On Explaining Behavior

Mary Litch
Department of Philosophy
425A Humanities Building
University of Alabama at Birmingham
Birmingham, AL 35294-1260
Phone: (205) 934-8907 Fax: (205) 975-6639
E-mail: litch@uab.edu

Paul Cohen
Department of Computer Science
Lederle Graduate Research Center
University of Massachusetts at Amherst
Amherst, MA 01003
Phone: (413) 545-3638 Fax: (413) 545-1249
E-mail: cohen@cs.umass.edu

On Explaining Behavior

Abstract

In his book *Explaining Behavior*, Fred Dretske distinguishes between types of representational systems based on whether the actions or intentions of an external agent need be mentioned in an explanation of the system s behavior. He argues that the relevant criterion in making this distinction is whether the system s behavior was structured by a learning process or by the activity of the system s designer/constructor. However, our work with learning mobile robots demonstrates that the distinction learned versus designed : (i) is not always well-defined, and (ii) cannot *by itself* be used for deciding whether the activities of an outside agent must be mentioned in an explanation of a system s behavior. We argue that Dretske s theory experiences similar difficulties even when the theory is applied to natural systems (e.g., humans). The problem arises because not all representational states fit neatly into one of the three idealized types he describes. There is a fairly broad class of states that fall somewhere between Type II and Type III. For system behavior involving these states, we ask the question anew: What is the relevant criterion for deciding whether the activities of an outside agent must be mentioned in an explanation of a system s behavior?

On Explaining Behavior

Introduction

In a series of books and articles,[i] Fred Dretske attempts to satisfy two pressing needs in the philosophy of mind by: providing a theory of content and demonstrating how content can be relevant in the explanation of behavior. Unlike most other philosophers writing on the same set of topics,[ii] Dretske has developed a theory and classification scheme that applies to all representational systems, not just those described as psychological . While *Explaining Behavior* makes a major contribution to philosophers understanding of content (especially psychological content), we are most interested in what Dretske has to say about *explanation*.

Dretske distinguishes between triggering and structuring causes. A triggering cause is the event that caused some effect. (In the token causal process C _ E, C is the triggering cause.) A structuring cause is the cause of C s causing E (i.e., whatever was responsible for arranging things such that C caused E). Consider the case of a pigeon trained via instrumental conditioning to peck at a lever whenever a certain light goes on. At time t, the light goes on and the (hungry) pigeon pecks at the lever. The triggering cause of the pecking is the pigeon s seeing that the light is on. (This will be some state in the pigeon s brain.) The structuring cause of the pecking is the set of past training trials that wired the pigeon s brain such that now (at time t) it pecks whenever it sees the light go on (and it is hungry). Note that structuring causal explanations are *historical* in nature: they advert to events that took place *before* the causing of the event to be explained. Dretske s contribution to the debate on the causal efficacy of mental content was in working out and then combining two separate parts of the puzzle: a theory of content and a description of the nature of content-adverting explanation. The first part of the puzzle was the development of an historical account of psychological content. According to this theory, psychological content is determined by the learning history of the system. Non-psychological content (the content had by non-learning representational systems) is determined by the intentions (and actions) of the designer/constructor of the representational system. Some philosophers of mind have argued that the physical properties of mental states screen off the mental properties (in particular, the *content* properties) of those mental states from playing a causal role.[iii] Dretske s second main contribution was noticing that this difficulty could be avoided by finding a role for content to play as *structuring* cause.

The Type II/Type III distinction is a distinction between kinds of representational states.[iv] However, the distinction is also relevant to Dretske s theory of explanation. Consider again the causal process C _ E. Assume that C is a representational state. If C is a Type II state, then the structuring cause of E will mention the actions of some outside agent. If, however, C is a Type III state, then the structuring cause of E will mention the learning history of the system that resulted in the C_ E connection.[v] In the remainder of the paper, we focus on the Type II/Type III distinction as it applies to explanation.

We begin by describing our work with learning mobile robots. We then argue that, at least according to one interpretation of the Type II/Type III distinction, our robot fits the bill for a Type III representational system. However, things are not quite so straightforward. We have noticed that there are serious difficulties in maintaining the Type II/Type III distinction in practice. We describe some problematic cases in the section entitled Breakdown in the Type II/Type III Distinction . For example, it is possible to manipulate the learning process to the extent that the resulting representational states, while receiving their control duties as the result of a learning process, nevertheless have their content assigned by an outside agent. Suppose a heavily manipulated learning process results in a C _ E connection. Event C occurs, and causes E. It is unclear whether the structuring cause of E is the learning by the system or the manipulation by the outside agent (or, perhaps, some combination or the two). One should not be misled into thinking that this problem only arises for artificial learning systems; while it is more obvious for artificial systems (because artificial systems supply more avenues for manipulation), difficulties in classification as Type II or Type III (and concomitant difficulties in picking out structuring causes) arise for *natural* systems as well.

The Robot and Some of What It Learns

In this section, we provide information on our robot, its controller and (most importantly) the learning algorithm that runs on top of and influences the controller. We also try to give some idea of what the robot is capable of learning. This description of the robot is necessary, first, for establishing that the robot can be characterized as a Type III representational system, and, second, for giving the background necessary for understanding the difficulties that arise in applying the Type II/Type III distinction in practice.

Our robot is a Pioneer-1 mobile robot, with several effectors and roughly forty sensors.[vi] The robot has learned numerous contingencies, including dependencies between its actions, the world state, and changes in the

world state by processing data gathered as it roams around our laboratory.[vii] In this section, we focus on one learning method, clustering by dynamics, and a primitive ontology of actions that it learned without supervision.

The robot s state is polled every 100msec., so a vector of 40 sensed values is collected ten times each second. These vectors are ordered by time to yield a multivariate time series. Figure 1[viii] shows four seconds of data from just four of the Pioneer s forty sensed values. Given a little practice, one can see that this short time series corresponds to the robot moving past an object. Prior to moving, the robot establishes a coordinate frame with an x-axis perpendicular to its heading and a y-axis parallel to its heading. As it begins to move, the robot measures its location in this coordinate frame. The ROBOT-X line is almost constant, while the ROBOT-Y line increases, indicating that the robot moved straight ahead. The VIS-A-X and VIS-A-Y lines indicate the horizontal and vertical locations, respectively, of the centroid of a patch of light on the robot s retina , a CCD camera. VIS-A-X decreases, indicating that the object s image drifts to the left on the retina, while VIS-A-Y increases, indicating the image moves toward the top of the retina. Then, both series jump to constant values simultaneously.[ix] In sum, the four-variable time series in Figure 1 indicates the robot moving in a straight line past an object on its left, which is visible for roughly 1.8 seconds and then disappears from the visual field.

Every time series that corresponds to moving past an object has qualitatively the same structure as the one in Figure 1 — namely, ROBOT-Y increases; VIS-A-Y increases to a maximum then takes a constant value; and VIS-A-X either increases or decreases to a maximum or minimum depending on whether the object is on the robot s left or right, then takes a constant value. ROBOT-X might change or not, depending on whether the robot changes its heading or not.

It follows that if we had a statistical technique to group the robot s experiences by the characteristic patterns in time series, then this technique would in effect learn a taxonomy of the robot s experiences. Clustering by dynamics is such a technique.[x] The end result of applying clustering by dynamics to the multivariate sensor data gathered by the robot as it wanders around the lab is the learning of prototypes. A prototype is a characteristic pattern in the robot s time series of sensor data as it performs some action. The details of the clustering by dynamics algorithm are described in an endnote.[xi] In a recent experiment, this procedure produced prototypes corresponding to passing an object on the left, passing an object on the right, driving toward an object, bumping into an object, and backing away from an object.[xii]

We claim that these prototypes were learned largely without supervision and constitute a primitive ontology of activities - the robot learned some of the things it can do. What supervision or help did we provide? We wrote the programs that controlled the robot and made it do things. We divided the sensor data time series into episodes (although this can be done automatically). We limited the number of variables that the dynamic time warping code had to deal with, as it cannot efficiently handle multivariate series of forty state variables. We did not label the episodes to tell the learning algorithm which clusters of activities it should consider. In fact, the only guidance we provided to the formation of clusters was a threshold statistic for adding an episode to a cluster. To reiterate, we did not anticipate, hope for, or otherwise coerce the algorithm to learn particular clusters and prototypes. Thus we claim that the robot s ontology of activities is its own.

Recently we have been trying to close the loop and have the robot learn enough about the preconditions and effects of its actions that it can plan to accomplish a goal. For instance, suppose the robot wants to drive the state of the bump sensor from low to high (i.e., it wants to bump into something); what should it do?[xiii] Previous work[xiv] discusses how the robot learns models of single actions from its prototypes. But planning is more than just executing single actions. Planning means reasoning about the effects of sequences of actions to achieve a goal state. To plan, the robot needs to transform prototypes into planning operators that specify the preconditions and effects of actions. The algorithm that performs this transformation is described in an endnote.[xv]

A single set of preconditions (one set for each prototype) is the vector of average sensor values that accurately predicts the future series of sensor values when a particular operator is applied. After learning, the robot will perform the operation specified by a prototype whenever both (i) its most recent time series of sensor values matches the set of preconditions for that prototype (within some tolerance), and (ii) it currently has a want that is satisfied by an effect of the operator associated with that prototype. We shall henceforth use the term preconditions satisfier (abbreviated PS) to refer to the data structure encoding the time series of sensor values, *when that time series matches any of the sets of learned preconditions*. (This term is applicable to the sensor state data structure only when a match occurs.)

PSs are Type III Representational States

We claim that our robot is a Type III representational system. More specifically, we claim that the robot's PSs satisfy the requirements for Type III status. A Type III state must: *indicate* some external condition,[xvi] have the *function of indicating* that condition, and have this *function assigned as the result of a learning process*.

We shall argue that PSs acquire their indicator functions through learning; although, as we discuss in the section entitled Breakdown in the Type II/Type III Distinction , there are problems in applying the learning criterion from Dretske s theory to our robot. Recall that, for Dretske, a state of a larger system *indicates* some external condition when there is a covariance between the system going into that state and the external condition s obtaining.[xvii] PSs indicate (because they covary with) external conditions. In the case of PSs, the conditions indicated involve the location of objects relative to the robot. PSs in the post-learning robot cause the robot to take specific actions to satisfy its wants. PSs have been given this control duty *because of what they indicate* about the state of the world (namely, that the location of objects relative to the robot is such that execution of these actions is likely to bring about satisfaction of a want). The specific control duty assigned to a PS is determined by a learning process — namely, by the learning algorithm that runs on top of the controller. Thus, PSs are Type III states.

The bulk of *Explaining Behavior* deals with the naturalization of content for beliefs of a rather primitive sort — beliefs about perceptually salient features of one s immediate environment.[xviii] In Chapter 6, Dretske allows that more perceptually remote beliefs will have their content determined, at least in part, by their functional (or conceptual) role in the production of output (including their *internal* relations to each other). [xix] Thus, one may object to our claim that PSs are Type III states as follows. There is an implicit requirement that must be satisfied in order to obtain Type III status (in addition to those mentioned above): the state must be one in a web of beliefs and desires that attains some critical level of complexity. This objection continues: without a complexity condition, punctate minds (e.g., minds containing only one belief) would be possible. However, a punctate mind is an impossibility.[xx] We reject this complexity condition, because the model of a cognitive agent that we have uppermost is not an adult human (for whom the complexity condition is appropriate), but an infant. We are trying to understand how mental content can be bootstrapped, given a small primitive set of wants and action types. A major goal of our project is to show how this bootstrapping is possible with limited innate structure. (This goal is shared by Dretske.[xxi]) Thus, our rejection of the complexity condition for Type III status is justified.

Explaining Behavior Involving Type II versus Type III States

As mentioned previously, Dretske s theory and classification scheme apply to all representational systems. This inclusiveness is justified, because explanations of all behaviors involving representational states share some general features in common. Although Type II and Type III states differ with respect to how their contents are determined (i.e., how their indicator functions are assigned), the explanation of behavior involving those states in terms of content is the same; for both, the content is explanatorily relevant to the system s behavior. Compare two robots. The first robot learns how to bump into an object using the method described in the section of this paper entitled The Robot and Some of What it Learns . The second robot comes with a fixed controller that accomplishes the same task. (In both cases, the robot possesses a state with the content there s an object in front of me .) Suppose both robot1 and robot2 bump objects at time t. Robot1 bumped an object (in part) because it was in a state with the content there s an object in front of me . Robot2 bumped an object (in part) because it was in a state with the content there s an object in front of me . Obviously, how the content is determined differs in the two cases, but this doesn t imply that explanation in terms of content is somehow less legitimate for robot2 than for robot1.[xxii]

The two cases diverge when one takes one step back and asks of robot1 and robot2: How did the state whose content is alluded to in the explanation acquire its content? For *learned* behavior (e.g., the stimulus-response pattern learned by robot1) this 2nd tier explanation is given in terms of an indicator function acquired through a learning process.[xxiii] For *unlearned* behavior, the 2nd tier explanation is given in terms of an indicator function that arises because of the intentions and actions of the designer or constructor as she structured the system.

One purpose that Type III representation serves within Dretske s framework is that it stops the chain of derived content; Type III states ground all intentionality. (The grounding of intentionality is one of the criteria for success of any theory of content naturalization.) However, for Type III states to serve this purpose, the authorship of Type III content must be isolated strictly within the individual Type III representational system. This condition is problematic for a wide variety of states that in all other respects appear to be clear candidates for Type III status.

Breakdown in the Type II/Type III Distinction

We argued above that PSs are Type III states; however, if one takes seriously the claim that the authorship of Type III content must be strictly isolated within the individual, it is not so clear that PSs are Type III states. While others have attacked Dretske for his insistence on the learning condition,[xxiv] our take on the issue differs from that taken by

others and is based less on philosophical concerns than on concerns relating to the application of the theory to a concrete system.

We have tried to avoid reverse engineering the robot or the robot s environment so as to coerce a particular result.[xxv] Even though we have minimized task-specific innate structure, we cannot avoid playing a significant role in the design of the robot s control and learning algorithms. In addition, we manipulated the robot s learning environment: the prototypes (and associated PSs) for moving toward and past objects were learned from trials that we set up. We varied the placement of objects relative to the robot and instructed the robot to move. Even so, we had no idea whether the robot s learning algorithm would produce prototypes, we did not anticipate the prototypes it produced, and we were pleasantly surprised that the prototypes made sense to us. So who is the author of the learned PSs contents? Dretske might take a hard line and say that the robot did not develop its prototypes all by itself, so PSs are not Type III states, but then he would also have to rule out concepts in a curriculum learned by human schoolchildren. In fact, curricular concepts are even more suspect than our robot s PSs, because teachers intend their students to believe something while our placement of objects around the robot was not intended to teach the robot any particular prototype (and, consequently, any particular PS). On the other hand, if our contribution to what the robot learned does not disqualify PSs from Type III status, then what other help are we allowed to provide? Suppose we classify prototypes as good and bad , and make the robot learn our classification rule. Does the fact that a learning algorithm inserts the rule qualify the state that predicts class for Type III status? Since the rule is entirely conventional, the fact that the rule was learned is not relevant to explaining the robot s behavior in correctly classifying prototypes as good or bad : the rule could just as well have been inserted by a programmer. We must also consider in-between cases like those in reinforcement learning, where a system learns our rule by seeking to maximize a reinforcement signal that may well be endogenous; for example, when I reward a child for saying please , she undoubtedly learns, mediated by her unique reinforcement function. But she is learning *my* rule.

Dretske wishes to distinguish types of representations based on whether their content is the result of learning or design, but the distinction learned versus designed : (i) is not always well-defined, and (ii) cannot *by itself* be used for deciding whether the activities of an outside agent must be mentioned in an explanation of a system s behavior. Dennett (1992) accuses Dretske of wanting do it yourself understanding , in which the authorship of understanding belongs entirely to the individual. Among our robots, at least, there are no such

individuals and they have no such understanding. A distinction between Type II representations, in which the functions of indicators are assigned, and Type III representations, in which the functions are learned, is not practical.

No aspect of our criticism of the Type II/Type III distinction has been the result of our focusing on an *artificial* learning system: a black and white distinction is just as untenable for natural learning systems. Parents of small children arrange their children s environment in a way not unlike the way we arrange our robot s environment. To what extent then are children s perceptually-based beliefs something less than full-blown Type III? Children are presumably the archetype of the Type III representational system. Although it was our work with robots that led us to notice that the Type II/Type III distinction is problematic, the difficulty is a general one, and arises when applying Dretske s theory to any learning system.

The Upshot for Explanation

So, when must the activities of an outside agent be mentioned in an explanation of a system s behavior? Unfortunately, we do not have a nice, neat answer to this question. Clearly, though, whether the behavior was learned is not by itself decisive.[xxvi] The root of the problem here seems to be that learning environments are more complicated than Dretske s simple Type II/Type III distinction assumes. We can think of Type III states as he envisions them (call them pure Type III states) as one end of a *continuum*, the other end of which is marked by pure Type II states. The representational states lying in between these two ends vary as a function of how large a role an outside agent played in structuring the system; however, at this juncture, we can discern no single condition or set of conditions that would allow the defining of some critical amount of outside involvement.

Figures

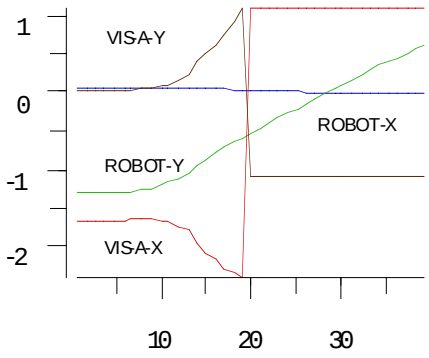

Figure 1. Time series of 4 sensors corresponding to the robot moving past an object on its left.

References

Cummins, R. 1989. *Meaning and Mental Representation*. MIT Press. Cambridge, MA.

Davidson, D. 1987. Knowing One s Own Mind . *Proceedings and Addresses of the American Philosophical Association*, pp. 441-458.

Dennett, D. 1991. Ways of Establishing Harmony in B. McLaughlin s (editor) *Dretske and His Critics*, Basil Blackwell. Cambridge, MA. 1991.

Dennett, D. 1992. "Do-It-Yourself Understanding". Reprinted in D. Dennett's *Brainchildren*. MIT Press. Cambridge, MA. 1998.

Dretske, F. 1988. *Explaining Behavior*. MIT Press. Cambridge, MA.

Dretske, F. 1991. Dretske s Replies in B. McLaughlin s (editor) *Dretske and His Critics*. Basil Blackwell. Cambridge, MA. 1991.

Dretske, F. 1993. "Mental Events as Structuring Causes of Behavior", in J. Heil and A. Mele's (editors) *Mental Causation*, Oxford University Press. Oxford, 1993.

Dretske, F. 1995. ""Does Meaning Matter?" in C. MacDonald and G. MacDonald's (editors) *Philosophy of Psychology* Vol. 1. Basil Blackwell. Cambridge, MA. 1995.

Fodor, J. and LePore, E. (editors) 1992. *Holism: A Shopper's Guide*. Basil Blackwell. Cambridge, MA.

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

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

Heil, J. 1992. *The Nature of True Minds*. Cambridge University Press. Cambridge.

Kim, J. 1989. Mechanism, Purpose, and Explanatory Exclusion , reprinted in J. Kim s, *Supervenience and Mind*, Cambridge University Press. Cambridge. 1993.

Millikan, R. G. 1984. *Language, Thought and Other Biological Categories*. MIT Press. Cambridge, MA.

Rosch, E. and Mervis, C. B. 1975. "Family resemblances: Studies in the internal structure of Categories", *Cognitive Psychology*, Vol. 7, pp. 573-605.

Sankoff, D. and Kruskal, J. B. (editors) 1983. *Time Warps, String Edits, and Macromolecules: Theory and Practice of Sequence Comparison*. Addison-Wesley. Reading, MA.

Stich, S. 1983. *From Folk Psychology to Cognitive Science*. MIT Press. Cambridge, MA.

Endnotes

[i] The most comprehensive sources in this series are Dretske, 1988 and 1993.

[ii] For example, Fodor, 1987 and 1990, Millikan, 1984, Cummins, 1989.

[iii] The most noteworthy example of a philosopher who holds that mental properties are screened off is Kim, 1989. For an overview of the debate, see Heil, 1992.

[iv] Dretske introduces the Type II/Type III distinction as applicable to whole representational systems, not individual states. However, the distinction only makes sense when applied to individual states within larger systems. We say this for two reasons. (1) The Type II/Type III distinction boils down to how an indicator function is assigned to a representational element. Having a particular indicator function is a property of states, not systems. (2) It is conceivable that a system has some states whose indicator functions are assigned by an external agent and some states whose indicator functions are the result of learning.

[v] There is a category of non-conventional representation that falls outside of the Type I/Type II/Type III scheme: representational states whose content is determined by a process of natural selection. Because the contents of such states are wholly non-conventional, they are not of Type II; however, Dretske explicitly disallows them from classification within Type III. (See *Explaining Behavior*, page 64.) Dretske s stated reason for excluding states with evolutionarily determined content from Type III is that this latter grouping is used to pick out psychological states, and Dretske thinks that, in order for a state to be a psychological state, its indicator function must be assigned as the result of a learning process during the life-span of the system.

[vi] The robot has two independent drive wheels, a trailing caster, a two degree of freedom gripper, and roughly forty sensors including five forward-pointing sonars, two side-pointing sonars, a rudimentary vision system, bump and stall sensors, and sensors that report the state of the gripper.

[vii] Reference removed for blind review.

[viii] The figure is on the page between the body of the paper and the bibliography.

[ix] These constant values are returned by the vision system when nothing is in the field of view.

[x] References removed for blind review.

[xi] First, one divides a long time series into segments, each of which represents an episode such as moving toward an object, avoiding an object, crashing into an object, and so on. Episode boundaries can be inserted by humans or by a simple algorithm that looks for simultaneous changes in multiple state variables. Obviously we prefer the latter technique (and apply it in [references removed for blind review]) because it minimizes human involvement in the learning process; however, for the experiment described here, episode boundaries were marked by us. We did not label the episodes in any way. Second, a dynamic time warping algorithm compares every pair of episodes and returns a number that represents the degree of similarity of the time series in the pair. Dynamic time warping is a technique for morphing one multivariate time series into another by stretching and compressing the horizontal (temporal) axis of one series relative to the other. (Sankoff and Kruskal, 1983) A number that indicates the amount of stretching and compressing is thus a proxy for the similarity of two series. Third, having found this similarity number for the series that correspond to every pair of episodes, it is straightforward to cluster episodes by their similarity. Agglomerative clustering is a method to group episodes by similarity such that the within-cluster similarity among episodes is high and the between-cluster similarity is low. Fourth, another algorithm finds the central member of each cluster, which we call the cluster prototype following Rosch and Mervis, 1975.

[xii] Reference removed for blind review.

[xiii] The robot s wants are implemented by a trivial algorithm that selects a sensor and tries to change the sensor s current value. Obviously, most human wants are more sophisticated, yet we think our simple algorithm is a primitive model of exploratory motor behavior in infants.

[xiv] Reference removed for blind review.

[xv] First, each episode is labeled with the cluster to which it belongs. Next, the first 1000 msec. time series of each state variable in each episode is replaced by its mean value. These are the initial conditions , the average values of state variables at the beginning (i.e., the first 1000 msec.) of each episode. Initial conditions are not the same as preconditions. To qualify as a precondition in an episode, an initial condition must at least make good predictions about how the episode will unfold. That is, an initial condition cannot be a precondition for a prototype if it is uncorrelated with that prototype across episodes. We have a batch of episodes, and each is labeled with its cluster membership, and each has a list of initial conditions, so it is a simple matter to run these data through a decision tree induction algorithm to find those initial conditions that best predict the cluster membership of the episodes. Since each cluster is represented by exactly one prototype, these predictive initial conditions are interpreted as preconditions for the prototypes.

[xvi] Strictly speaking, a Type III state need not indicate what it has the function of indicating. (The bifurcation between indicating and having the function of indicating is what allows Dretske to explain how misrepresentation is possible.) However, Dretske does insist (Dretske, 1988, pp. 59-60) that, at a minimum, to have the function of indicating that F, a state must be (in theory?) able to indicate that F. There is no harm in stiffening the requirement for Type III status to require that a state in fact indicates some external condition, because if PSs satisfy this stiffened requirement, then they automatically satisfy some lesser requirement.

[xvii] Thus, the physical states of a thermometer (in particular, the level of mercury in the thermometer) indicate the external temperature because the level of mercury covaries with the external temperature.

[xviii] Dretske, 1988, page 138.

[xix] Dretske, 1988, pp. 150-151.

[xx] One finds this sort of argumentation in the papers collected in Fodor and LePore, 1992.

[xxi] See Dretske (1988, pp. 106-107).

[xxii] In a certain derived sense, then, it is the fact that C [the bending of a bimetallic strip in a thermostat/furnace system] means what it does, the fact that it indicates the temperature, that explains (through us, as it were) its *causing* what it does. And its causing, or being made to cause, what it does *because* it means what it does is what gives the indicator the function of indicating what it does. An internal indicator acquires genuine (albeit derived) meaning — acquires a *representational content* of Type II — by having the fact that it indicates F, determine its causal role in the production of output. (Dretske, 1988, page 87)

[xxiii] [D]uring the normal development of an organism, certain internal structures *acquire* control over peripheral movements of the systems of which they are a part. Furthermore, the explanation, or part of the explanation, for this assumption of control duties is not (as in the case of [non-learning] artifacts) what anyone *thinks* these structures mean or indicate but what, in fact, they *do* mean or indicate about the external circumstances *in which* these movements occur and *on which* their success depends. In the process of acquiring control over peripheral movements (in virtue of what they indicate), such structures acquire an indicator function (Dretske, 1988, page 88)

[xxiv] See, for example, Dennett (1991 and 1992), Davidson (1987) and Stich (1983). For Dretske s response, see Dretske (1988, p. 105), Dretske (1991, pp. 201-202) and Dretske (1995, pp. 116-117).

[xxv] The charge of reverse engineering arises in both the traditional AI and connectionist approach; the tweaking of network parameters that goes on in connectionist research is not qualitatively different from the reverse engineering in traditional AI learning systems.

[xxvi] This is a co-authored paper. At least one of us is willing to use learning as a necessary condition for Type III status.