Philosophy of Mind

Identity Conditions for Indicator State Types within Dretske's Theory of Psychological Content Naturalization

Mary Litch
University of Alabama-Birmingham
litch@uab.edu

ABSTRACT: Within the context of Dretske's theory of psychological content naturalization, as laid out in *Explaining Behavior*, the concept of an indicator state type plays a pivotal role. Providing a general (and non-circular) description of the identity conditions for being a token of an indicator state type is a prerequisite for the ultimate success of Dretske's theory. However, Dretske fails to address this topic. Thus, his theory is incomplete. Several different approaches for specifying these identity conditions are possible; however, each is inadequate.

Of the various theories for psychological content naturalization put forward within the past two decades, I believe that a Dretske-style approach that explains the content of a mental state in terms of the causal history of past tokens of that state holds out the most promise of giving us a workable theory describing the role that content plays in learned behavior. While I favor this general approach, the particular theory laid out by Dretske in *Explaining Behavior* has a shortcoming that must be addressed before his theory can be applied to real systems: Dretske fails to provide an analysis of identity conditions for being a token of an indicator state type. The shortcoming is serious because of the critical role that past tokens of an indicator type play in fixing the content of a current token of the indicator type — without identity conditions, there is no way to specify which previously tokened states among the many that have been instantiated during the learning period of the organism are of that indicator type.

I begin with a very brief review of Dretske's theory from *Explaining Behavior*. Some organisms possess indicator states (i.e., internal states that indicate whether some external conditions hold). For example, organism O may token an instance of I (the internal indicator state type) whenever external conditions F obtain. Prior to learning, I indicates F does not mean F. Let's suppose that external conditions F are relevant in some manner to O's continued functioning, perhaps because environments in which F obtains are environments that are relatively inhospitable for O. Let's also suppose that O is capable of learning using reinforcement information (via operant conditioning), such that future tokenings of I come to cause movements that are appropriate to conditions F. (My use the evaluative term "appropriate" here rests on two assumptions: (1) that the learned pattern of movements given I tends to maximize the pleasure and/or minimize the pain experienced by O, and (2) that the prior evolutionary history of O's species has resulted in O being

hardwired such that there is a strong correlation between experienced pain (by O) with inhospitableness and experienced pleasure with hospitableness of the environment.) After this learning process, I no longer just indicates F, it also means F. Note that the representational content of I is learned Type III (i.e., the representational content has no conventional aspect).

One advantage of Dretske's version of content acquisition is that it explains how content can be efficacious as a structuring cause. Even though, as triggering cause, the content of a particular psychological state is explanatorily irrelevant, still it was what the indicator state indicated about external conditions in the past that played a role in the restructuring of an organism's nervous system, such that this (current) token of that psychological state type produces this particular movement. It is generally assumed that a necessary condition for causal efficacy is counterfactual relevance. We can see that, on Dretske's account, this condition is satisfied, for the content does make a difference in the movements of O. Had a token of I meant something else (by virtue of indicating some other external conditions), then the reinforcement information O received would (likely) have been different, thus resulting in O's having a different post-learning configuration in accordance with which tokens of I would (likely) fail to cause the movements that Is came to cause in the actual world.

As mentioned above, I believe that Dretske's theory for the naturalization of content holds the most promise of providing an analysis of content acquisition resulting from learning. As it stands, however, there is an important shortcoming in Dretske's theory: the lack of clear identity conditions for the indicator state types. In the remainder of this paper, I shall examine this shortcoming, making explicit its relevance for the overall success of Dretske's program. I then examine several ways in which a proponent of Dretske's theory may attempt to address the shortcoming, and argue that each of these is inadequate.

Recall that, according to Dretske's theory, the content of a (just-tokened) post-learning mental state depends upon the environmental conditions that obtained when previous tokens of this same type occurred — these environmental conditions played a role in the reinforcement received by the organism, resulting ultimately in physical changes to the organism such that now tokens of that indicator state type produce bodily movements appropriate to the sensed environment. Given the important place that past instances of an indicator state type have in Dretske's theory of content, it is surprising that he fails to mention how these state types are to be picked out. One possible explanation for this is his blurring of the type/token distinction throughout *Explaining Behavior*. A more likely explanation is that he is supposing something like this: tokens of I are physical states, so the most straightforward identity condition for I-hood is being a token of a particular physical state type. (Presumably, we do not need to go all the way down to basic physics to specify the state type for all instances of I; rather, one assumes that the state type can be described in the vocabulary of neurophysiology.) The trouble with this identity condition is that there is no reason to believe that exactly the same internal (neuro/)physical state type will be tokened reliably whenever the external conditions F obtain. Dretske's theory might not demand flawless tokening of I when F, but his theory requires at least many occasions of a tokening of I when F during learning. However, neurological research into learning in real central nervous systems suggests that even this weak requirement will not, in the general case, be satisfied. If sharing the same (neuro/)physical state type is supposed to "link together" the tokens of I during the learning period with post-learning Is, such that the content of the latter is partially grounded in the causal etiologies of the former, then Dretske's theory is in serious trouble.

Perhaps my dismissal of (neuro/)physical state type as identity condition was too hasty, for there is a way of describing I that ensures at least that identical sensed conditions result in

the tokening of the same (neuro/)physical state type. If one specifies that I occurs on the periphery of O's central nervous system (put crudely, I is the unprocessed raw sense data being fed into the organism's brain), then one is guaranteed: same sensed external conditions, same (neuro/) physical state type tokened.

There are, though, two important reasons to reject this version of the identity conditions for I-hood. (I combine these two reasons under the banner the "too specific" objection.) The first reason to reject it is that it makes representational content too specific. If psychological generalizations are to be possible, then the representational content that emerges after the learning process will have to be at least moderately general. Consider how severely limited possible learned content would be under this assumption: the sensed input (or, at least that part of the sensed input that constitutes the indicator state) would have to be identical in order for the same (neuro/)physical state to emerge from one exposure to the next. Any slight change would cause a physically different state to be tokened — a state that, by definition, was not the indicator state in question. The resulting content would likewise be very specific, in order to circumscribe the conditions F that I indicates, for conditions F would not just involve some perspective-free external state-of-affairs, but also lots of other (very specific) conditions about the spatial relation of O to objects mentioned in F, and (depending on how I is specified) perhaps many other irrelevant ambient conditions. Dretske cannot non-circularly say "while the actual state- of-affairs that obtains whenever an I is tokened is severely circumscribed, still the representational content of I is the general state-of-affairs, G, because it is the function of I to indicate G". He cannot do this because the function of I is cashed out solely in terms of the learning history of O; in particular, in terms of the external state-of-affairs F that led to I's being tokened and that played a role in the reinforcement signal received by O (and the concomitant changes in O's internal structure). We, as external observers of O during the learning process, may assign to I the (Type II) representational content of G; but, to O, I comes to represent the very specific state-of-affairs F.

I want to reiterate that it is inappropriate for us, as outside observers of the system, to stipulate the content of the indicator state, and, with a wave of our hand, generalize the content by removing the context-specificity. If the end result of the learning process is a Type III representational state, then there is a non-observer-relative fact-of-the-matter with respect to the content of that state. Perhaps an example from the literature on learning in connectionist systems will help to bring this point home. Oftentimes failed experiments are more instructive than successful ones. A case in point is the experiment during the mid-1960s involving the "optical perceptron" reported in Christiansen and Chater(92). The purpose of the experiment was to determine whether a connectionist net was capable of learning the concept picture-with-tank. The training regimen consisted of providing encoded pictures as input, some of which were scenes containing tanks, and some of which weren't. The net was trained to output one value when the input encoded a tank-scene, and to output a different value when the input encoded a scene without a tank. Training went well — the net not only learned to correctly discriminate between tank-scenes and non-tank-scenes for all of the scenes presented during the training period, but it could also correctly discriminate on a few novel scenes taken at the same time as the pictures used during training. Then, the researchers photographed some more tank-scenes and non-tank-scenes. They were surprised to find that their trained net failed miserably at discriminating between the two scene types. On closer examination of the features generalized by the network, they found that it was really discriminating between high-intensity light scenes and low-intensity light scenes. On re-checking the original photographs used in training, they discovered that the tank pictures were all shot during the same time period of bright sunlight, and the non-tank pictures were all shot during the same time period of less-bright sunlight. The moral here is two-fold: (1) there is a matter-of-fact (independent of external

observers) with respect to the content of a Type III representational state, and (2) to assume that one can stipulate away context-specificity is wholly unjustified.

There is another reason to reject the identification of an indicator state type with a (neuro/)physical state type on the periphery of O's central nervous system. This is the second part of the "too specific" objection. Even if one is willing to allow highly specific contents as the end-product of learning, there are still problems for Dretske's theory. His theory assumes that O will have tokened Is many times in the past, and that (as a result) O has enough reinforcement information to place I in the causal pathway leading to some movement appropriate to conditions F. But if I is only tokened a few times (or maybe only one time) during O's learning period, O will not have enough information to place I correctly, as required. The extreme specificity of I ensures that it will rarely be tokened, because it is only rarely that an organism receives exactly the same sensed data. This is a twist on the standard "poverty of the stimulus" argument.

Thus, the identification of an indicator state type with an internal (i.e., non-peripheral) (neuro/)physical state type flies in the face of recent results from neuroscience, while the identification of an indicator state type with a peripheral (neuro/)physical state type is subject to the "too specific" objection.

What other options are open to a proponent of Dretske's theory for psychological content naturalization? The first approach one might take is to bite the bullet on specificity, but to make it less objectionable by construing specificity as a form of context-relativity. Anyone who buys a Dretske-style explanation of content acquisition is already willing to have contents be context-relative — maybe this specificity is just another aspect of context-relativity. So, the extreme specificity of content is transformed from a flaw of the causal-historical view of content into a feature. This tack shows a misunderstanding of what context-relative means as an aspect of any causal-historical account of content. If the contents of my mental states are in part a function of my past learning history, then those contents will be relative to certain contingent facts about what states-of-affairs I found myself in and my particular reinforcement schedule. So much is undeniable and not in itself a drawback of the theory of content according to which content is relative to the causal-historical contexts of the agent. The "too specific" objection is not an objection against this aspect of Dretske's theory, but rather an objection against over-specificity of the learning contexts. On this account, the content of my mental states would be relative not only to the particular objects with which I had interacted, but also to my exact location relative to those objects, the ambient lighting conditions present at the time I saw the objects (for cases in which the indicator state has a visual component), etc. This is not good old context-relativity in the benign sense, and does not constitute a bullet that one should be too willing to bite, for its acceptance implies the illegitimacy of the view of psychology as the search for (at least weakly) universalizable generalizations.

A second approach that one might take to circumvent the "too specific" objection is to change the identity conditions for I-hood to membership in a set of (neuro/)physical state types. The problem with this approach is explaining how being a member of a set can make a difference, over and above being a collection of individual instantiations of (neuro/)physical state types. It will still be the case that each of the member state types in the set will only rarely be instantiated. Without some way of showing how being a member of this set makes a causal difference to the learned behavior (such that the ultimate content acquired by tokens of I can legitimately be adverted to in psychological explanations), it is unclear how this approach could avoid the "too specific" objection.

Maybe, one might reply, there is some causally relevant property that binds together the various (neuro/)physical state types in the set (and that could in turn serve to bind together the tokens of those state types) — namely, being tokens that indicate F. The main problem

with using indicates F as the criterion for inclusion of a (neuro/)physical state type in the set that constitutes the indicator is circularity. As this is also the main problem with the third failed attempt discussed below, I postpone a closer examination of it until after I offer the third attempt at thwarting the "too specific" objection.

All of the individual tokens of I have one feature in common: they all indicate that external conditions F obtain. Maybe this is the identity condition for I-hood that we should adopt: a state is an instance of the indicator type if it indicates F. In its favor, this does circumvent the "too specific" objection, but at a huge cost. The theory of content that results is circular: it attempts to explain how a physical state (picked out by virtue of the fact that that state indicates F) comes to means F. The advantage of describing I-hood in terms of a state's (neuro/)physical state type is that it does not presuppose what it is attempting to explain: it does not presuppose that there is something (prior to learning) about the state that refers to some external state-of-affairs. At least (neuro/)physical state type descriptions do not advert to external conditions.

What we need is some way of picking out instances of I that is syntactic (i.e., some way of picking out instances of I that can, in theory, be cashed out in a non-extrinsic physicalist vocabulary), and that can be used to provide a causal explanation for why current tokens of I produce the movements that they produce. Identifying I-hood by its indicated external conditions fails to satisfy the first criterion above, whereas identifying I-hood by membership in a set of (syntactically-described) (neuro/)physical state types fails to satisfy the second. We, as observers of O, are also not allowed to bridge the gap between set membership and causal relevance by a judicious choice of the (neuro/)physical state types that make up the set. We are assuming that O will eventually be a Type III representational system requiring no outside agents to ground the content of its representational states.

So, where does this leave us (and Dretske)? The lack of an analysis of the indentity conditions for indicator state types is a serious shortcoming for Dretske's theory. Of the three attempted analyses I considered here, all are seen to be inadequate. Obviously, this does not show that an adequate analysis is impossible. (Indeed, I have argued elsewhere that, with the aid of the connectionist model of operant conditioning, an adequate analysis can be offered). However, the above argument at least demonstrates that Dretske's theory as it now stands is incomplete.

References

Baker, L. 1991. "Dretske on the Explanatory Role of Belief", Philosophical Studies 63, pp. 99-111.

Christiansen, M. and Chater, N. 1992. "Connectionism, Learning and Meaning", *Connection Science* 4, Nos. 3-4, pp. 227-252.

Dretske, F. 1988. *Explaining Behavior*, Cambridge, Mass: MIT Press.

Dretske, F. 1993. "Mental Events as Structuring Causes of Behavior" in *Heil and Mele* (93), pp. 121-136.

Fodor, J. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*, Cambridge, Mass: MIT Press.

Fodor, J. 1990. *A Theory of Content*, Cambridge, Mass: MIT Press.

Heil, J. and Mele, A. (editors) 1993. *Mental Causation*, Oxford: Oxford University Press.

Kim, J. 1991. "Dretske on How Reasons Explain Behavior" in *McLaughlin* (91), pp. 52-72.

Llinas, R. and Churchland, P. S. (editors) 1996. *The Mind-Body Continuum: Sensory Processes*, Cambridge, Mass: MIT Press.

McLaughlin, B. (editor) 1991. *Dretske and His Critics*, Oxford: Blackwell.

Merzenich, M. and deCharms, R. 1996. "Neural Representations, Experience, and Change" in *Llinas and Churchland* (96).

Millikan, R. 1984. *Language, Thought and Other Biological Categories*, Cambridge, Mass: MIT Press.

Poland, J. 1994. *Physicalism: The Philosophical Foundations*, Oxford: Clarendon Press.

Ramsey, W. 1997. "Do Connectionist Representations Earn Their Explanatory Keep?" *Mind and Machine*, Vol. 12, No. 1, pp. 34-66.