The International Journal of Comparative Psychology, Vol. 2, No. 3, Spring 1989

PROSPECTS FOR ANIMAL MODELS OF MENTAL REPRESENTATION

Shawn Lockery and Stephen Stich
University of California, San Diego

ABSTRACT: A major goal of physiological psychology is to determine the physical basis of mental representation. Animal models are essential to this project. Dretske's influential analysis of the concept of mental representation suggests that operant and classical conditioning involve mental representation. This analysis comports well with known physiological mechanisms of conditioning, but fails to capture necessary features of mental representation at the human level. We conclude that the applicability of animal models to the problem to human mental representation is more restricted than previously thought.

What is the relationship between the study of animal psychology and the study of human psychology? Staddon (1988) reminds us that there is a long tradition that answers this question by appealing to the idea that in various psychological domains, animals can be used as models for people. And, of course, in this tradition it is human psychology that is ultimately of interest. Staddon is not comfortable with this tradition. He argues that it has been a baleful influence on the study of animal psychology while yielding relatively little insight into human psychology. The alternative that Staddon recommends is that "psychology as a basic science should be about intelligent and adaptive behavior, wherever it is to be found," and thus that animals should be "studied in their own right for what they can teach us about the nature and evolution of intelligence, and not as surrogate people, or tools for the solution of human problems." This is a view we wholeheartedly endorse. Much of what we say in this paper can be read as illustrating Staddon's theme.

The domain in which we propose to develop our illustration is the rich literature on intentionality and mental representation. Our thesis will be that much recent theorizing about the concept of mental representation and its role in psychological explanation has been led astray by its anthropocentric focus. Most theorists simply presuppose that mental representation is a single phenomenon, and that the paradigm case of the phenomenon is to be found in conscious human thought. Typically, the psychological processes of animals are seen as relevant to the understanding of intentionality only in so far as they provide simple models for

Address all correspondence to Shawn Lockery, The Salk Institute for Biological Studies, Box 85800, La Jolla, CA 92138.

full-blown human intentionality. It is our contention that this anthropocentric approach is in two ways unfortunate: first, it yields relatively little insight into human intentionality, and second, it obscures the fact that there may be very different kinds of semantic or intentional phenomena and thus that significantly different notions of representation may all have a substantive role to play in the explanation of intelligent and adaptive behavior. The specific target of our critique will be the elegant and sophisticated account of mental representation developed by Fred Dretske (1988) in his recent book, *Explaining Behavior.*

This paper is divided into three parts, the first of which is largely expository. In it we sketch Dretske's account of mental representation, and the role that it plays in psychological explanation. Dretske's story is developed against the background of some abstract and very schematic assumptions about what goes on in certain sorts of conditioned learning. So a natural question to ask is how well Dretske's cartoon of conditioning comports with what is known about the underlying neurobiology. This is the project pursued in the second section. The answer is that despite its simplifications, Dretske's sketch of learning meshes quite well with the emerging neurobiological details. In the third section, our question will be how useful Dretske's account of mental representation is likely to be for the understanding of the paradigm cases of human intentionality. Our contention is that Dretske's account is a poor model for the sort of intentionality and intentional explanation that loom large in human psychology. On the brighter side, however, we argue that Dretske has isolated a semantic or representational notion that can be of use in animal psychology. If this is right, then it is to be welcomed on its own merits, not disparaged because it fails as a model of human intentionality.

AN OVERVIEW OF DRETSKE'S PROJECT

The subtitle of Dretske's book is *Reasons in a World of Causes*, and that subtitle provides a convenient jumping off point for our description of Dretske's project. Following tradition, Dretske begins with a discussion of human behavior and its explanation. Fred, who is sitting in the living room, gets up and walks into the kitchen. How can this behavior be explained? Common sense psychology often provides a ready answer. Fred walked into the kitchen because he wanted a drink, and he believed he could get one there. This explanation, which provides Fred's reasons for walking into the kitchen, invokes a pair of intentional or representational states. His belief represents the world as being in a certain way. (Had Fred believed, instead, that there was nothing to drink in the kitchen, he would have remained in the living room, or gone to search

elsewhere). And his desire has as its object some future state—Fred getting a drink—which may, or may not, ultimately come to pass.

But it is also the case that when Fred walked into the kitchen, he did so because various muscles contracted; they, in turn, were "responding to a volley of electrical impulses emanating from the central nervous system." (p. ix). Walking into the kitchen was something Fred's body did, and ultimately we expect that the movements of his body will be explained by neuroscience and biology. If this is right, Dretske notes, then "one seems driven, inevitably, to the conclusion that, in the final analysis, it will be biology rather than psychology that explains why we do the things we do" (1988, p. x). "What, then," Dretske asks,

> . . . remains of my conviction that I already know, and do not have to wait for scientists to tell me, why I went to the kitchen? I went there to get a drink, because I was thirsty, and because I thought there was still a beer left in the fridge. However good biologists might be, or become, in telling me what makes my limbs move the way they do, I remain the expert on what makes me move the way I do. Or so it must surely seem to most of us. To give up this authority, an authority about why we do the things we do, is to relinquish a conception of ourselves as human agents. This is something that we human agents will not soon give up (p. x).

The ultimate goal of Dretske's project is to preserve a legitimate role for intentional psychology by showing how two schemes for explaining behavior—the intentional (or "psychological," as Dretske sometimes says) and the neurobiological—can co-exist. He wants to find an explanatory role for reasons in a world of causes.

A central element in Dretske's account of the explanatory role of reasons is his distinction between triggering causes and structuring causes. Before explaining that distinction, we need to think a bit about the relation between behavior and bodily movement. Most behavior involves bodily movement: Fred reaches for a beer; Bonnie turns the steering wheel of her car to avoid hitting a child; a rat depresses a lever with its paw when the light in its cage goes on in order to get a bit of food. However, not all bodily movement will plausibly count as behavior. When Clyde pushes Bonnie's arm, or I move the rat's paw, the resulting bodily movements are not part of Bonnie's behavior or the rat's. The distinction, Dretske urges, turns on the location of the cause of the movement. In genuine cases of behavior, a salient aspect of the cause of the movement is some event or process internal to the organism. By contrast, when I move the rat's paw the salient causes are external to the rat. In the terminology Dretske recommends, the term "behavior" is reserved for a process in which some internal state or event causes a bodily movement —schematically, a behavior is a process in which an internal state C causes a movement M. The movement itself is the visible product of the behavior; it is the output of the process.

Now when behavior is construed as a process, the question of why a particular piece of behavior occurred can be construed in two very different ways. On one reading it is a request for some explanation of what began the process and kept it going. This is what Dretske calls a triggering explanation. In the case of the rat pressing the lever, the explanation would begin with the light. It would detail how the light brings about some internal perceptual state C, and how C leads to the production of certain movements of the rat's paw. Ultimately, one would hope to be able to tell the entire story, from retinal stimulation to paw movement, at the neurobiological level. However, there is another aspect of the rat's behavior that needs to be explained. In addition to asking how a certain stimulus brings about an internal perceptual state, and how that internal state leads to certain movements, we can ask why the organism is so structured that the internal state leads to those movements rather than others. Why is C connected to M, rather than to M'? In asking this question, we are seeking what Dretske calls a structuring explanation, rather than a triggering explanation.

Since the distinction between triggering and structuring explanations is central to Dretske's account, we'd do well to consider another case in which the contrast emerges quite vividly. Suppose we are sitting at a friend's home and, as the day gets warmer, suddenly his automatic garage door motor is turned on, and the garage door opens. Why did this happen? Well, the rising temperature bent the bimetallic strip in the thermostat on the wall. That closed a circuit, which enabled current to flow to the garage door motor. (Note the pattern: an environmental E caused an internal C which caused the movement M.) The temperature triggered the opening of the door, and with a bit of effort we could tell an elaborately detailed story about the physical processes that subserve each of these causal connections. Having been told all of this, however, there is still something important that needs to be explained. Why is the thermostat connected to the garage door? Why isn't it connected to the furnace, or the air conditioner, or the doorbell, for that matter? Here, of course, all sorts of answers are possible. Perhaps our friend has found that opening the garage door cools off the house. Perhaps it is a way of letting the dog out in warm weather. Perhaps the thermostat was wired to the garage door as a joke. Whatever the explanation, it will be very different from the triggering explanation, because what we want explained is not the mechanism by which an environmental event leads to a certain movement. What we want to know is why the system is structured in this way rather than in some other way.

Recall that Dretske's fundamental problem is to find a place for intentional (or semantically interpretable) phenomena in the explanation of behavior. How could the fact that an internal state means some-

thing, or represents the world as being a certain way, contribute to the explanation of behavior? The answer that Dretske proposes focuses on structuring causes. Certain internal states of organisms lead to certain movements—the organisms are hooked up in that way—because they represent a particular fact or state of affairs. To see how this sort of explanation works, we need two further notions; indication and representation.

Indication, for Dretske, is the basic building block out of which more complex semantic or representational notions are built. One state of affairs indicates another if the occurrence of the latter is strongly correlated with the occurrence of the former. Thus, for example, in certain sorts of trees, the fact that the fifth ring in a cross section of the trunk is significantly wider than the other rings indicates that the tree grew more vigorously in its fifth year of life than in any other year. This notion of indication has a pair of important features. First, indication is a perfectly naturalistic notion; there is nothing spooky or mysterious about it. Second, indication is a very promiscuous relation. One state of affairs can be an indicator for many others. Thus, for example, if the tree whose rings we are examining lived in an arid environment, the larger than average fifth ring may also indicate that rainfall was significantly above average in the fifth year of the tree's life.

Let's turn now to representation. For Dretske, representation is a relation that obtains between an indicator and a state of affairs when the indicator has the function of indicating that state of affairs. Here again, nonbiological examples provide intuitive illustrations. In a thermostat, the degree of curvature of the bimetallic strip indicates the ambient temperature. Moreover, in the thermostat, it is the function of the strip to indicate the temperature. In this case it is easy enough to be very explicit about the function of the bimetallic strip, since an engineer designed the thermostat with just that indicator function in mind. It was because the strip indicates the temperature that the engineer wired it up in the way he did. Note here that we have a structuring explanation which appeals to the protosemantic indicating function of the bimetallic strip. An explanation of why the thermostat is hooked up in the way it is invokes the fact that the bimetallic strip indicates temperature.

Now, of course, there is a sense in which this is not a terribly interesting case of a structuring explanation, if our ultimate aim is to understand the explanatory role of representations, since the structuring is done by a person, whose beliefs, goals and other intentional states remain unexplained. But a central step in Dretske's attempt to find an explanatory role for intentional phenomena is his contention that there are phenomena to be found even in relatively simple organisms whose explanation is structurally analogous to the one just given for the ther-

mostat. There are biological processes in which an internal indicator comes to be hooked up to a movement controlling mechanism because of what it indicates.

The paradigm for Dretske's semantically involved structuring explanations is provided by operant conditioning. In operant conditioning, Dretske maintains, an internal indicator comes to have causal control over a movement producing mechanism: a C gets linked to an M. Moreover, this happens because the C is an indicator of some environmental feature. Consider the example of the rat in the Skinner box. At the beginning of the rat's training, a light goes on, and there is some internal state of the rat that indicates this fact. (Indeed, there are probably lots of different internal states that indicate the light being on. Indication, recall, is just lawful correlation). But at the outset none of the internal indicator states cause the rat to depress the lever. Sooner or later, however, the rat will happen to depress the lever when the light is on, and the result will be a reward. Note that the reward is contingent on two things: the light must be on and the lever must be pressed. Under these circumstances, some internal indicator of the light being on, some C, will gradually come to cause a kind of bodily motion it did not cause before. If it were not for the fact that the reward was causally contingent on the light being on, the internal indicator of the light would not have ended up linked to the movement-producing mechanism. So a semantic property of C—the fact that it is an indicator of the light being on—plays an essential role in a structuring explanation. If we want to know why C is hooked up to M—why the system is structured in this way—then the fact that C indicates the light being on is going to be an important element in the explanation. The pattern of this explanation is indicated schematically in Figure 1.

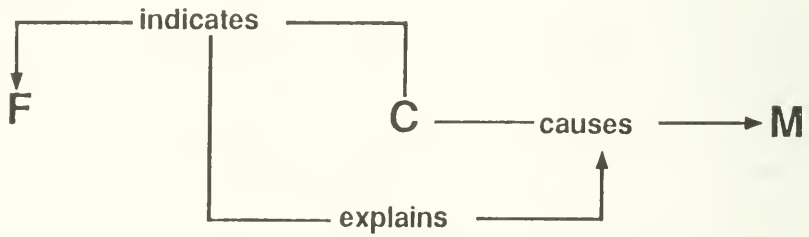

FIGURE 1. Causal and explanatory relations underlying indication and representation (after Dretske, 1988, p. 84). F is an event or condition of the environment and C the internal state of the organism which "indicates" that F occurs. In addition to indicating F, C comes to "represent" F, when part of the explanation of the causal link between state C and movement M is the fact that C indicates F.

"Once C is recruited as a cause of M," Dretske maintains,"—and recruited as a cause of M because of what it indicates about F—C acquires, thereby, the function of indicating F. Hence, C comes to repre-

sent F. C acquires its semantics, a genuine meaning, at the very moment when . . . the fact that it indicates F . . . acquires an explanatory relevance" (p. 84). Moreover, according to Dretske, once C is hooked up to M in this way, it acquires the status of a genuine belief (or proto-belief). That is because, in Dretske's view (borrowed from Ramsey 1931 and Armstrong 1973), a belief is an internal map "by means of which we steer" (p. 79). A bit less metaphysically, "beliefs are representational structures that acquire their meaning, their maplike quality, by actually using the information it is their function to carry in steering the system of which they are a part" (p. 81). "What you believe, i.e., the semantic content of your belief is relevant to what you do because beliefs are precisely those internal structures that have acquired control over output, and hence become relevant to the explanation of system behavior, in virtue of what they, when performing satisfactorily, indicate about external conditions" (p. 84).

FINDING THE C's AND THE M's: DRETSKE MEETS NEUROSCIENCE

Dretske's formulation of learning offers a plausible way of making explicit how semantic relations enter our explanations of behavior. Clearly, however, his account makes some rather demanding assumptions about the underlying neurobiology. Should any of these turn out to be unwarranted, Dretske's attempt to show how intentional psychology and neurobiology can co-exist would be undermined. We might wonder, for example, whether the neurobiologists find anything like Dretske's C's (explicit, identifiable indicators of environment conditions) or his M's (brain sites responsible for particular behaviors). An equally important question is how C's and M's become connected as the animal is conditioned. Can we really maintain that a C-M connection is established *because of what C indicates?* How are we to phrase this in neurobiological terms?

Ideally, to answer these questions, one would like to have a complete understanding of how operant conditioning—Dretske's example—works at the neurobiological level in a wide range of species. One could then determine whether it operates in general the way his account of the explanatory role of semantics supposes. Unfortunately, the neurobiology of learning and memory is still in its infancy and many decades away from what we require as far as operant conditioning is concerned. Nevertheless, at least one general mechanistic principle has begun to emerge. And while we may know little about operant conditioning, the situation is considerably better for other forms of learning, particularly in invertebrate species where these questions can be addressed somewhat more directly than in the human or even mammalian cases.

A central question for Dretske of course is whether indicators exist in the nervous system and whether these function during learning as he would suppose. We have known for a long time that states of the central nervous system co-vary with environmental stimuli and so "indicate" them. This issue has been elegantly explored by Mountcastle and coworkers (1957) in the somatosensory system and in the well known studies of Hubel and Wiesel (1977) in the visual system. Thus, it is practically part of the neurobiological canon that something like Dretske's C's do exist. Roughly the same can be said for movements. It has been known since the work of Sherrington that artificially stimulating appropriate regions of the brain and spinal cord can be sufficient to produce the twitches, scratches, blinks, and so forth that characterize a number of reflexes, including conditionable behaviors (e.g., Mauk and Thompson, 1984). These regions could in principle serve as the required M's, at least for simple behaviors.

The question then becomes what role C's and M's might have in learning. At present, neurobiologists can tell us very little about how operant conditioning occurs (though for promising results at the invertebrate level see Hawkins et al. 1985; Cook and Carew 1988). A great deal more is known about Pavlovian or classical conditioning. It is therefore tempting to ask whether results in this area run against the grain of Dretske's semantic indicator. Of course, we first must show that there is sufficient formal similarity—in light of Dretske's theory—between operant and classical conditioning.

The central difference between operant and classical conditioning lies in what governs reinforcement during training. In operant conditioning, reinforcement is contingent on something the animal does. In our earlier example, the reinforcing event (food) is withheld until the rat treads on a lever. In a classical conditioning experiment, reinforcement is completely independent of anything the animal does; it is linked instead to environmental cues or stimuli. Thus, in Pavlov's experiment, the reinforcer (meat powder) was contingent on the ringing of a bell but independent of the dog's movements.

This contrast is shown diagrammatically in panels A and B of Figure 2, where training contingencies introduced by the experimenter are shown as single arrows. Within each panel, the diagram on the left shows the potential connections that exist prior to conditioning (before). The diagrams on the right show the conditioned state (after). Notice that in operant conditioning, the reinforcing event R is linked to a movement M3, while in classical conditioning it is linked to a stimulus F3. A second difference is in the range of conditionable responses. The particular reinforcer in a traditional classical conditioning experiments restricts the conditioned behavior to movements produced by the reinforcer itself. Thus Figure 2B contains a single M, while in 2A, many M's could be connected to particular C's.

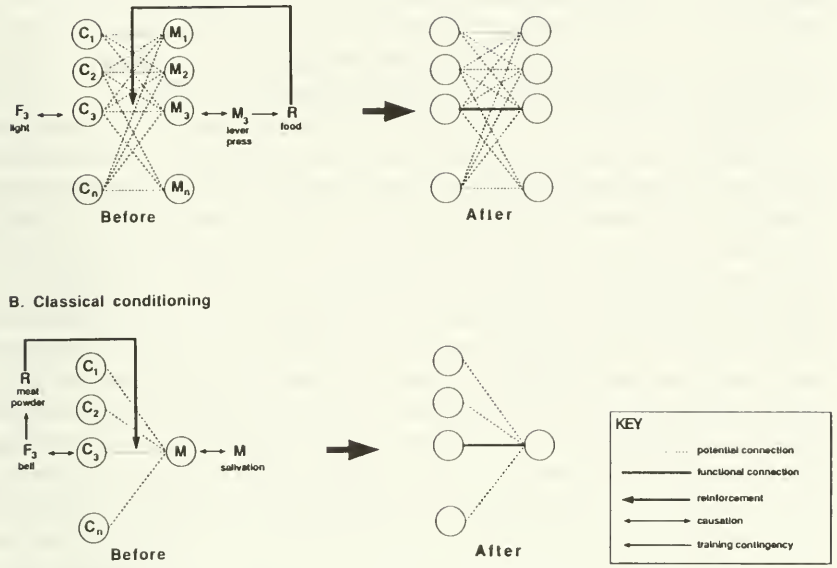

FIGURE 2. Contrasts between operant and classical conditioning. In operant conditioning (A), the reinforcing event R is contingent on a particular movement M3, but R could in principle have been contingent on any of the range of movements M1 to Mn, thus increasing the number of conditionable responses. In classical conditioning (B), R is independent of M and only the M produced by R becomes conditioned. F, C, and M are as in Figure 1; symbols are given in the key.

Returning briefly to the themes of the first section, let's ask what is special about operant conditioning, such that mental representation gets a foothold here. There are two elements to the answer. First, we have seen that for representation to be involved, "structuring" events must be among those in need of explanation. In the thermostat example, it was only of interest to ask why this sensor was wired to the garage door because everyone knows that insofar as the laws of physics are concerned, it could have been attached to any electrical appliance, limited only by the electrician's caprice. The situation is similar for operant conditioning. In the previous example the rat comes to lever-press (M) in the presence of the light (indicated by C), but we know that we could have conditioned him to a different light, or a tone, or a buzzer and so on. Similarly, lever-pressing is not the only conditioned response that could have become associated with the light. The second element of the answer is that indicator properties of C must be an important part of the explanation of the structuring event. When we inquire why *this* C became attached to *this* M (and not some other C), part of the answer must be that only this C indicated the relevant aspects of the environment. In the operant conditioning example, C3 indicates the light F3. C3 becomes associated with lever-pressing because of what the *light*

happens to be correlated with. As this particular operant experiment is conducted, the experimenter's M3–R contingency is only in force when the light is on. Only light was coupled to the M3–R contingency, therefore only the indicator of the light, C3, gained control over M3.

Despite their differences, both the operant and classical treatments are equally well viewed as cases where structuring explanations are required. One can just as easily ask why *this* C becomes connected to *this* M (and not some other C to some other M) in classical conditioning as in operant conditioning. This is because in classical conditioning too there are many potential C–M links, as Figure 2 suggests. Nor should the differences between the experimental contingencies underlying operant and classical conditioning obscure the fact that in both cases, the C that becomes associated with M is the one that indicates the relevant environmental event F, and that the relevant F is determined by the same considerations in the two cases. The relevant F is always the stimulus that is correlated with reinforcement. The fact that in one case (classical conditioning) F is correlated with the reinforcing event itself, while in the other (operant conditioning) F is correlated with the contingency of two other events (M and R) does not affect the character of the structuring explanation given. It is still which F a C indicates that determines its associability. Thus, there is an important formal similarity between operant and classical conditioning. Each is an instance of selective connection of C's to M's, and selectivity is in both cases achieved because only one C has the right indicator role.

We are now in a position to ask how well what is known about classical conditioning comports with the Dretskean models of Figure 2. The system in which the neural mechanism of classical conditioning has been pursued most thoroughly is the gill withdrawal reflex of the marine mollusk *Aplysia*. The *Aplysia* nervous system is simple by human standards yet this animal, or its close phylogenetic relations, performs favorably under a wide range of complex conditioning procedures including second-order conditioning, blocking, operant conditioning, food aversion training, and conditioned emotional response (for reviews, see Byrne 1987; Carew and Sahley 1986).

In response to a moderate tactile stimulus to a fleshy spout called the siphon, the animal withdraws its gill apparatus into a cavity on its back. This is accomplished in large measure by the activity of siphon sensory neurons that synapse directly on motor neurons that produce withdrawal of the gill. Shocking the animal's tail causes a much stronger gill withdrawal and can be used as a reinforcing stimulus. When activity of a siphon sensory neuron is paired with tail shock in a classical conditioning experiment, gill responses are enhanced to that sensory neuron but not another siphon sensory neuron whose activity was not paired with tail shock (Walters and Byrne 1983; Hawkins et al. 1983). Physiolog-

ical studies have subsequently localized the site of classical conditioning to the sensory neuron to motor neuron synapse.

A model for how classical conditioning occurs in *Aplysia* is shown in Figure 3. For simplicity, only two of the more than 20 sensory neurons are shown. Each sensory neuron (C3 and C2) can be considered an explicit indicator of touch (F1 and F2) to a particular region of the siphon. Reinforcement is subserved by a facilitating interneuron that responds to tail shock. When the experimenter introduces a specific contingency between activity in sensory neuron 1 and tail shock, the connection between this neuron and the motor neuron (M) is selectively strengthened, as the diagram on the right illustrates using the size of triangles to represent relative synaptic strength. The striking formal similarities to Dretske's mode of selective recruitment are clear from comparison of Figures 2B and 3. In each case a reinforcing mechanism acts to enhance just one of the possible C-M connections. In particular, the C-M connection is strengthened only for the indicator of the stimulus that in turn predicted the reinforcing event.

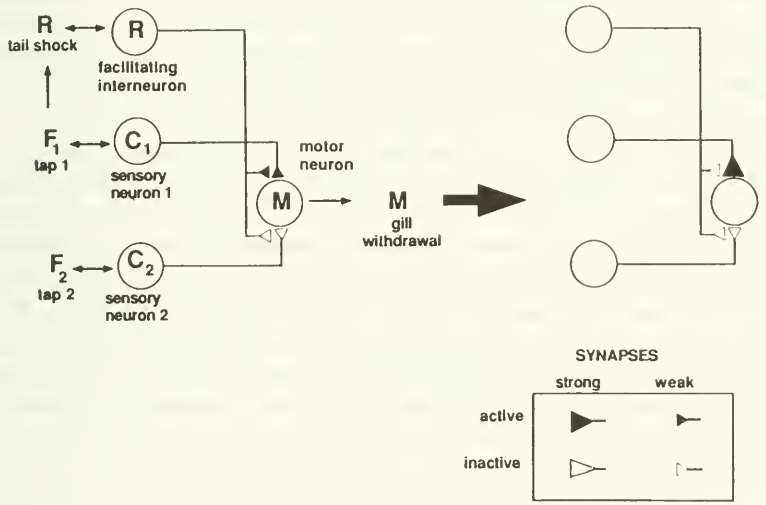

FIGURE 3. Model of classical conditioning in the gill withdrawal reflex in *Aplysia*. Circles represent the indicated sensory, motor, and facilitating interneurons contributing to the reflex. Triangles represent connections (synapses) between neurons. As shown in the key, active synapses are black and the size of a triangle indicates the strength of the connection. Each sensory neuron responds to stimulation (tap) of a different location on the skin. When stimulation of a particular location (e.g., location 1) is reinforced by tail shock, the connection from the corresponding sensory neuron to the motor neuron producing gill withdrawal is specifically enhanced.

Let us return to the example of the garage door wired to the thermostat, now with an eye for physical mechanism. Initially, of course, the thermostat and door were wholly unconnected. The connection was

established by the electrician who strung the wires from one to the other. One might call this the *ex nihlo* model of learning since the components start off with no connection of any kind. The consensus that is emerging from studies of how learning actually occurs points to a somewhat different model (Byrne 1987). We have already seen in *Aplysia* that learning consists of selectively enhancing existing connections. This also seems to be the case in less completely understood reflexes from a wide range of species, invertebrate and vertebrate alike. It is as though the wires are already in place but the contacts weak. This would be bad news for Dretske if his account of semantics required the *ex nihlo* model as his thermostat example would suggest. But there is no reason to suppose this is the case. As we have seen, for a particular C to count as a mental representation, all that is required is that its indicator properties be an important part of the structuring explanation. In particular, it is enough to show that the selectivity inherent in the structuring event (why this C and not *that* one) is attributable to the indicator properties of the C that wins out. The details of the resulting physical events that eventuate in the required connectivity causation are immaterial, so long as the causal role of C is essential to explaining what initiated them. It therefore does not matter whether, because of its particular capacities for indication, a certain C "grows" an entirely new connection or undergoes the strengthening of one already in place.

It is too early of course to say precisely how general the neural mechanisms of conditioning will turn out to be. But given the trend toward conservation during evolution, there is at least some reason to expect that the mechanics of classical conditioning will be fundamentally the same in other systems. It is interesting in this regard that the basic principles of simple classical conditioning can be used, at least theoretically, to construct models of more complex forms of conditioning, including operant conditioning. This means that the theme of selective recruitment operating on explicit indicators of sensory events may prove quite general indeed.

PROTO-BELIEFS AND INTENTIONALITY

Dretske, as we have seen, takes the case of operant conditioning as a simplified model for the sort of mental representation to be found in conscious human beliefs. The "proto-beliefs" that rats or sea slugs acquire in operant conditioning are much the same, Dretske maintains, as Dretske's own belief that there is a beer left in the fridge. The only major difference Dretske sees between the proto-belief of the sea slug and the full-fledged belief of the human is that humans have more of

them. The beliefs of adult humans are embedded in much richer networks. However, what we propose to argue in this section is that Dretske's "proto-beliefs" are actually very poor models for the sorts of beliefs presupposed by common-sense psychological explanations of human action.

At the core of our argument is the fact that adult human belief, as conceived of by commonsense psychology and as exploited in reason-giving explanations of human behavior, is capable of being very specific in the way it represents the world. What Fred believes is that there is a bottle of beer in the fridge, not that there is a bottle of ale, or stout. Oedepus believed that Jocasta would be a good person for him to marry. He did not believe that his mother would be a good person for him to marry. It is even possible for a person to believe that p and not believe that q, despite the fact that p and q are logically equivalent. The standard examples here involve propositions whose logical equivalence is not obvious, and not known to the person in question. Another facet of the precision that is possible in the common sense concept of belief is that people can have beliefs about things that they cannot reliably identify. Fred may believe that the gem in his wife's wedding ring is a diamond, though he has no idea how to distinguish real diamonds from fake ones. It is our contention that Dretske's notions of mental representation and proto-belief are incapable of attaining this sort of fine grained discrimination. Thus Dretske's constructs will not do as models for full-blown human intentionality.

At the heart of the problem is the fact that indication is a promiscuous relation. An internal state that indicates one external state of affairs will typically indicate many others as well—recall the example of the tree ring. As an indicator state, a C in Dretske's schematic formulation, comes to be a full-fledged representation of some state of affairs, F (and a proto-belief that F is the case), when it acquires the function of indicating F. But the same indicator will typically also indicate various other states of affairs: G, H, etc. How are we to determine which of the various states of affairs that C indicates it comes to represent? As we have seen, Dretske's answer is that C represents the state of affairs it has the function of indicating. However, it is our contention that the notion of function is simply not sufficiently discriminating to distinguish between the various state of affairs that C indicates.

All this will be a bit clearer if we consider a specific example. We'll focus on one that Dretske introduces himself. Monarch butterflies store a noxious substance from the milkweed plants on which they feed. When a bird eats a monarch butterfly it finds it foul tasting, and quickly learns to avoid monarchs in the future. Exploiting this fact, another species of butterfly, the viceroy, has evolved to resemble the monarch. However, the

viceroy has not evolved the monarch's system of storing a noxious substance. A bird that eats a viceroy is not punished by a foul taste. The viceroy is a mimic, an evolutionary freeloader, it gets by on looks alone.

Now consider the situation of the bird that has eaten a number of monarch butterflies and learned to avoid them. We may suppose that it has also eaten a viceroy or two, with no ill effects. Let's also suppose that this bird has encountered and eaten a monarch which, for one reason or another had not stored any noxious substance, once again with no ill effects. Now, if Dretske is right, there is some internal state—some C—in the bird that is lawfully correlated with monarch butterflies. Thus C indicates monarchs. However, C also indicates a larger class: the class consisting of monarchs and viceroys. The bird avoids viceroys because viceroys trigger state C. C also indicates a smaller class: the class of noxious monarchs; it was members of this class that provided the positive punishment that led to C being made a cause of avoidance behavior. When the bird has been conditioned—when C has come to cause avoidance behavior that it did not previously cause—what does C represent? Does it represent monarchs? or just noxious monarch? or monarchs and viceroys? Dretske answers this sort of question as follows:

> C will normally indicate a great many things other than F. Its indication of F is, therefore only "one component" of its natural meaning.[1] Nonetheless, it is this single component that is promoted to representational status . . . because it is C's indication of F, not its indication of (say) G or H, that explains it causing M. Hence, it becomes C's function to indicate F, not G or H (p. 84).

Making the substitutions relevant to the present example we obtain:

> C will normally indicate a great many things other than *noxious monarchs*. Its indication of *noxious monarchs* is, therefore only "one component" of its natural meaning. Nonetheless, it is this single component that is promoted to representational status . . . because it is C's indication of *noxious monarchs*, not its indication of (say) *monarchs* or [the class] *monarchs and viceroys*, that explains its causing M. Hence, it becomes C's function to indicate *noxious monarchs*, not *monarchs* or [the class] *monarchs and viceroys*.

Thus we see Dretske's answer is that C must represent only the reinforceable class, the noxious monarchs. But to draw the line here runs against the grain of common sense. Consider the example of the rat undergoing operant conditioning of lever presses to the light. Suppose lever pressing is reinforced by food, still only when the light is on,

1. As Dretske makes clear elsewhere (p. 55), "to mean" in the sense of natural meaning is, for him, synonymous with "to indicate."

but now on a less than 100 percent schedule. Drawing the obvious analogies, the light flashes associated with reinforced lever presses are equivalent to the noxious monarchs, and so it is only this class that the relevant C represents. But surely any coherent theory of animal information processing ought to maintain, to the contrary, that what C represents, and what the animal learns about, is the light—any and all of its flashes, not just the subclass of its reinforcement associated occurrences. Without this, there is no account of the animal's persistence in lever pressing when the light is on, yet lever pressing is not reinforced.

Another reason it seems more natural to say that C represents all the light flashes is that the C triggering mechanism (whatever it is) has no means of telling reinforcement associated flashes from flashes unrelated to reinforcement. The experimenters can tell the difference since obviously it is within their powers to inspect the state of the Skinner box on each trial, but from the point of view of the C in question, one flash is like any other. Of course, one monarch is like any other too, and here enters a new problem. From the point of view of the relevant C, one *monarch* is like any *viceroy.* This means that C represents monarchs—both kinds—and viceroys. The problem is that there appears to be no stopping. All things that trigger C (in a context where enough of them are associated with reinforcement to support conditioning) are represented by it. This means that C never misrepresents—the bird never has false beliefs. But the possibility of believing falsely is a property of human psychology just as essential as fine-grainedness of belief. On neither reading of representation do we find a suitable model of two truly central aspects of human mental representation.

It appears then that the notion of function is not strong enough to provide a well motivated way of deciding which of several quite different alternatives C represents. And if that's right, then Dretske's notion of representation and proto-belief will not be sufficiently fine grained to serve as the foundation for anything much like the commonsense notion of belief. If Fred believes that the insect he is looking at is a monarch, that belief is very different from the belief that the insect he is looking at is either a monarch or a viceroy. Indeed, the latter belief could well be true when the former is false.

This completes our case for one of the two claims we promised to illustrate in our introductory remarks. Along with Staddon (1988), we suspect that the study of intelligent and adaptive behavior in animals has provided relatively little insight into human psychology or human intelligence. Dretske's analysis of representation and proto-belief seems to us to be a case in point. His goal was to shed light on the structure of the intentional concepts we use in common-sense reason-giving explanations, and to explicate the strategy invoked in such explanation. But if we are right, then Dretske's efforts have only partially succeeded. The animal model has thrown light on beliefs of only the crudest sort and is

destined to blur the fine distinctions that make human beliefs truly human.

On the positive side, however, it seems to us that Dretske has at least made an intriguing beginning at explicating a strategy of explanation that is of importance in understanding adaptive behavior in animals. A complete neurobiological account detailing each step in the causal process from stimulus to behavior will not explain everything that needs explaining. It will not tell us why one rat presses the lever when the light goes on and its identical twin does not. It will not tell us why one hungry bird avoids viceroy butterflies while a conspecific does not. Explaining these facts requires more than circuitry; it requires an appeal to the history of the organism and the environment in which that history unfolded. The complete story about why certain birds avoid viceroy butterflies has two parts: First, they have an internal state which, in their environment, indicates both noxious monarchs and harmless viceroys. Second, there is a neural mechanism, the details of which are gradually becoming clear, which results in the indicator triggering avoidance behavior in just such circumstances. Plainly there is something quasi-intentional about such explanations. They involve both the internal workings of the organism and correlations with the environment. It is less clear whether any more richly intentional notion will play a role in the explanation of adaptive behavior in animals—whether notions like representation, misrepresentation and belief which play a central role in human reason-giving psychology have any work to do in the explanation of animal behavior. In our view the answer to this question can't be settled by a priori speculation. Only careful research and theory building will do. Should it turn out that the answer is negative, however, we won't be surprised or disappointed. Understanding adaptive behavior in animals is a profoundly interesting project even if it does not provide a useful model for purposive behavior in humans.

ACKNOWLEDGMENT

We wish to thank Peter Godfrey-Smith, William Ramsey, and Valerie Walker for discussion. S. Lockery was supported by a NSF predoctoral fellowship.

REFERENCES

Armstrong, D. M. 1973. *Belief, truth and knowledge.* (CUP, Cambridge).

Byrne, J. H. 1987. Cellular analysis of associative learning. *Physiological Reviews* 67:329-439.

Carew, T. J. and Sahley, C. L. 1986 Invertebrate learning and memory: from behavior to molecules. *Ann. Rev. Neurosci 9,* 435-487.

Cook, D. G. and Carew, T. J. 1988. Operant conditioning of identified neck muscles and individual motor neurons in *Aplysia. Abstra. Soc. Neurosci 14*, 607.
Dennett, D. C. 1987. *The intentional stance, Ch. 8 (MIT, Cambridge).*
Dretske, F. 1988 *Explaining behavior* (MIT, Cambridge).
Hawkins, R. D., Abrams, T. W., Carew, T. J, and Kandel, E. R. 1983. A cellular mechanism of classical conditioning in *Aplysia:* activity-dependent amplification of presynaptic facilitation. *Science* 219, 400-405.
Hawkins, R. D., Clark, G. A., Kandel, E. R. 1985. Operant conditioning and differential classical conditioning of gill withdrawal in *Aplysia. Abstr. Soc. Neurosci. 11*, 796.
Hubel, D. H., and Wiesel, T. N. 1977. Functional architecture of macaque visual cortex. *Proc. Roy. Soc. (Lond) Series B 198*, 1-59.
Mauk, M.D. and Thompson, R. F. 1984. Classical conditioning using the inferior olive as the unconditioned stimulus. *Abstr. Soc. Neurosci. 10*, 122.
Mountcastle, V. 1957. Modality and topographic properties of single neurons of the cat's somatic sensory cortex. *J. Neurophysiology 20*, 408-431.
Ramsey, F. P. 1931. *The foundations of mathematics, and other logical essays* (RKP, London).
Staddon, J. E. R. 1988. Animal psychology: the tyranny of anthropocentrism. In P. P. G. Bateson and P. H. Klopfer (eds) *Perspectives in Ethology.* Vol. 8: Whither Ethology Plenum, London.
Walters, E. T., Byrne, J. H. 1983. Associate conditioning of single sensory neurons suggests a cellular mechanism for learning *Science 219*, 405-408.