TELEOFUNCTIONALISM AND PSYCHOLOGICAL EXPLANATION

Pacific Philosophical Quarterly 87 (2006), 403-421

Jason Bridges
University of Chicago

Abstract: Fred Dretske's teleofunctional theory of content aims to simultaneously solve two ground-floor philosophical puzzles about mental content: the problem of naturalism and the problem of epiphenomenalism. It is argued here that his theory fails on the latter score. Indeed, the theory insures that content can have no place in the causal explanation of action at all. The argument for this conclusion depends upon only very weak premises about the nature of causal explanation. The difficulties Dretske's theory encounters indicate the severe challenges involved in arriving at a robust naturalistic understanding of content.

Content epiphenomenalism is the view that mention of intentional content has no legitimate place in causal explanation—that, for example, one cannot causally explain a person's actions by citing what she believes and desires. Since explanations of just this sort appear to be a fixture of our everyday thought about other people and ourselves, content epiphenomenalism is a radical and surprising doctrine. Nonetheless, a number of arguments, based on intuitive-sounding metaphysical principles, can be marshaled in its favor.

Naturalism is the view that everything that happens or is the case is within the explanatory reach of the natural sciences. This is not merely to deny the existence of the supernatural. As most contemporary philosophers understand it, naturalism demands in addition the 'naturalization' of any putative fact or phenomenon expressed in a non-natural-scientific discourse—that is, the construction of a theory that displays the fact or phenomenon in question as constituted by states and properties that belong to the ontology of some natural science. Since our everyday ('folk') psychology of beliefs, desires and other content-bearing states and occurrences is not a natural science on any plausible construal of that category, naturalists who do not want to dismiss folk psychology as an illusion are obligated to 'naturalize' beliefs, desires and the rest.

The foregoing briefly summarizes two issues about mental content that have received a great deal of attention in contemporary philosophy of mind. Unsurprisingly, in light of the bent toward specialization in the contemporary philosophical literature, work on each of these problems has tended to proceed in independence from work on the other. It is thus a distinctive merit of Fred Dretske's theory of mental content that it aims to solve both problems at once. Indeed, Dretske's guiding methodological thought is that considering the two problems together will yield new insights pointing the way to a joint solution.

The theory that results from this exercise is a form of *teleofunctionalism*: it takes the contents of inner physical states to be constituted by facts about the purpose, or function, of those states. An inner state's function is said in turn to be constituted by facts about the past processes of reinforcement that shaped that state's causal role in the organismic system. And finally, the linkage to reinforcement history is said to yield an innovative account of the non-epiphenomenal character of content: Dretske argues that philosophers tempted toward content epiphenomenalism have overlooked the special kind of causal explanation of behavior yielded by reflection on reinforcement history.

The aim of this paper is to show that Dretske's theory fails in its dual goal. Rather than turning away the threat of content epiphenomenalism, it makes content epiphenomenalism inevitable.

Commentators on Dretske have made this charge before.[1] I revisit the issue for two related reasons. First, it seems to me these commentators did not make a good case for the charge; their arguments contain substantial errors or lacunae. This is evinced by Dretske's mounting of effective counter-responses.[2] The issue remains unresolved, and a careful and convincing account of why Dretske's account cannot succeed in its goal has yet to be provided.

The second reason is this. After an initial flurry of articles in the years following the publication

[1] See several of the articles in McLaughlin, 1991, especially the piece by Cummins.

of Dretske's *Explaining Behavior*, discussion of his account has largely died down. This reflects a general shift in the literature away from content epiphenomenalism and related issues about content, and toward questions about the status of qualia and other elements of consciousness. There is nothing wrong with trends; interests wax and wane. But the questions of the role of mental content in explanation and of naturalism's prospects are ground-floor problems for the philosophy of mind, and will remain so. A close examination of one of the most inventive and interesting attempts to solve these problems in recent decades can help provide a fresh look at these fundamental issues.

1. A brief synopsis of Dretske's account of the causal-explanatory role of content

A. The role of information in the explanation of behavior

The carrying of *information* is an extrinsic property of tokenings of states. It is a form of nomic dependency: an occurrence of a state carries the information that F iff it is a law of nature (perhaps just a *ceteris paribus* one) that the state occurs only when it is the case that F.[3] For Dretske, the first step in meeting the threat of content epiphenomenalism is to grasp the role of information in the explanation of behavior. That role comes into view when we reflect on the familiar phenomenon of operant conditioning (a.k.a., instrumental learning).

Imagine an organism, call him O, whose circumstances are such that engaging in behavior M yields a beneficial result R when and only when condition F obtains in the local environment.[4] (O might, for example, be a pigeon that gets a pellet for pecking a button when and only when a certain light is on.) Suppose O's nervous system has sufficient plasticity and sophistication that repeated instances of his receiving R when performing M in condition F, coupled with repeated instances of

[2] See especially his replies in McLaughlin, 1991.

[3] Dretske himself idiosyncratically construes information in terms of probabilistic dependency. Nothing we discuss here turns on which construal we adopt. Note also that in the writings that will be our focus here, Dretske sometimes speaks of "indication" rather than "information": an occurrence indicates F iff it carries the information that F.

his not receiving R when performing M in the absence of F, gradually makes it the case that O can be counted upon to perform M in and only in the presence of F. How does this happen? Dretske suggests that there is an obvious answer: 1) there is a neurophysiological state C of O tokenings of which indicate (carry the information that) F, and 2) O's brain is gradually rewired so that Cs (i.e., tokenings of C) are what cause performances of M. In Dretske's terminology, Cs are "recruited" as causes of M. Dretske does not offer an account of how internal indicators are recruited as causes of behavior—indeed, he finds this process a "complete mystery"—but he is certain it happens nonetheless: "Learning cannot take place *unless* internal indicators of F are harnessed to effector mechanisms in some appropriate way. Since this learning *does* occur, the recruitment *must* take place" (Dretske, 1988, p. 98).

The significance of the process of recruitment for our topic emerges when we consider how things stand once the process is complete. Suppose at this stage that a C causes a performance of M. What is the explanation of the M-performance? As Dretske sees it, this question could be taken in either of two ways. First, it might be construed as asking for an account of the neurophysiological process that led from the C to the M. This Dretske calls a *triggering explanation* of the M, for it explains how we get from the C, the *triggering cause*, to the M. For the purposes of this explanation, the explanatorily relevant property of the C will just be that it is a C—that it is a token of that neurophysiological type.

But the question can be taken another way—not as seeking a blow-by-blow of the C→M neurophysiological process, but as asking why O's brain is wired in the first place such that Cs initiate this process. Cs cause Ms because that's how O's brain is structured, but we may well ask how O's brain came to be structured in this way. Answering this question requires identifying the *structuring cause* of the M—that is, the source or origin of the structure implicated in the process by

[4] This paragraph and the next summarize section 4.4 of Dretske, 1988.

which the C triggered the M. And according to Dretske, this answer cannot itself be provided in solely neurophysiological terms. Rather, an adequate structuring explanation of the M must appeal to the fact that Cs indicate, or at least used to indicate, F. As we have seen, it is because tokens of C indicated F—the environmental condition necessary and sufficient for M's yielding the reward R—that these tokens were recruited as causes of performances of M (Dretske, 1988, p. 101).

The upshot of this train of thought is that an informational property of certain internal items—namely, the Cs—proves relevant to the complete explanation of O's behavior, for it is crucial to the structuring explanation of that behavior.[5]

B. From information to teleofunction to content

Since Dretske's aim is to give a naturalistic account of content that shows content to be relevant to the explanation of behavior, and since he has a story about why informational properties are so relevant, why not just identify content properties with informational properties—why not hold, for example, that having the content *F* consists in indicating F? That content properties are informational properties is the defining claim of *informational semantics*, an approach to naturalized semantics which, although Jerry Fodor became its primary apostle, was originally developed by Dretske himself (Dretske 1981/1999).

But by the time of the writings that concern us in this paper, Dretske had become convinced that informational semantics was incorrect, and he had been led to this conclusion in part precisely

[5] In my view, Dretske never gives a satisfactory general account of the structuring/triggering distinction. Things he says in different places on this score seem to me inconsistent and incomplete. But all that matters for our purposes is his application of the distinction to the case of operant conditioning; accordingly that is what I focus on in the text. One caveat: Dretske tends to speak of structuring causes as causes of processes rather than as causes of particular events. He does so because this dovetails with his view that behavior is to be identified with processes eventuating in bodily movements rather than with the bodily movements themselves. Because I think this view of behavior is quite implausible, and because the view is not essential for understanding the shape or motivation of Dretske's account of the explanatory role of content (as evinced by 1993, in which Dretske presents his account of content without

because of the line of thought just traced. According to that line of thought, informational properties of tokens of internal states of a creature bear on the explanation of its behavior only by helping to explain why these states are recruited as causes of behavior—why the brain gets wired so that occurrences of that state trigger behavior. It is thus an informational property of *past* Cs, in particular their indicating F, that matters for C's recruitment for the causal role that it now possesses. The informational properties of *present* Cs are no part of the explanation of why present Cs cause Ms; indeed, tokenings of C might have ceased to indicate F and the explanation would still go through. Thus if we identify an item's having the content *F* with its indicating F, it follows that it is not a present C's own possession of content that matters for the explanation of its impact on behavior; the relevant content properties are that of past Cs. Even if past and present tokens of C had the same content, it would be the past Cs' having that content, and not the present Cs' having that content, that explains why present Cs cause what they do. And this result falls short of what we want. Our aim, in seeking to refute content epiphenomenalism, is to show that a given thought (belief, desire, etc.) causes the action that it does because *that* thought has the content that it does—not because, say, some long-ago thought had the same content. It is, we assume, my currently thinking a thought with the content *Fire!* that explains my now running out of the theater, not my having had thoughts with that content in the distant past. As Dretske puts the point: "If the behavior M is to be explained by what the agent believes, it will have to be explained by what present tokens of C mean. How does the meaning of *the current* C explain *its* causing M?" (Dretske, 1990b, p. 830)

Dretske's response to this question is to abandon informational semantics in favor of (although he himself doesn't use this term) *teleofunctional* semantics. On Dretske's revised account, an item's possession of a given content is a matter of what information the item is *supposed* to carry, in the

assuming the process-view of behavior), I shall ignore this aspect of Dretske's exposition here.

sense of what information the item has the *function* of carrying. In particular, rather than holding that an item's possession of the content *F* consists in its indicating F, Dretske suggests that its possession of that content consists in its having the function (or "teleofunction") of indicating F.[6]

We assign functions to things, but no one assigns functions to states of our brain. Any such functions must have been acquired naturally, without the intercession of intentional agents. Dretske's suggestion is to look to the very reinforcement history that we have seen to be central to understanding the explanatory relevance of informational properties. If an organism gets a reward R by performing a behavior M when and only when condition F obtains in the environment, and if C is an internal state whose tokenings indicate F, then C may end up being recruited as the cause of M. Once the process of recruitment is complete, it provides us with a basis for saying that present Cs cause M-performances because past Cs indicated F. And, we may plausibly add, it provides us with a basis for saying that present Cs have the *function* of indicating F. Says Dretske: "Since these indicators are recruited for control duties *because* of the information they supply, supplying this information becomes part of their job description—part of what they, once recruited, are *supposed to do*" (Dretske, 1988, p. 99).

This move seems to offer an ingenious solution to our puzzle. Since informational properties of tokenings of internal states like C are relevant only to structuring explanations of behavior, a given C's own informational properties do not help explain any performance triggered by that C. It is the informational properties of past tokens of that type, past Cs, that help explain the performance. But on Dretske's view, the content of a given token is not tied to that token's informational properties. It is tied rather to the informational properties of past tokens of that type, indeed, to the informational properties of the very tokens that figure in the structuring explanation

[6] The point of the unappealing piece of jargon "teleofunction" is to avoid confusing the idea of what an item is supposed to do with the idea of an item's causal role as it is understood by functionalists. I shall sometimes avoid the jargon, however, and take the sense of "function" to be clear from context.

of the behavior triggered by the current token of that type. For a present C to have the function of indicating F just *is*, on this view, for there to be a structuring explanation of the behavior triggered by that C that appeals to the fact that C was recruited for its causal role because past Cs indicated F. The fact that past Cs indicated F, coupled with the fact that their doing so explains why the brain is now structured so that current Cs cause the behavior that they do, are precisely what enable us to say that current Cs are *supposed* to indicate F. Thus by equating present Cs' having the content *F* with their possession of that function, Dretske ensures that present Cs' having the content *F* is constitutively connected to past Cs' indicating F. It follows, says Dretske, that anything explained by the latter is explained by the former. Present content is exactly as relevant to the explanation of behavior as past information, for present content is constituted by the explanatory role of past information.

Dretske summarizes the maneuver in the following passage (note the Gricean terminology: "non-natural meaning" = content; "natural meaning" = information).

> [T]he idea is that the sort of meaning attaching to psychological states (particularly belief) is a historical property: to say that S believes F is to say that there is some C inside S that means (non-naturally) F, and this, in turn, is to say something about the way past Cs—in virtue of what they meant (natural meaning)—changed the way the system was causally organized. Present meanings explain present behavior, but only because both meaning and behavior (at least the structuring explanations of behavior) are backward looking phenomena (Dretske, 1991, p. 216).

We may encapsulate the salient features of Dretske's account as follows:

> **Dretske's account of content:** Where C is an inner physical (neurophysiological, computational) state of an organism and F is an environmental condition, that current tokens of C have the content *F* consists in its being the case that current Cs cause behavior because past Cs indicated F.[7]

[7] It should be noted that this formulation of Dretske's account incorporates an ambiguity in his own presentation. The question is whether the informational properties of past Cs—and so, on Dretske's view, the content of present Cs—explain why present Cs cause the specific behaviors they do, or merely explain why present Cs cause any behavior at all. I believe that neither resolution of the ambiguity is satisfactory: the latter gives us less than we seek, and the former runs up against difficulties having to do with the holistic character of propositional-attitude explanation. But I do not have the space to address this topic here and in any case we do not have to decide: the problem I will outline arises on either interpretation. I will continue to characterize Dretske's target explanandum simply as being that presents Cs cause

2. Explanatory relevance

As we have seen, there are two parts to Dretske's account of how content can be relevant to the causal explanation of behavior. The first part is his attempt to show, through reflection on operant conditioning, that the informational properties of tokens of an internal state can bear on the explanation of what later tokens of that state cause. The second part attempts to piggyback upon this story an account of how the contents of tokens of an internal state bear on the explanation of what those tokens cause. In what follows, we will grant the first part of Dretske's story. I will argue that even so, the second part cannot be sustained. Indeed, Dretske's account achieves the opposite of its intent: it ensures that content cannot be relevant to the causal explanation of behavior. If Dretske's account of the nature of content is accepted, content cannot be anything but epiphenomenal.

What are we asking when we ask whether content, or anything else for that matter, is "relevant" to the causal explanation of something? We may wish to say, as many philosophers do, that questions about causal-explanatory relevance are questions about which properties of causes are the properties *in virtue of which* those causes cause their effects. But now the question becomes: what is it for a given property of a cause to be that "in virtue of which" it causes what it does? One answer to this question familiar from the literature is that the causally relevant properties of a given cause, the properties in virtue of which it produces an effect, are those that satisfy certain *counterfactuals*. The basic thought is that a property p of a cause E_1 counts as causally relevant to an effect E_2 iff E_2 wouldn't have occurred had E_1 not had p. The basic thought, unfortunately, turns out to be inadequate—it fails to capture our intuitions about a range of cases—and those sympathetic to the

behavior, and leave open questions about the specificity of the behavior explained.

counterfactual approach have been led to proffer increasingly baroque alternatives.[8]

My own view is that the counterfactual approach is flawed in principle.[9] But we do not need to wade into this issue here. To set the stage for the argument against Dretske's view, we need only register an exceedingly minimal claim about what is at stake in questions of causal-explanatory relevance. This claim is consistent with the counterfactual approach as well as with its rejection.

The claim I have in mind is articulated by Dretske in the following passages:

> Though the causal barriers separating mind and body are thus removed (thanks largely to Donald Davidson), *explanatory* barriers remain intact…From a practical standpoint, it may be useful to know what a person is thinking if he generally does A whenever he thinks T, but if he doesn't do A, in part at least, *because* he thinks T, why, in our philosophical (not to mention scientific) efforts to understand why people do what they do, should we be interested in *what* (or even *that*) they think? (Dretske, 1990a, p. 5-6)
>
> A satisfactory model of belief should reveal the way in which *what we believe* helps to determine *what we do*…We can, following Davidson (1963), say that reasons *are* causes, but the problem is to understand how their being reasons contributes to, or helps explain, their effects on motor output….In exploring the possibility of a causal role for meaning one is exploring the possibility of a *thing's having a meaning* being a cause... (Dretske, 1988, pp. 79-80).

The principle Dretske asserts is simply this: if content epiphenomenalism is to be avoided, it must sometimes be the case that people behave as they do *because* their thoughts have the contents they do. Or as he puts it in the second passage, it must sometimes be the case that an item's *having a given content* is a cause of behavior.

Now, I take these to be two ways of saying the same thing: to say that x's having property p is a

[8] For a proposal roughly equivalent to what I call the basic thought, see Lepore and Loewer, 1987. For a sophisticated late-edition model, see Yablo, 2003.

[9] More specifically, what I take to be flawed is the attempt to explain causal relevance in terms of counterfactual dependence (rather than simply taking criteria of the latter sort to be a defeasible guide to the former). The problem is that counterfactually-framed conditions cannot provide an *independent* basis for adjudicating judgments about the relevance of properties to causal explanations. Our judgments about these counterfactuals tacitly rely upon assumptions about which properties are relevant to which kinds of causal explanation: absent any assumptions of this sort, we cannot interpret, much less evaluate, the relevant counterfactuals. If this point is not generally recognized, that is because otherwise admirable work in the formal semantics of counterfactuals can serve to suppress it. In the case of a proposal like Yablo's, the problem lies in the notions, taken as primitive, of the naturalness of properties and the nearness of possible worlds. I strongly doubt such notions have intelligible application in abstraction from any causal-explanatory facts about anything.

cause of y is to say that x caused y because x has p. We need to register a caution here, however. Although Dretske's usage of the word "cause"—in which an item's having a property is said to be a cause, and is said to be so in virtue of what it causally explains—is perfectly good English, it is not part of the linguistic register of contemporary analytic philosophy. In the years since Davidson's seminal work on the difference between statements that assert the existence of causal relationships (between events) and statements that explain these relationships (perhaps by citing properties or facts), philosophers have largely come to reserve "cause" for items that bear the causal relationship so construed rather than for items that explain instances of that relationship. To avoid possible confusion, I shall rely on Dretske's first formulation of the principle.

We can put the 'account' of causal-explanatory relevance suggested by Dretske's passages as follows:

> Where E_1 is an event that causes an effect E_2 and p is a property of E_1, E_1's having p is *explanatorily relevant* to E_1's causing E_2 iff E_1 causes E_2 because E_1 has p.

Dretske's claim in the quoted passages is that content epiphenomenalism is avoided only if the contents of inner items are sometimes explanatorily relevant, in this sense, to their causing the behavior that they do.

It is hard to imagine anyone disagreeing with this claim. Surely no one disputes that the issue is whether inner items cause what they do because they have the content they do. As Stephen Schiffer puts it, "The explanatory role, the 'causal relevance', of psychological properties is nothing over and above their ability to occur in true 'because'-statements" (1991, p. 10). Perhaps "nothing over and above" is meant here to signal skepticism about the possibility of saying anything more about the nature of explanatory relevance than the simple point about "because"-statements, skepticism that, as we have noted, not all philosophers will share. But surely no one could disagree that this is the *first* thing to be said about explanatory relevance. At any rate, no one should disagree with it, for we shall later verify, it is obviously correct.

In spite of its truistic status, the requirement that ascriptions of content figure in "because"-statements of the form we have identified is sufficient to generate a decisive difficulty for Dretske's account. The argument for this claim is the topic of the next two sections.

3. The explanatory role of content: redundant?

Let C be a state current tokens of which have the content *F*. If content epiphenomenalism is false, then according to the view of explanatory relevance we have just seen Dretske to endorse, contentful items must sometimes cause behavior *because* they have the content they do. Let us suppose that current C tokens are such contentful items. Thus:

> **1.** *Denial of content epiphenomenalism:* Current Cs cause behavior because they have the content *F*.

Now, according to Dretske's account, that current Cs have the content *F* consists in the fact that C was recruited for its current behavior-causing role because past Cs indicated F. Abstracting from the details of the particular process of "recruitment" Dretske envisions, we have summarized his view thusly:

> **2.** *Dretske's account of content:* That current Cs have the content *F* consists in the fact that they cause behavior because past Cs indicated F.

From **1** and **2** follows:

> **3.** Current Cs cause behavior because (currents Cs cause behavior because past Cs indicated F).

The principle presupposed in the inference from **1** and **2** to **3** can be put as follows: if a state of affairs S obtains because x has property p, and if x's having property p consists in x's having property q, then S obtains because x has property q. If x's being p *consists in* its being q—if x's being q is *what it is* for x to be p—then whatever is explained by x's being p is explained by its being q. The explosion caused the beeping of the Geiger counter because the explosion was radioactive. Being radioactive consists (let's say) in emitting ionizing radiation. Thus the explosion caused the beeping of the Geiger counter because it emitted ionizing radiation. This inference cannot fail to be sound. Emitting ionizing radiation is just what it *is* to be radioactive. If the explosion's being radioactive explains its effect on the Geiger counter, then so does its emitting of ionizing radiation.

Or consider an example involving an extrinsic property. Suppose Euthyphro is right that being pious consists in being loved by the gods. It follows that if Odysseus causes the sword to glow because he is pious, then Odysseus causes the sword to glow because he is loved by the gods. If the one is a reason for Odysseus's causing the sword to glow, then so must be the other. Indeed, given the constitutive connection between the properties, it seems most apt to say that they are the *same* reason.

I take the principle to be obvious and not to stand in need of an extended defense. It is constantly, if tacitly, relied upon by Dretske and other naturalists. They *have* to assume it: it is implicit in the naturalist's (as well as the special scientist's) policy of taking constitutive claims as pivots upon which to shift the burden of causal explanation from a higher-level to a lower-level discourse.

Now, if **3** is false then so must be one of the two premises from which it was inferred. **1**, again, is what Dretske thinks we must affirm if we wish to reject content epiphenomenalism. And **2** encapsulates Dretske's account of content. Hence if **3** is false, Dretske cannot have what he seeks: a naturalistic account of content that allows for the rejection of content epiphenomenalism.

Is **3** true or false? It certainly looks peculiar. But perhaps it can nonetheless be motivated. Recall that we are accepting, for sake of argument, the first part of Dretske's account: his story about the role that informational properties of past tokens of a state play in structuring the causal relations of current tokens. We are granting, then:

> **4.** Current Cs cause behavior because past Cs indicated F.

3 and **4** are obviously related. There is a significant overlap in form: the forms of **4** and **3** respectively are "p because q" and "p because (p because q)". The resemblance might encourage us to suppose that the repetition of "p because" in **3** is semantically otiose, and thus that **3** asserts essentially the same thing as **4**. This appears to be Dretske's suggestion in the following passage (the passage in which he comes closest to acknowledging the difficulty I am now working to bring out):

> On [my] theory of meaning, the meaning of the current C consists, in part, in what past tokens of C indicated. If this C means F, then it is the function of C (tokens) to indicate F. If it is their function to indicate F, then (in the case of beliefs) past tokens of C must have successfully indicated F (that is how current tokens acquired their *function* of indicating F). On this account of meaning, meaning is a historical property. Later tokens of C mean F if earlier tokens of C indicated F and, in virtue of this fact, structured the causal relations of later tokens (including this token). Hence, anything explained by the fact that earlier tokens indicated F will be explained (albeit redundantly and indirectly) by the fact that current tokens *mean* F. To explain behavior by current meaning is just to explain what indicational facts (about earlier tokens) *were* relevant in recruiting the current belief (the current token of this type) for this kind of causal service (Dretske, 1990b, p. 831).[10]

Perhaps something in the neighborhood of the structural overlap we have noticed between **3** and **4** is behind Dretske's remark that "anything explained by the fact that earlier tokens indicated F will be explained (albeit *redundantly* and indirectly) by the fact that current tokens mean F" (my emphasis). If **3** is simply a redundant formulation of **4**, then presumably it is true if **4** is. Since we are granting **4**, we would have to grant **3** as well.

The problem with this reading is that **3** and **4** are not versions of the same claim: they offer *different* explanations of why current Cs cause the behavior they do. **3** simply cannot be viewed as a wordier version of the explanation offered by **4**.

Consider, by way of analogy, the following claims:

> **i.** Vera is upset because she was denied admission to the university.
> **ii.** Vera is upset because she's not a U.S. citizen.
> **iii.** Vera is upset because (she was denied admission to the university because she's not a U.S. citizen).[11]

These are three different explanations of why Vera is upset. None of them entail or presuppose any of the others. **iii** is consistent with the falsity of both **i** and **ii**. Vera may be content to be a non-U.S. citizen, relieved she was denied admission to the university (her parents forced her to apply),

[10] For an earlier, less explicit statement of this idea, see 1988, p. 84.

[11] We need the parentheses to distinguish **iii** from the different assertion, "(Vera is upset because she was denied admission to the university) because she's not a U.S. citizen."

and nonetheless be upset by the fact that she was denied admission *because* she's not a U.S. citizen. The source of her upset is not that she's not a U.S. citizen or that she was denied admission to the university, but that the one causally explains the other. Conversely, both **i** and **ii** are consistent with the falsity of **iii**. Vera may be upset because she was denied admission to the university, or upset because she's not a U.S. citizen, and *not* be upset because (she was denied admission to the university because she's not a U.S. citizen). She might think it perfectly appropriate to deny admission to people on the basis of their citizenship, and so be untroubled by the application of that policy to her case.

Now consider the following:

iv. Vera is upset because (Vera is upset because she's not a U.S. citizen).

iv, unlike **iii**, entails **ii**: **iv** says, in effect, that **ii** explains why Vera is upset, and **ii** could not do that unless it were true. Quite generally, "because" is factive: a sentence of the form "____ because ____" implies the truth of both of its constituent statements. But, and this is the important consideration for our purposes, **ii** does not entail **iv**. Vera could be upset because she's not a U.S. citizen and not be upset because (she's upset because she's a U.S. citizen). She may find her being upset because she's not a U.S. citizen an appropriate reaction to that circumstance, not in itself something upsetting. Thus **iv** makes a stronger claim than **ii**, and cannot be regarded as a redundant statement of it.

It is hard to see on what ground we could deny that what goes for **ii** and **iv** goes for any statement pair "p because q" and "p because (p because q)". And so it is hard to see why it should not go in particular for **4** and **3**. Granted, **3** implies **4**. But **4** does not imply **3**. And if **3** were in some other sense a "redundant" formulation of **4**, shouldn't **iv** equally be in that sense a redundant formulation of **ii**? Why not? But **iv** isn't a redundant formulation of **ii** in any conceivable sense. The proper conclusion is that **3**, like **iv**, must stand on its own two feet; each is a stronger claim than the other member of its pair (**4** and **ii** respectively), and would require an additional defense.

Once we acknowledge that **3** is not a redundant formulation of **4**, and so disperse whatever air of plausibility surrounds it in virtue of its being mistakenly so regarded, it emerges as a proposition of dubious coherence. **iv** is an intelligible claim because the two appearances within it of "Vera is upset" can be understood as describing two distinct facts or states of affairs. Assertions of the form 'S is e because p', where S is a person and 'is e' is a predicate for an emotional state ("is upset", "is angry", "is happy"), generally implicate not only that S is e because p but also that S is e *about* the fact that p—that S is e *that* p. Since mental states with intentional objects are partly individuated by those objects, the most natural reading of **iv** implies that Vera is in two distinct emotional states. Moreover, given the respective intentional objects of these states, it is quite comprehensible why the explanatory connection posited between them should obtain: we are all familiar with emotional states directed at, and explained by, other emotional states.

I submit that to the extent that we ever find ourselves endorsing statements of the form, 'p because (p because q)', which is not in any case very often, a comparable circumstance will prevail: the two appearances of the statement substituting for "p" will describe two different states of affairs. It is not easy to think of non-intentional examples, but suppose that the process of a certain atmospheric condition's x causing rain is itself a condition likely to cause more rain later on. Then perhaps we may be led to say on some occasion that it rained because (it rained because condition x obtained). The comment makes sense given the context, because we are talking about two different intervals of rain.

But there is no interpretation available along these lines to secure the cogency of **3**. In both cases we are talking about the same thing: the fact that current Cs cause behavior. **3** maintains that *current* Cs cause behavior because (*current* Cs cause behavior because past Cs indicated F). That's like saying that it's raining here and now because (it's raining here and now because condition x obtained). In both cases, what is asserted is first, that fact A explains fact B, and second, that the fact that fact A explains fact B itself explains fact B. I think it is clear that we never make such claims in actual life, never in fact advance such explanations, and correlatively, have no notion of what the basis or significance of a given claim of that form could be.

At the very least, then, we need some positive account of why we should accept an explanation like **3**. Other than a hint of the discredited suggestion that **3** is a redundant formulation of **4**, nothing of the sort is to be found in Dretske. Lacking such an account, and with excellent reason to doubt even the intelligibility of **3**, we should reject it. But then either **1** or **2** is false, for they jointly imply **3**. In particular, if **2** is true, then **1** must be false. This is the conclusion I advertised at the beginning of the previous section: Dretske's account of content yields content epiphenomenalism.

I close this section by offering a partial diagnosis: Dretske's failure to perceive this problem is facilitated by an ambiguity sometimes present in his writings. Although the distinction between content properties and informational properties is crucial for Dretske's solution to the threat of content epiphenomenalism, there are places in his writings where it is not clear which kind of property he is speaking of. This ambiguity is encouraged by the Gricean terminology of natural meaning (information) and non-natural meaning (content). And it may in turn encourage overlooking the difference I have tried to bring out in contrasting **3** and **4**.

I will mention what I take to be the most interesting instance of this ambiguity—most interesting because it bears on a long-standing debate about 'biosemantics', the view that a state's or structure's content is fixed by its evolutionary-biological function. In the course of a multi-decade discussion with some of his fellow naturalists (e.g., Dennett and Millikan), Dretske has been unwavering in his insistence that the functions in virtue of which internal items possess their content must be learning-theoretic, not evolutionary-biological. Dretske does not deny that natural selection has a hand in explaining why a creature's brain is configured as it is. He even allows that informational properties of internal states and structures may help explain why these states and structures were selected for certain causal roles. But as he sees it, the crucial sticking point for biosemantics is that such selectional explanations will not show that the informational properties of structures in a given organism are relevant to the explanation of *that* organism's behavior. So far as natural selection goes, it will be the informational properties of structures of that type in *ancestors* of a given organism that will explain why the organism in question has a nervous system in which that structure has a particular causal role. And so, Dretske argues, processes of natural selection do not provide the material for a naturalistic solution to the threat of content epiphenomenalism:

> This, then, is *not* a case where the meaning of an organism's internal states makes a difference, a causal difference, to what *that* organism does. We have, to be sure, the extrinsic properties of structures (the information they carry) making a difference in the world. They do not, however, make the *right* difference to qualify as a belief. Nothing, on this account of things, turns out to be a semantic engine, something whose *own* behavior is driven by the meanings of *its* internal states (Dretske, 1991, p. 207).

What sense of "meaning" does Dretske intend in this passage? If he is speaking of natural meaning, then we may grant that natural selection is not a process in which the meanings of a given organism's states have a hand in determining that organism's behavior. We may also grant that instrumental learning *is* such a process. But the kind of meaning that bears on the question whether an organism can be said to be a 'semantic engine' is non-natural meaning—content. Since Dretske denies the information-semantic premise that an item's possession of non-natural meaning consists in its possession of natural meaning, showing that the natural meaning of an organism's internal states bears causally on what it does is not on its own sufficient for showing that the organism is a semantic engine, an engine driven by content.

If, on the other hand, Dretske is speaking in this passage of non-natural meaning, then he is certainly correct that *given his account of content, which ties facts about content to facts about instrumental learning and not to facts about natural selection,* natural selection is not a process whereby meaning makes a difference to what the organism does. But this begs the question. It is equally true that, had Dretske adopted a view by which processes of natural selection determine content, then (non-natural) meaning would be exactly as relevant to the explanation of the aspects of behavior shaped by evolution as it is, on his actual view, to the aspects of behavior shaped by learning.

In the quoted passage (and elsewhere), Dretske fails to distinguish between past information and present content. Such a move goes hand in hand with eliding the difference I have been trying to bring out.

4. The explanatory role of content: indirect?

Consider again Dretske's remark that "anything explained by the fact that earlier tokens indicated F will be explained (albeit redundantly and indirectly) by the fact that current tokens *mean* F". We have been putting weight on the first of the two parenthetical qualifications to "explained". But it might seem that we get a different read on Dretske's remark if we emphasize rather the second. And it might seem that we can then dissociate Dretske from the view that sentences like **3** are redundant formulations of sentences like **4**, a view we have seen not to withstand scrutiny.

Focusing on the second parenthetical qualification and filtering out the first, Dretske's remark comes to this: whatever is explained by the fact that earlier Cs indicated F is "indirectly" explained by the fact that present Cs have the content *F*. Since Dretske does not elaborate on the significance of this qualification, any interpretation of the meaning of this claim will have to be speculative. But perhaps the idea is something like this. First of all, the content of an inner item does not have explanatory relevance, in the sense identified above, to that item's causing behavior. One cannot say of a current C that it causes behavior *because* it has the content that it does. But although ascriptions of content do not themselves figure in 'because'-statements explaining inner items' causal relations, they put us in a position to deduce pertinent 'because'-statements, statements that mention not content but information. As we might put the idea, ascriptions of content *encode* explanatory relationships between facts about the informational properties of past tokens of a state and facts about the causal doings of present tokens of that state. If we know the code, we can retrieve the relationships.

Here is one way of elaborating this idea:

> Where E_1 is an event that causes an effect E_2 and p is a property of E_1, E_1's having p *indirectly explains* E_1's causing E_2 iff a) there is some fact or state of affairs S such that a) E_1 causes E_2 because S obtains and b) the fact that E_1 has p, coupled with the right account of what it is for an event to have p, implies that E_1 causes E_2 because S obtains.

It is undeniable that Dretske's account of content entails that content "indirectly explains" behavior in just this sense. According to that account, that current Cs have the content *F* implies the existence of a state of affairs S such that current Cs cause behavior because S obtains. (The state of affairs in question is that past Cs indicated F.) The suggestion we are now taking from Dretske's remark is that content bears on inner items' causing behavior *only* in this sense—that in particular, it does not also have explanatory relevance as that was defined above. On this interpretation, then, Dretske is not committed to **1**, and hence not to **3**. The question whether **3** can be regarded as a redundant formulation of **4** or can be motivated in some other way is rendered moot. Thus we seem to have managed to find an interpretation of Dretske's account of the explanatory role of content that avoids the objection made in the last section.

Unfortunately, this interpretation is untenable. It is at odds with Dretske's own explicit remarks about what is required to defeat content epiphenomenalism. Moreover, Dretske's remarks on this score are quite correct: a satisfactory repudiation of content epiphenomenalism must show that content can be explanatorily relevant, in the sense identified earlier, to the production of behavior. It does not suffice to bequeath to content the "indirect" role just delineated.

With regard to the first point, we have already seen that Dretske portrays his goal as that of showing that content has explanatory relevance, in our sense, to the causal explanation of behavior. The crucial question, says Dretske, is whether a person ever does something *because* he has a thought with a particular content. If it is never the case that a person performs "A [an action], in part at least, *because* he thinks T, why, in our philosophical (not to mention scientific) efforts to understand why people do what they do, should we be interested in *what* (or even *that*) they think?" The interpretation currently on the table thus cannot be squared with Dretske's pronouncements on what an adequate refutation of content epiphenomenalism demands.

Moreover, Dretske is right that a satisfactory repudiation of content epiphenomenalism must make room for the explanatory relevance, in our sense, of content. Here is where the difference in kind between the current notion of explanatory relevance and the aforementioned counterfactual accounts of causal relevance is crucial. The latter are substantive, tendentious pieces of philosophical analysis. But all that is meant here in saying that an item's content has explanatory relevance to the item's causing behavior is that the item causes behavior *because* it has the content that it does. And all *this* means is that the item's having that content *explains* its causing behavior. To say "E_1 caused E_2 because E_1 is p," is to say no more or less than "E_1's being p explains E_1's causing E_2." It is not to endorse or imply any particular account of what this explanatory connection comes to. It is simply to assert its existence. (Note that asserting "E_1 caused E_2 because E_1 is p," does not even rule out the possibility of additional explanations of E_1's causing E_2. That x explains y doesn't preclude z's explaining y as well.) Hence to deny that an item causes behavior because it has the content that it does is to deny that its having that content explains why it causes behavior.

Recall again the claim we extracted from Dretske's passage: "anything explained by the fact that earlier tokens indicated F will be explained…indirectly…by the fact that current tokens *mean* F." We should press this question: is "indirect" explanation a form of explanation? Does the claim, "C's having content *F* indirectly explains C's causing M," imply, "C's having content *F* explains C's causing M"? If the answer is yes, then the claim also implies, "C causes M because C has content *F*." But then C's having content *F* is said to have explanatory relevance, in the current sense, to C's causing what it does, and Dretske's view is saddled with the problematic implication identified in the previous section. But if the answer is no, then the fact that C's having content *F* "indirectly explains" C's causing M is of no interest to us or to Dretske. Our interest is in the place of content in causal explanations of behavior. The question of content epiphenomenalism is the question of whether content causally explains.

5. Conclusion

Dretske offers an account of content intended to meet two criteria: 1) that it portray an inner item's possession of content as consisting in naturalistically specifiable properties of the item, and 2) that it be consistent with, indeed make a positive case for, the denial of content epiphenomenalism. Dretske begins by working to establish a role in the explanation of behavior for informational properties. The explanatorily relevant informational properties turn out not to be properties of the inner tokens whose effect on behavior we are seeking to explain: they are properties of tokens that obtained at an earlier time. Nonetheless, reflection on the role that these properties play in explaining a current token's causal relations simultaneously yields an account of the relevance of that token's content to the explanation of its causal relations. All we need to do is to take a current token's possession of content to be constituted by the existence of the relevant historical, information-based explanation of the token's causal relations.

This proposal is ingenious. But it doesn't work. This was made clear by our tracing through the implications, given Dretske's account, of placing ascriptions of content in 'because'-statements purporting to causally explain behavior. The upshot was putative explanations of behavior of a reflexive, second-order form that we were unable to make sense of, let alone find any reason for accepting. Nor was there any prospect of distinguishing the question of content epiphenomenalism from the question of whether ascriptions of content can figure in 'because'-statements that express causal explanations of behavior.

I want to close by indicating two directions in which I believe the results of this paper generalize. First, there is a long-standing debate in the philosophy of mind about whether functional properties of inner items (in the functionalist's sense) are explanatorily relevant to those items' causal relations. I believe that the right case for a negative answer to this question, one that avoids the mistakes of those extant in the literature, can be constructed along lines similar to the argument presented here.

The second direction of generalization concerns the naturalistic project as such. As we noted at the beginning, Dretske believes that tackling the challenges of content epiphenomenalism and naturalism together will serve to bring into view new insights about both. The current discussion confirms Dretske's belief. In particular, the results reached here are evidence that that the naturalistic approach makes the challenge of content epiphenomenalism more difficult to surmount.

The aim of non-eliminativist naturalism of the sort Dretske espouses is not to supersede but to vindicate our ordinary 'folk' psychology of the mind. The aim is to show that the complex contours of the concepts of folk psychology—of belief, desire, intentional content and so forth—can be redrawn using materials taken only from the natural sciences. It seems obvious on the face of it that this is an extremely difficult task, for a satisfying naturalization will have to reproduce all the features of mental phenomena that are crucial to our folk understanding of them. Conventional naturalist methodology works to obscure this difficulty, for discussions of naturalistic views tend to focus on the question of extensional adequacy. In the case of naturalistic accounts of content, the debate between proponents and critics of a given account is usually given over to the question of whether that account assigns the right contents to states and occurrences. The apparent assumption is that so long as this criterion is met, the rest will take care of itself.

As we have seen, Dretske's attitude is less complacent. From the get go, he has his sights set not just on extensional adequacy but on another fundamental feature of content: the role it plays in our everyday causal explanations of thought and behavior. It is thus suggestive of the severe difficulties faced in the quest for a less superficial naturalism that the account at which Dretske arrives renders content explanatorily inert.[12]

University of Chicago

References

Cummins, Robert. (1990). "The Role of Mental Meaning in Psychological Explanation." In B. McLaughlin (ed.), pp. 102-117.

Davidson, Donald. (1963/1980). "Actions, Reasons and Causes." In his *Essays on Actions and Events.* Oxford: Oxford University Press, pp. 3-19.

Dretske, Fred. (1981/1999). *Knowledge and the Flow of Information.* CSLI Publications.

Dretske, Fred. (1988). *Explaining Behavior.* Cambridge: MIT Press.

Dretske, Fred. (1990a). "Does Meaning Matter?" in Villanueva (ed.), pp. 5-15.

Dretske, Fred. (1990b). "Replies to Reviewers." *Philosophy and Phenomenological Research* 50, pp. 819-39.

Dretske, Fred. (1991). "Dretske's Replies." In B. McLaughlin (ed.), pp. 180-221.

Dretske, Fred. (1993). "Mental Events as Structuring Causes of Behavior." In *Mental Causation*, J. Heil and A. Mele (eds.). Oxford: Oxford University Press, pp. 121-136.

Dretske, Fred. (1994/2000). "If You Can't Make One, You Don't Know How It Works." In Dretske, 2000c, pp. 208-226.

Dretske, Fred. (2000a). "Minds, Machines and Money: What Really Explains Behavior." In Dretske, 2000c, pp. 259-274.

[12] In "Does Informational Semantics Commit Euthyphro's Fallacy?", in *Noûs*, 40:3 (2006), 522-547, I argue that informational semantics also fails to accommodate the role of content in everyday psychological explanation.

Dretske, Fred. (2000b). "Norms, History and the Constitution of the Mental." In Dretske, 2000c, pp. 242-258.

Dretske, Fred. (2000c). *Perception, Knowledge and Belief.* Cambridge: Cambridge University Press.

Fodor, Jerry. (1990). "Reply to Dretske's "Does Meaning Matter?"" In Villanueva (ed.), pp.28-35.

Lepore, Ernest and Barry Loewer. (1987). "Mind Matters." *Journal of Philosophy* 84, pp. 630-642.

McLaughlin, Brian, ed. (1991). *Dretske and his Critics.* Oxford: Basil Blackwell.

Schiffer, Stephen. (1991). "Ceteris Paribus Laws." *Mind* 100, pp. 1-17.

Villanueva, Enrique, ed. (1990). *Information, Semantics & Epistemology.* Oxford: Basil Blackwell.

Yablo, Stephen. (2003). "Causal Relevance." *Philosophical Issues* 13, pp. 316-328.