlying relation (indication) that both has a causal role and gives rise to meaning or representation.[16] But such a thesis is fully compatible with an epiphenomenal view of representation or meaning.

In sum: To accept the suggestion for getting out of the circle would be to give up the account of the causal role of meaning. And since, for Dretske, causal role is explanatory role, it would be to give up the

account of the explanatory role of meaning. In that case, although the view would be free of circularity, Dretske would be back where he started with respect to understanding "how ordinary explanations, explanations couched in terms of an agent's *reasons*, explain."[17]

NOTES

[1] Fred Dretske, *Explaining Behavior: Reasons in a World of Causes* (Cambridge MA, MIT/Bradford, 1988): 52. Hereafter, references to this work will appear in the text as "EB" followed by a page number.

[2] In EB, Dretske construes behavior, not as bodily motion, but as a process, $C \rightarrow M$, which rsults in bodily motion M. But his account is equally applicable to a more standard construal of behavior as M.

[3] One may say, derivatively, that structuring causes explain tokens, if any, of the $C \rightarrow M$ process (type) that they structure. But we can not confine attention to actual tokenings since structuring causes also must explain why if a token of C had occurred, it would have produced a token of M.

[4] "Metaphysics and Mental Causation" in *Mental Causation*, John Heil and Albert Mele, eds. (Oxford, forthcoming).

[5] Indeed, there is a vast metaphysical picture about the nature of causality, the "closed causal order," events and so on that underlies this attention to causal explanation. For a good discussion of the metaphysical background, see Jaegwon Kim, "Explanatory Realism, Causal Realism, and Explanatory Exclusion" in *Realism and Antirealism (Midwest Studies in Philosophy XII*, Peter A. French et al, eds. (Minneapolis: University of Minnesota, 1988): 225–239.

[6] Since what is called "natural meaning" obviously would not suffice for giving belief an explanatory role, I shall use 'meaning' to apply to what is called "non-natural meaning."

[7] To be more precise, what (D1) defines is the lawful dependency relation between types in virtue of which a token of C indicates a token of F. I am calling this relation *(type) indication*, because Dretske's account of meaning rests on the notion of C's indicating F, where C and F are types. The diagram on p. 84 would misrepresent his view if C and F were not types.

[8] Although I doubt that normal conditions can be specified in any non-question-begging way, I shall not discuss the matter here.

[9] Although I do not want to presuppose any particular detailed account of indication, I believe that they all preclude the possibility of misindication.

[10] I have doubts about Dretske's deployment of discriminative learning. (i) Dretske's crude stimulus-response model of discriminative learning seems far too impoverished to account for acquiring meaning. (A richer account of discriminative learning, according to which S—R connections are themselves mediated by contentful internal states, may be more adequate to understanding how having the natural function of indicating F is acquired, but the richer account would violate the strictures of Dretske's naturalism). (ii) If learning is to be the mechanism by which an organism comes to represent *Fs* (i.e., acquires the natural function of indicating F), then it is crucial to be able to distinguish the learning period (in which all tokens of C indicate F) from the postlearning period (in which some tokens of C may fail to indicate F). I see no hint of any principled demarcation.

[11] Notice that any state that satisfies (ii) of (D3′) ipso facto satisfies (i).

[12] Notice that one can not avoid the circle by saying that C's indicating F is a

structuring cause of some behavior, but C's explanatory role derives from C's causing some other behavior. First, this appears to be a confusion between triggering and structuring causes. Second, if we stick to structuring causes, as we must to avoid equivocation, the same behavior in which C has a causal role (in virtue of the fact that C's indicating F is a structuring cause of it) is to be explained by appealing to what C means. If we couldn't use a single instance of 'some behavior' to explicate both (2) and (3), then (1) would be false.

[13] I am not hereby denying an explanatory role to belief; I am only saying that such denial is Dretske's best way out of the circle.

[14] The equivocation could just as well be stated in terms of natural and non-natural meaning: Non-natural meaning is required for representation, but it is only natural meaning that is the structuring cause of the $C \rightarrow M$ process.

[15] Thus, Dretske should not associate his view with what Stich calls the Strong Representational Theory of the Mind. (EB, 81)

[16] An anonymous referee suggested that what I have shown is that what we might call 'efficient structuring causes' are (in the first instance) indicator properties and not representional properties. However, what we might call 'sustaining structural causes' — i.e., "events and states that are causally contributory to the maintenance of the power of a representation to produce some bodily movement" — might (and presumably normally would) include the fact that the representation has the function of indicating some property. This move seems to me tantamount to embracing the circle.

[17] At a conference on Mental Causation in Bielefeld, West Germany, in March 1990, I read a recent ancestor of this paper in response to Dretske's "Mental Events as Structuring Causes." Derk Pereboom and Hilary Kornblith made useful criticisms of my Bielefeld paper. I have benefited from helpful discussions with Fred Dretske and wish to thank him for comments on a draft of this paper after the conference.

University of Massachusetts at Amherst
and Middlebury College
Department of Philosophy
352 Bartlett Hall
Amherst, MA 01003
USA