Drutske

Chapter 4
The Explanatory Role of Belief

Armstrong (1973), following Ramsey (1931), has described beliefs as maps by means of which we steer. In the last chapter, we examined the maplike character of representations—the way they indicate, or have the function of indicating, the content and the nature of one's surroundings. But beliefs are not merely maps; they are maps *by means of which we steer*. And if this metaphor is to have any validity, as I think it does, then what makes the map a map—the fact that it supplies information about the terrain through which one moves—must, in one way or another, help to determine the direction in which one steers. If a structure's semantic character is unrelated to the job it does in shaping output, then this structure, though it may *be* a representation, is not a belief. A satisfactory model of belief should reveal the way in which *what we believe* helps to determine *what we do*.

The job of this chapter is to supply this account, to show that there are *some* representations whose role in the determination of output, and hence in the explanation of behavior, is shaped by the relations underlying its representational content or meaning. Such representations, I submit, are beliefs.

4.1 *The Causal Role of Meaning*

Something possessing content, or having meaning, can *be* a cause without its possessing that content or having that meaning being at all relevant to its causal powers. A soprano's upper-register supplications may shatter glass, but their meaning is irrelevant to their having this effect. Their effect on the glass would be the same if they meant nothing at all or something entirely different.

What is true of the soprano's acoustic output is true of reasons—those content-possessing mental states (belief, desire, fear, regret) we invoke to explain one another's behavior. We can, following Davidson (1963), say that reasons *are* causes, but the problem is to understand how their being reasons contributes to, or helps explain, their effects on motor output. It has been pointed out often enough that although reasons may cause us to behave in a certain way, they may not, *so described*, explain the behavior they cause (McGinn 1979; Mackie 1979; Honderich 1982; Robinson 1982; Sosa 1984; Skillen 1984; Follesdal 1985; Stoutland 1976, 1980; Tuomela 1977). McGinn (1979, p. 30) puts it this way: "To defend the thesis that citing reasons can be genuinely explanatory, we need to show that they can explain when described *as reasons*." The fact that they have a content, the fact that they have a *semantic* character, must be relevant to the kind of effects they produce. If brain structures possessing meaning affect motor output in the way the soprano's acoustic productions affect glass, then the meaning of these neural structures is causally inert. Even if it is there, it doesn't *do* anything. If having a mind is having *this* kind of meaning in the head, one may as well not have a mind.

Haugeland (1985, p. 40) notes that this problem is merely a reenactment within a materialistic framework of an old problem about mind-body interaction. Materialists think to escape this difficulty by claiming that a thought, like everything else, is merely a physical object—presumably (in the case of a thought) a neural state or structure. That may be so, of course, but what about the *meanings* of these physical structures? Are they, like the mass, charge, and velocity of objects, properties whose possession could make a difference, a *causal* difference, to the way these neural structures interact? If meaning, or something's *having* meaning, is to do the kind of work expected of it—if it is to help explain *why* we do what we do—it must, it seems, influence the operation of those electrical and chemical mechanisms that control muscles and glands. Just how is this supposed to work? This, obviously, is as much a mystery as the interaction between mind stuff and matter.

My task is to show how this embarrassment can be avoided within a materialist metaphysics. I will *not* try to show, of course, that meanings *themselves* are causes. Whatever else a meaning might be, it certainly is not, like an event, a spatio-temporal particular that could cause something to happen. It is, rather, an abstract entity, something more in the nature of a universal property such as redness or triangularity. Trying to exhibit the causal efficacy of meaning itself would be like trying to exhibit the causal efficacy of mankind, justice, or triangularity. No, in exploring the possibility of a causal role for meaning one is exploring the possibility, not of meaning itself being a cause, but of a *thing's having meaning* being a cause or of the *fact that something has meaning* being a causally relevant fact about the thing. In considering its effect on the glass, is the sound's having a meaning a causally relevant fact about the sound? Is it the sound's *having meaning* that explains, or helps explain, why it broke the glass?

We will see that there are *some* processes—those in which genuine cognitive structures are developed—in which an element's causal role in the overall operation of the system of which it is a part is determined by its indicator properties, by the fact that it carries information. The element

does *this* because it indicates *that*. This connection between a structure's meaning and its causal role, though not direct, is, I shall argue, the connection that underlies the explanatory role of belief. 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.[1]

We are, remember, looking for an *explanatory* role for belief and, hence, an explanatory role for the semantic properties of a structure. If a symbol's meaning is correlated with the symbol's physical properties—if the semantics of symbols is faithfully reflected in their syntax, plus or minus a bit, as Fodor (1980) puts it—then meanings may turn out to be predictively useful without being explanatorily relevant. If I know that the high note is the only passage in the aria that has a certain meaning, I can predict that the glass will shatter when a passage with a certain meaning is sung. The fact that the words have this meaning, however, will not *explain* why the glass shattered. Rather, a sound's having a certain meaning will co-occur with something else (that sound's having a sufficient pitch and amplitude) that *does* explain this physical effect. It may even turn out, if the semantic features co-occur often enough with the right syntactic features, that useful generalizations (useful for *predictive* purposes) can be formulated in semantic terms. It may even be useful, perhaps even essential for methodological purposes, to catalog or index the causally relevant formal properties of our internal states in terms of their causally irrelevant meanings (see, e.g., Loar 1981; Pylyshyn 1984). But this, even if it turns out to be a fact, will not transform meaning into a relevant explanatory notion. If beliefs and desires explain behavior in this way, then *what* we believe and desire (the *content* of our beliefs and desires), however useful it might be for predicting what we are going to do, will not be a part of the explanation of what we do. What will then be relevant are the physical properties of the things that have these meanings, not the fact that they have these meanings. On this account of the explanatory role of meaning, meaning would be as relevant—i.e., wholly *irrelevant*—to explanations of human and animal behavior as it now is to explanations in the science of acoustics.

This, of course, is precisely why computer simulations of mental processes sometimes appear to be more than they are, why it sometimes

that it takes for meaning to explain

1. I will be developing a version of what Stich (1983) calls the *strong* Representational Theory of the Mind. His criticisms of this theory are often based on its uselessness to cognitive science in promoting generalizations about human behavior. Such criticisms of the strong RTM are irrelevant to my project. Ordinary belief (and desire) attribution—what Stich calls Folk Psychology—though it is in the business of *explaining* behavior, is *not* in the business (as is cognitive *science*) of looking for explanations of very general application. I shall return in due course to other, more relevant, criticisms (e.g., the replacement argument) that Stich makes of representational theories.

appears that what a computer does with the symbols it manipulates depends on what these symbols mean. Though it can be disputed, let us agree that the symbols a computer manipulates *have* meanings. If, then, we devise a program for manipulating these symbols that preserves, in some relevant way, the semantic relations between their meanings, it will appear that what these symbols mean makes a difference to what happens to them. It will appear, in other words, that what the computer does—what it displays on the monitor, what it tells the printer to print, or, if we are dealing with a robot, what motors and solenoids it activates—is explicable in terms of the meanings of the elements on which it operates. It will appear, in other words, as though these symbols mean something *to the computer.* The robot went *there* because it *thought* this and *wanted* that. This, of course, is an illusion. It is an illusion that good programming is devoted to fostering. What *explains* why the device printed "Yes" in response to your question is not the fact that the computer knew this, thought that, had those facts in its data base, made these inferences, or indeed understood anything about what was happening. These semantic characterizations of the machine's internal operations may be predictively useful, but only because, by deliberate design, the meanings in question have been assigned to elements which, in virtue of possessing quite different (but appropriately correlated) properties, explain the machine's output. In Dennett's familiar terminology, the modern computer is a machine that is deliberately designed to make adoption of the intentional stance, a stance wherein we ascribe thoughts and desires, a *predictively* useful stance. The mistake lies in thinking that anything is *explained* by adopting this stance towards such machines.[2]

If this is the best that can be done for meaning—and a good many philosophers, for varying reasons and to varying degrees, have concluded that it is (see, e.g., Loar 1981 Fodor 1980, 1987a; Pylyshyn 1984; Stich 1983; Churchland 1981; Dretske 1981[3])—then the case for beliefs and desires as explanatory entities in psychology is exactly as strong as the case for the explanatory role of meaning in the science of acoustics.

2. Searle (1980) has dramatized this point in a useful and (I think) convincing way. Some of Block's (1978) examples make a similar point. Dennett's (1969) distinction between the (mere) storage of information and its intelligent storage makes, I think, basically the same point in a more oblique way. For more on the relevance of meaning to the explanation of machine behavior, see Dretske 1985, 1987; Haugeland 1985; Cummins 1987.

3. In Dretske 1981 I did not think that information, or (more carefully) a signal's carrying information, could itself be a *causally* relevant fact about a signal. I therefore defined the causal efficacy of information (or of a signal's carrying information) in terms of the causal efficacy of those properties of the signal in virtue of which it carried this information. For epistemological purposes (for purposes of defining knowledge) I think this characterization will do, but I no longer think it suffices for understanding the role of belief or meaning in the explanation of behavior. It makes meaning and information, and hence belief, epiphenomenal.

But something better can be done, and it is my purpose in this chapter to do it—to describe the way those relations that underlie an element's meaning, the relations that enable it to *say* something about another situation, figure in the explanation of the containing system's behavior. What we need is an account of the way reasons, in virtue of being reasons, in virtue of standing in semantically relevant relations to other situations, causally explain the behavior that they, in virtue of having this content, help to rationalize.

In pursuit of this end it is important that we avoid effects that are achieved through the mediation of intermediate cognitive processes or agents. So, for example, my automobile's gas tank gets filled with gasoline when I, at the right time and place, make sounds with a certain meaning, when I say "Fill it up, please." If I produce sounds with a substantially different meaning, the tank doesn't get filled. And if, at a different time and place, I produce completely different sounds with the same (or a similar) meaning (e.g., "Benzina, per favore"), the same result is achieved. So it looks like it is not the sounds I produce but their meaning that is having the desired effect. It is *what* I say, not *how* I say it, that explains, or helps to explain, why my gas tank gets filled.

I say we must avoid effects like this. The project is to understand how something's having meaning could itself have a physical effect—the kind of effect (e.g., muscular contraction) required for most forms of behavior—and to understand this *without* enlisting the aid of intelligent homunculi in the head, *without* appealing to hypothetical centers of cognitive activity who, like filling-station attendants, *understand the meaning* of incoming signals. Meaning itself, not some convenient but purely hypothetical understander-of-meaning, has to do the work. To introduce intermediaries who achieve *their* physical effects (on motor neurons, say) by understanding (= knowing the meaning of) the stimuli impinging on them is to interpolate into our solution the very mystery we are seeking to unravel. For to speak of an understander-of-meaning is to speak of something *on which* meaning, and differences in meaning, have an effect. An understander-of-meaning *is* the problem, not something we can use in a solution.

Earlier chapters have put us in a position to confront this problem with some realistic hopes for progress. The chief result of chapters 1 and 2 was that behavior, *what* we are trying to explain when we advert to such content-bearing entities as beliefs and desires, is *not* the physical movements or changes that are the normal *product* of behavior. What we are trying to explain, causally or otherwise, is not why our limbs move but why we move them.

So the explanandum, what is to be explained, is why some process occurred, why (in the case of a structuring cause) M (rather than some

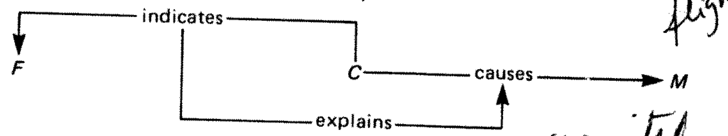

Figure 4.1

other result) is being produced by an internal C. Furthermore, given the results of chapter 3, this causal relationship between C and M, if it is going to be explained by something like the meaning of C, will have to be explained by the fact that C indicates, or has the function of indicating, how things stand elsewhere in the world. It will not be enough merely to have a C that indicates F cause M. We want the fact that it indicates F to be an explanatorily relevant fact about C—the fact about C that explains, or helps explain, *why* it causes M. What needs to be done, then, is to show how the existence of one relationship, the relationship underlying C's semantic character, can explain the existence of another relationship, the causal relationship (between C and M) comprising the behavior in question. With F standing for a condition that C indicates, what we need to show is illustrated in figure 4.1.

Once C is recruited as a cause of M—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 *represent* F. C acquires its semantics, a genuine meaning, at the very moment when a component[4] of its natural meaning (the fact that it indicates F) acquires an explanatory relevance. This, indeed, is why beliefs are maps *by means of which we steer.* An indicator element (such as C) becomes a representation by having part of *what* it indicates (the fact that it indicates F) promoted to an explanatorily relevant fact about itself. A belief is merely an indicator whose natural meaning has been converted into a form of non-natural meaning by being given a job to do in the explanation of behavior. What you believe 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.

What we must do, then, is show how the explanatory relationship depicted in figure 4.1, the relation between C's indicating F and C's causing

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

M, can come about in some natural way. Once this is done, we will have a model of the way beliefs *might* figure in the explanation of behavior—and, hence, a model of the way reasons *could* help to determine what we do. The modesty (reflected in the qualifiers "might" and "could") is necessary because nothing has yet been said about the way *desire* and other motivational states fit into this explanatory picture. We pick up the phone not only because we *think* it is ringing but also because we *want* to answer it when it rings. This is a topic for the following chapter.

Aside from this gap, however, there will doubtless be deeper questions about the adequacy of our account of belief. Even if it can be shown that certain internal indicators can acquire an indicator function, hence a *meaning* or a *content*, in the process by means of which this content is made relevant to the explanation of behavior, it may be wondered whether such simple, almost mechanical, models of belief could ever provide a realistic portrait of the way reasons function in everyday action. Can one really suppose that our ordinary explanations of human behavior have this kind of tinkertoy, push-pull quality to them? Maybe for rats and pigeons it will do, but in explaining a person's weekly attendance at church, the sacrifices of a parent, or an act of revenge are we really talking about the operation of internal indicators? Indicators of what? Salvation? A divine being? An afterlife? Justice?

This challenge—a very serious and understandable challenge, even among those who are otherwise sympathetic to naturalistic accounts of the mind—will be confronted (with what success I leave for others to judge) in the final chapter. What we are after in the present chapter is something less ambitious: an account, however oversimplified and crude it might have to be, of the basic cognitive building blocks. What we are after in this chapter and the next are the elements out of which intentional systems, systems whose behavior can be explained by *reasons*, are constructed. How these basic elements might be combined to give a more realistic portrait of intelligent behavior I leave for later.

4.2 *Why Machines Behave the Way They Do*

To illustrate the structure of relations depicted in figure 4.1, it is useful to begin with simple artifacts. Though instruments and machines don't have beliefs and desires, much less do things *because* of what they believe and desire, they nevertheless *do* things. And *some* of this behavior is explicable, indirectly at least, in a way analogous to the way we explain the behavior of animals. Since these explanations make essential use of the purposes and beliefs of those who construct and use the device, nothing of deep philosophical interest—nothing that helps one understand the ultimate nature of purpose and belief—is revealed by the existence of such explanations.

Nonetheless, there are certain revealing similarities between these explanations and the ones that are of real interest, and it is to highlight these similarities that I begin with these artificial examples.

In an earlier chapter I described the behavior of a thermostat. A drop in room temperature causes a bimetallic strip in this instrument to bend. Depending on the position of an adjustable contact, the bending strip eventually closes an electrical circuit. Current flows to the furnace and ignition occurs. The thermostat's behavior, its turning the furnace on, *is* the bringing about of furnace ignition by events occurring in the thermostat—in this case (it may be different in other thermostats), the closure of a switch by the movement of a temperature-sensitive strip.

In asking why the device turned the furnace on, we are asking why these internal events—whatever, in detail, they happen to be—caused furnace ignition. As we saw in chapter 2, the drop in room temperature, though it caused the bimetallic strip to bend and, in this way, caused the furnace to ignite, and though it may therefore be identified as the *triggering* cause of this process (and, therefore, of the product of this process: furnace ignition), is not the *structuring* cause of this behavior. The drop in room temperature causes a *C which* (given the way things are wired) causes *M*. It, so to speak, initiates a process which has *M* as its outcome. But it does not cause *C* to cause *M*. It does not, therefore, help us to understand why the thermostat behaves this way—why it turns the furnace on rather than, say, opening the garage door or starting the dishwasher.

But if the drop in room temperature is not, in this sense, the cause (the *structuring* cause) of thermostat behavior, if it did not cause the thermostat to turn the furnace on, what did? *We* did. The movement of the bimetallic strip caused furnace ignition because that is the way it was designed, manufactured, and installed. *We* arranged things *so that* the movement of this temperature-sensitive component would, depending on the position of an adjustable setting, close an electrical circuit to the furnace, thereby causing furnace ignition. We wanted furnace ignition to depend on room temperature in some systematic way, so we introduced an appropriate causal intermediary: a switching device that was at the same time a thermometer, something that would cause furnace ignition depending on what it indicated about room temperature. If anyone or anything is responsible for *C*'s causing *M* and, hence, for the thermostat's behaving the way it does, it is we, its creators.

So (referring to figure 4.1) *we* caused *C* to cause *M*. We did so, however, because of some fact about *C*. The bimetallic strip was made into a furnace switch, into a cause of *M*, because it has a special property: its shape varies systematically with, and therefore indicates something about, the temperature. The strip is given a causal role to play, assigned (as it were) control duties in the operation of this thermoregulatory system, because of what it indicates about a certain quantity. Ultimately, then, the strip causes what it does because it indicates what it does.[5]

The bimetallic strip is given a job to do, made part of an electrical switch for the furnace, *because* of what it indicates about room temperature. Since this is so, it thereby acquires the *function* of indicating what the temperature is. We have a representational system of Type II. An internal indicator (of temperature) acquires the function of indicating temperature by being incorporated into a control circuit whose satisfactory operation, turning the furnace on *when* the temperature drops too low, depends on the reliable performance of this component in indicating the temperature.[6] We can speak of (Type II) representation here, and therefore of misrepresentation, but only because the device's internal indicators have been assigned an appropriate *function*: the function of telling the instrument what it needs to know in order to do what it is supposed to do.

In a certain derived sense, then, it is the fact that *C* 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 and confers on it, therefore, the status of a *representation*. An internal indicator acquires genuine (albeit derived) meaning—acquires a *representational* content of Type II—by having its natural meaning, the fact that it indicates *F*, determine its causal role in the production of output. In terms of figure 4.1, the situation looks something like figure 4.2. The indicator relation (between *C* and *F*) becomes the relation of representation insofar as it—the fact that *C* indicates *F*—explains the causal relation between *C* and *M*.

This account of the behavior of a thermostat is infected with intentional and teleological notions, and thus does not represent significant progress in our attempt to understand the causal efficacy of meaning. As figure 4.2 reveals, *C*'s causal efficacy is achieved through the mediation of agents (designers, builders, installers) who *give C* a causal role in the production of *M* because they *recognize C*'s dependence on *F* and *want M* to depend on *F*.

5. I am ignoring the fact that the bimetallic strip is only *part* of the furnace switch, the other part consisting of an adjustable contact point—adjustable to correspond to "desired" temperature (desired by *us*, of course, not the thermostat). In speaking of the cause of furnace ignition, then, there are really two separable factors to be considered: the configuration of the bimetallic strip (representing *actual* temperature) and the position of the adjustable contact point (corresponding to *desired* temperature). I ignore these complications now since I am, for the moment, interested only in developing a model for belief. I will return to this point later when considering the role of desire in the explanation of behavior.

6. See, e.g., Cummins 1975: "When a capacity of a containing system is appropriately explained by analyzing it into a number of other capacities whose programmed exercise yields a manifestation of the analyzed capacity, the analyzing capacities emerge as functions." (p. 407 in Sober 1984b)

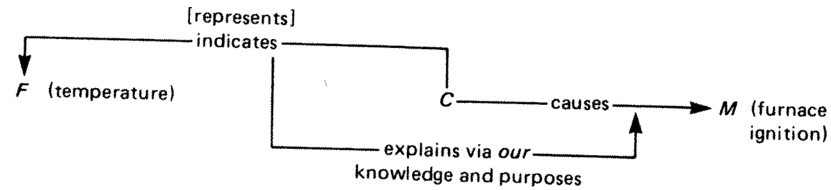

Figure 4.2

The intrusion of *our* purposes into this explanatory story is especially obvious if we consider circumstances in which the designers are confused—circumstances in which *C*, although it does not depend on *F* in the requisite way and therefore does not indicate anything about *F*, is nonetheless *thought* to depend on *F*. If this should occur, there is little question but that *C* would (or might) be given exactly the same causal role to play. In such a case, *C* would not indicate *F*; yet, because of our false beliefs, *C* would still (be made to) cause *M*.

Nevertheless, the case of the thermostat and those of various other control devices are suggestive. They suggest a way that the relations underlying genuine meaning, the indicator relations out of which Type II and Type III representations are fashioned, might figure in the explanation of a state's (*C*'s) acquiring certain control duties and, hence, in the explanation of the behavior (*C*'s causing *M*) of the containing system (the system of which *C* is a part).

It is these suggestive leads that I mean to develop in the rest of this chapter. The idea will be that during 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 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 and, hence, the capacity for misrepresenting how things stand. This, then, is the origin of genuine meaning and, at the same time, an account of the respect in which this meaning is made relevant to behavior.

We can come a bit closer to getting what we want—getting *us* (intentional agents) out of the explanatory picture—by looking at the way detector mechanisms are developed for control purposes in plants and animals. In some of these cases natural selection plays a role similar to that which *we* play with artifacts. The chief difference is that natural selection does not literally *design* a system. There is nothing comparable to a human agent's installing components and assigning control functions because of what things are capable (or what the designer *thinks* they are capable) of doing. For this reason the evolutionary development of control mechanisms, because it gets along without the assistance of any intentional agent, promises to come much closer to our ultimate objective: a completely naturalized account of the explanatory relation illustrated in figure 4.1. It will turn out that this is *still* not quite what we need, but the respects in which it falls short are illuminating.

4.3 *Explaining Instinctive Behavior*

It seems plausible to suppose that certain patterns of behavior—those commonly thought of as instinctive, innate, or genetically determined—involve internal triggering mechanisms that were developed over many generations because of the adaptive advantage of reacting quickly, reliably, and in a stereotypical way to recurring situations. If *M* is always, or almost always, beneficial in conditions *F*, why not hard-wire the system to produce *M* when *F* occurs?

We have already spoken of plant behavior. Some of this behavior depends on the operation of internal indicators. As was noted in chapter 2, it is important that certain trees shed their leaves at the approach of cold, dry weather. In order that this be done in a timely way, it is essential that whatever it is *in* the tree (*C*) that initiates the chemical activity leading to leaf removal (*M*) itself be (or be coupled to) a mechanism sensitive to seasonal changes: perhaps a biological clock of some sort; perhaps a thermal sensor responsive to the gradual temperature gradients characteristic of seasonal change; perhaps a photoreceptor signaling the shortening of days as winter approaches. This is the only way that such activities as dormancy, leaf abscission, and flowering can be synchronized with the external conditions in which these behaviors are beneficial to the plant.

It is interesting in this connection to listen to the biologists Raven, Evert, and Curtis (1981, p. 529) describe a plant's informational needs:

> After periods of ordinary rest, growth resumes when the temperature becomes milder or when water or any other limiting factor becomes available again. A dormant bud or embryo, however, can be "activated" only by certain, often quite precise, environmental cues. This adaptation is of great survival importance to the plant. For example, the buds of plants expand, flowers are formed, and seeds germinate in the spring—*but how do they recognize spring* [my italics—F.D.]? If warm weather alone were enough, in many years all the plants would flower and all the seedlings would start to grow during Indian summer, only to be destroyed by the winter frost. The same could be said

> for any one of the warm spells that often punctuate the winter season. The dormant seed or bud does not respond to these apparently favorable conditions because of endogenous inhibitors which must first be removed or neutralized before the period of dormancy can be terminated.

In such cases it seems reasonable to suppose that whatever it is in the plant that causes the buds to expand, the flowers to form, and the seeds to germinate in the spring is something that was selected for this job *because* it tended to occur at the right time, *when* the plant profited from the kind of activity (growth, germination, etc.) that it brought about. In other words, the chemical trigger for growth, germination, flowering and leaf removal was selected for its job, over many generations, because of its more or less reliable[7] correlation with the time of year in which this activity was most beneficial to the plant. Here again we find a structure's causal role in the production of output explained, in part at least, by its indicator properties.

We earlier saw how predaceous fungi capture, kill, and consume (eat?) small insects and worms. The mechanisms these plants use to trap their prey embody sensitive indicators (*C*) of movement (*F*). These indicators, once activated by movement, cause a rapid swelling (*M*) of a ring that "grasps" or "holds" the prey. More sophisticated plants have more discriminating sensors. The Venus flytrap, for instance, comes equipped with sensitive hairs on each half-leaf. When an insect walks on the leaf, it brushes against these hairs, triggering a traplike closing of the leaves. The leaf halves squeeze shut, pressing the insect against the digestive glands on the inner surfaces of the leaves. This trapping mechanism is so specialized that it can distinguish between living prey and inanimate objects, such as pebbles and small sticks, that fall on the leaf by chance. Once again, leaf movement (*M*) is caused by an internal state (*C*) that signals the occurrence of a particular kind of movement, the kind of movement that is normally produced by some digestible prey. And there is every reason to think that this internal trigger was selected for its job *because* of what it indicated, because it "told" the plant what it needed to know (i.e., *when* to close its leaves) in order to more effectively capture prey.

7. Elliott Sober has pointed out to me that for selection to take place all that is needed is for the triggering state to be *better* correlated with the appropriate season than are the corresponding states in competing plants. A state need not be reliably correlated with spring—hence, need not indicate the arrival of spring—in order to be correlated sufficiently well with the arrival of spring to confer on its possessor a competitive advantage. In cases where the correlation (with spring) is not of a sort to support the claim that there is an indication of spring, there will always be an indication of something (e.g., an interval of mild weather) which will (via its past correlation with the arrival of spring) explain its selection. The indicator properties are still relevant to the thing's selection, just not its indication of *spring*. I return to this point in section 4.4.

Explaining a plant's behavior (its closing its leaves, trapping an insect, or strangling a nematode) by describing the event that, by activating the internal indicator, brings about leaf movement, enclosure of insect, or strangulation of nematode, is merely a way of describing the triggering cause of the plant's behavior: the condition (*F*) the internal indication of which (by *C*) led (presumably by natural selection) to *C*'s causing *M*. But though the movement of an insect on the plant's leaves triggers a process that culminates in closure of the leaves (*M*), it does not explain why the process has this, rather than another, outcome. If we want a *structuring* cause of plant behavior, an explanation of why the plant did *this* then, rather than an explanation of why it did this *then*, we have to look for the cause, not of *C*, not of *M*, but of *C*'s causing *M*. And here, just as in the case of the thermostat, we find the explanation coming back to some fact about *C*. It is a fact about *C*'s status as an indicator—the fact that it registers the occurrence of a certain kind of movement, the kind of movement that is usually (or often, or often *enough*) made by a digestible insect—that explains why, over many generations, *C* was selected, installed, or made into a cause of *M*. Because *M* is beneficial to the plant when it occurs in conditions *F* (but not generally otherwise), some indicator of *F* was given the job of producing *M*. It is this fact about *C* that explains, via natural selection, its current role in controlling leaf movement in the same way a corresponding fact about the bimetallic strip in a thermostat explains, via the purposes of its designers, its causal role in regulating a furnace.

As with plants, so with animals. The noctuid moth's auditory system is obviously designed with its chief predator, the bat, in mind. The moth's ear does not relay information about a host of acoustical stimuli that are audible to other animals. Prolonged steady sounds, for example, elicit no response in the receptor. The bat emits *bursts* of high-frequency sound, which are what the moth's receptors are "designed" to pick up and respond to. The moth's ear has one task of paramount and overriding importance (Alcock 1984, p. 133): the detection of cues associated with its nocturnal enemy. And its behavioral repertoire is equally constrained and simple: it turns away from low-intensity ultrasound (the bat at a distance) and dives, flips, or spirals erratically to high-intensity ultrasound (the bat closing in).

Why did the moth's nervous system develop in this way? Why did it inherit neural wiring of this sort, wiring that automatically adjusts the moth's orientation (relative to the incoming sound) and, hence, its direction of movement so as effectively to avoid contact with the source of that sound? The answer, obviously, is to enable moths to avoid bats. Inspection of the comparatively simple wiring diagram of the moth's central nervous system reveals that the motor neurons that adjust orientation, and hence the moth's direction of movement (*M*), are controlled, through a network of interneurons, by structures that indicate the *location* (distance and direction) of the sound source (*F*). What the theory of evolution has to tell us about these cases (and these cases are typical of motor control systems throughout the animal kingdom) is that *C*'s production of *M* is, at least in part, the result of its indication of *F*. *M* is produced by an indicator of *F* because such an arrangement confers a competitive advantage on its possessor. If you want *M* to occur in conditions *F* but not generally otherwise, and if *F*, left to its own devices, won't produce *M*, then the best strategy (indeed the only strategy) is to make an indicator of *F* into a cause of *M*. If the organism already has an indicator of *F*, *make it* into a cause of *M*. If it doesn't have such an indicator, *give it* one. This is the course that engineers follow in designing control systems such as the thermostat. It is also the course that nature takes, in its own nonpurposeful way, in the design of plants and animals.

Though the evolutionary development of control systems for the instinctive or innate behavior of animals does not, like figure 4.2, involve an interpolated *agent*, it nonetheless fails to meet the explanatory requirements of figure 4.1 for another reason. As Cummins (1975) notes, natural selection (assuming this is the chief pressure for evolutionary change) does not explain why organisms *have* the properties for which they are selected any more than Clyde's preference for redheads explains why Doris, his current favorite, has red hair. It is, if anything, the other way around: her having red hair explains why Clyde selected her. The neural circuitry in a particular moth, the connections in virtue of which an internal sign of an approaching bat causes evasive wing movements, is, like other phenotypical structures, to be causally explained by the genes the moth inherited from its ancestors. This isn't to suggest that there is a sharp distinction between nature and nurture, between genetic and environmental determinants of behavior, but it is to suggest that the explanation for the control circuitry in *this* moth—the explanation for why *this* *C* is causing *this* *M*, why the moth is now executing evasive maneuvers—has nothing to do with what *this* *C* indicates about this moth's surroundings. The explanation lies in the moth's genes. They (given anything like normal conditions for development) determine that *C*, *whatever* it in fact happens to indicate about the moth's surroundings, will produce *M*.

Elliott Sober (1984a, pp. 147–152), applying a distinction of Richard Lewontin (1983), contrasts selectional explanations with developmental explanations. In explaining why all the children in a room read at the third-grade level (Sober's example), one explains it developmentally by explaining why each and every child in the room reads at this level. Or one can explain it selectionally by saying that *only* children reading at the third-grade level were allowed in the room (selected for admission into the room). The latter explanation does not tell us why Sam, Aaron, Marisa, et

al. read at the third-grade level. In effect, it tells us why all of them read at the third-grade level without telling us why any one of them reads at that level. Sober correctly diagnoses this difference in explanatory effect by pointing out that the difference between a selectional and a developmental explanation of why all the children in the room read at the third-grade level is a *contrastive phenomenon* (Dretske 1973; Garfinkel 1981). It is, in effect, the difference between explaining why (all) my friends imbibe martinis, an explanation that requires my telling you something about them, and explaining why I have (only) martini imbibers as friends, an explanation that requires my telling you something about me.

The moth has the kind of nervous system it has, the kind in which an internal representation of an approaching bat causes evasive movements, because it developed from a fertilized egg which contained genetic instructions for this kind of neural circuitry, circuitry *in which* the occurrence of *C* will cause *M*. This is a developmental explanation, a causal explanation of why, in today's moths, tokens of type *C* produce movements of type *M*. These genetically coded instructions regulated the way in which development occurred, channeling the proliferation and specialization of cells along pathways that produced a nervous system with these special features. Even if through a recent freak of nature (recent enough so that selectional pressures had no time to operate) the occurrence of *C* in contemporary moths were to signal not the approach of a hungry bat but the arrival of a receptive mate, *C* would still produce *M*—would still produce the same evasive flight manuevers. What *C* indicates in *today's* moths has nothing to do with the explanation of what movements it helps to produce. And the fact that tokens of *C* indicated in remote ancestors the approach of hungry bats does not explain—at least not causally (developmentally)—why *this* (or indeed, why *any*) *C* produces *M*. Rather, it explains (selectionally) why there are, today, predominantly moths in which *C* causes *M*.

The moth's behavior is, like so much of the behavior of simple organisms, tropistic. Tropisms are simple mechanical or chemical feedback processes or combinations of such processes that have the interesting property of looking like organized motivated behavior. According to Jacques Loeb (1918), who first described tropisms in plants and simple animals, the working of all tropisms can be explained with two principles: symmetry and sensitivity. Caterpillars emerge from their cocoons in the spring, climb to the tips of tree branches, and eat the new buds. This apparently purposeful behavior has a simple explanation in terms of Loeb's two principles. Rachlin (1976, pp. 125–126) describes it thus:

> The caterpillars are sensitive to light and have two eyes, symmetrically placed one on each side of the head. When the same amount of light comes into the two eyes, the caterpillars move straight ahead;

but when one of the eyes gets more light, the legs on that side move more slowly. The result is that the caterpillars tend to orient toward the light—which in nature invariably is strongest at the tops of trees. Thus, whenever they move, they move toward the tops of the trees, ending up at the tip of a branch. When, in his experiments, Loeb put lights at the bottom of the trees, the caterpillars went down, not up, and would starve to death rather than reverse direction. When the caterpillars were blinded in one eye, they traveled in a circle like a mechanical toy with one wheel broken.

A symmetrical placement of light-sensitive indicators, each indicator harnessed to an appropriate set of effectors, is capable of explaining most of this behavior. Though a plant doesn't have a nervous system, similar mechanisms help explain the climbing behavior of some plants. And they are equally at work in guiding the moth away from the bat.

Such tropistic behavior has a rather simple mechanical basis. And the blueprint for the processes underlying this behavior is genetically coded. The behavior is instinctive—i.e., not modifiable by learning. But it is not the simplicity of its explanation that disqualifies such behavior from being the behavior of interest in this study. *Reasons* are irrelevant to the explanation of this behavior, not because there is an underlying chemical and mechanical explanation for the movements in question (there is, presumably, some underlying chemical and mechanical explanation for the movements associated with all behavior), but because, although indicators are involved in the production of this movement, *what they indicate*—the fact that they indicate thus and so—is (and was) irrelevant to what movements they produce. If we suppose that, through selection, an internal indicator acquired (over many generations) a biological function, the function to indicate something about the animal's surroundings, then we can say that this internal structure *represents* (or *misrepresents*, as the case may be) external affairs. This is, in fact, a representation of Type III. But it is *not* a belief. For to qualify as a belief it is not enough to *be* an internal representation (a map) that is among the causes of output, something that helps us steer. *The fact that it is a map*, the fact that it *says* something about external conditions, must be relevantly engaged in the way it steers us through these conditions. What is required, in addition, and in accordance with figure 4.1, is that the structure's indicator properties figure in the explanation of its causal properties, that what it *says* (about external affairs) helps to explain what it *does* (in the production of output). That is what is missing in the case of reflexes, tropisms, and other instinctive behaviors. Meaning, though it is there, is not relevantly *engaged* in the production of output. The system doesn't do what it does, *C* doesn't cause *M*, *because* of what *C* (or anything else) means or indicates about external conditions. Though *C* has

meaning of the relevant kind, this is not a meaning it has *to* or *for* the animal in which it occurs. That, basically, is why genetically determined behaviors are not explicable in terms of the actor's reasons. That is why they are not *actions*. What (if anything) one wants, believes, and intends is irrelevant to what one does.

The distinction between developmental and selectional explanations is not, therefore, merely the difference in what behaviorial biologists call *proximate* factors and *ultimate* factors (Alcock 1984, p. 3; Grier 1984, p. 21). What they mean by ultimate factors (the selectional explanations one finds in sociobiological "explanations" of behavior, for instance) are not factors that figure in the causal explanation, proximate *or* remote, of the behavior of any individual. In such cases an internal state, *C*, which means (indicates) that a hungry bat is approaching and which even (let us say) has the *function* of indicating this (in virtue, let us suppose, of its evolutionary development in this *kind* of moth), *does*, to be sure, cause orientation and wing movements of an appropriate (evasive) sort. *C* (something that indicates the approach of a bat) causes *M* (bat-avoidance movements). Nevertheless, it is not *C*'s meaning what it does (*F*) that explains why it causes this (*M*). In this case the internal state *has* a semantics—something it is (given its evolutionary development) *supposed* to indicate—but the fact that it indicates this, or is supposed to indicate this, is irrelevant to an understanding of why it actually does what it does. A selectional explanation of behavior is no more an explanation of an individual organism's behavior—why *this* (or indeed *any*) moth takes a nosedive when a bat is closing in—than is a selectional account of the antisocial behavior of prison inmates an explanation of why Lefty forges checks, Harry robs banks, and Moe steals cars. The fact that we imprison people who forge checks, steal cars, and rob banks does not explain why the people in prison do these things.

4.4 *Putting Information to Work: Learning*

To find a genuine case where an element's semantic character helps to determine its causal role in the production of output—a case where what the (internal) map *says* helps explain what kind of (external) effects the map has—one must look to systems whose control structures are actually shaped by the kind of dependency relations that exist between internal and external conditions. The places to look for these cases are places where individual learning is occurring, places where internal states *acquire* control duties or *change* their effect on motor output as a result of their relation to the circumstances on which the success of this output depends.

There are many forms of learning, or what generally passes as learning, that have little or nothing to do with the meaning, if any, of internal states.

If learning is understood, as it sometimes is, as *any* change in behavior (or, perhaps, any *useful* change of behavior) brought about by experience, then habituation and sensitization may qualify as elementary forms of learning. Roughly speaking, habituation is a decrease, and sensitization an increase, in response to a repetitive stimulus. Such changes are often mediated by relatively peripheral mechanisms. For example, the change in movements produced by a certain stimulus may be due entirely to receptor (or muscle) fatigue. It seems fairly clear that if there are internal maps that help us steer, one isn't likely to find them playing a significant role in explaining the behavior resulting from changes of this kind.[8]

It is only when we get to a form of learning whose success depends on the deployment and the use of internal indicators that it becomes plausible to think that the causal processes constitutive of behavior may actually be explained by facts about what these indicators indicate. And this means that we must look to kinds of learning in which the *correlations* (*contingencies*, as they are sometimes called) underlying the indicator relationship play a prominent role. We must look, in other words, to certain forms of associative learning if we are to find the kind of explanatory relationship depicted in figure 4.1. Only (but, as it turns out, not always) in this kind of learning do we find internal states assuming control functions *because* of what they indicate about the conditions in which behavior occurs. Only here do we find *information*, and not merely the structures that carry or embody information, being put to work in the production and the control of behavior.

Consider the following common problem, whose general form I shall call The Design Problem: We want a system that will do M when, but only when, conditions F exist.[9] How do we build it? Or, if we are talking about an already existing system, how do we get it to behave in this way?

In very general terms, the solution to The Design Problem is always the

8. Staddon (1983, p. 2) sees no hard and fast line separating learning from other kinds of behavioral change: "... we do not really know what learning is." Experience can change behavior in many ways that manifestly do not involve learning: "... a change brought about by physical injury or restraint, by fatigue or by illness doesn't count. Short-term changes, such as those termed *habituation*, *adaptation*, or *sensitization*, are also excluded—the change wrought must be relatively permanent. *Forgetting* has an ambiguous status: The change is usually permanent and does not fall into any of the forbidden categories, yet it is paradoxical to call forgetting an example of learning. Evidently it is not just *any* quasi-permanent change that qualifies. *Learning* is a category defined largely by exclusion." (ibid., pp. 395–396)

9. In order to minimize the use of symbols I will hereafter (in this and later chapters) let "M" do double duty. I shall, as before, let it stand for some external movement; but I shall also let it stand for behavior, the process of producing movement. It will, I hope, always be clear which is intended. When I speak of behavior M, or of someone's *doing* M, I should be understood as referring to *the production of* M (by some internal state C).

same. Whether it is the deliberate creation of an engineer, the product of evolutionary development, or the outcome of individual learning, the system S must embody, and if it doesn't already embody it must be supplied with, some kind of internal mechanism that is selectively sensitive to the presence or absence of condition F. It must be equipped with something that will indicate or register the presence of those conditions with which behavior is to be coordinated. We have already taken note of the way this works with artifacts: If you want a device that will turn the furnace on when the temperature gets too low, (a particular instance of The Design Problem), this device must be supplied with a temperature indicator. We have also noted how it works with instinctive behavior: If you want young animals to stop or change direction when they encounter cliffs, they must, sooner or later, be supplied with a mechanism sensitive to steep (downward) depth gradients—a "cliff" indicator. If you want chickens to hide from hawks (another instance of The Design Problem), you have to give them an internal hawk indicator, or at least an indicator of something (e.g., a certain silhouette in the sky) that is sufficiently well correlated with the approach of a hawk to make concealment a beneficial response when there is a positive indication. The same is true of learning. If you want a rat to press a bar when and only when a certain tone is heard, a pigeon to peck a target when and only when a light is red, or a child to say "Mommy" to and only to Mommy, then the rat needs a tone indicator, the bird a color indicator, and the child a Mommy indicator. Only if such indicators exist is it possible to solve The Design Problem. You can't get a system to do M in conditions F unless there is something in it to indicate when these conditions exist.

In the case of learning, this is merely to say that you must begin with a system that has the appropriate sensory capacities. The system must have a way of getting the information that condition F obtains if it is going to learn to do M in conditions F. The rat must be able to hear, able to distinguish one tone from another, if it is to learn to respond in some distinctive way to a particular tone. The pigeon must be able to *see*, to distinguish visually, one color from another if it is to learn to peck when the light is red. The child must be able to see Mommy, or at least sense her presence in some way, before she can be taught to say "Mommy" when Mommy is present. If Mommy has a twin sister who regularly babysits for the child, this learning is going to be impaired or, depending on the degree of resemblance, impossible. It will be slower because the infant's Mommy detector has been neutralized by the presence of the twin. If the child's powers of discrimination are such that she cannot tell the difference between Mommy and Auntie, the child *cannot* learn to say "Mommy" in the prescribed way (i.e., only to Mommy), for she no longer has a Mommy indicator. It would be like trying to teach a tone-deaf rat to respond to middle C or a color-blind bird to peck at red targets.

So the first requirement for a solution to The Design Problem is that the system be equipped with an F indicator. Once this requirement is satisfied, all that remains to be done is to harness this indicator to effector mechanisms in such a way that appropriate movements (M) are produced when and only when the indicator positively registers the presence of condition F. This is something the engineer accomplishes by soldering wires in the right places. This is something nature accomplishes in the case of instinctive behavior by selecting systems whose wires are already secured, if not soldered, in the right place (or, if not in the right place, at least in a place that is more nearly right—a place that confers on its possessor a competitive advantage). And, finally, this is something that is accomplished in certain forms of learning by the kind of *consequences* attending the production of M.

By the timely reinforcement of certain output—by rewarding this output *when*, and generally *only when*, it occurs in certain conditions—internal indicators of these conditions are recruited as causes of this output.[10] Just *how* they are recruited by this process may be (and to me is) a complete mystery. The parallel distributed processing (PDP) networks, networks of interconnected nodes in which the strength of connections between nodes is continually reweighted (during "learning") so that, eventually, given inputs will yield desired outputs, provide intriguing and suggestive models for this recruitment process (Hinton and Anderson 1981; McClelland and Rumelhart 1985). In these models, the internal indicators would be patterns of activation of the network's input nodes, and recruitment would proceed by selection (by appropriate reweighting between nodes) of the desired input (i.e., an F indicator) for an appropriate activation of effector mechanisms (M). But no matter how the nervous system manages to accomplish this trick, the fact that it does accomplish it, for many animals and for a variety of different behaviors, is obvious. Learning cannot take place *unless* internal indicators of F are harnessed to effector mechanisms in some appropriate way. Since this learning *does* occur, the recruitment *must* take place. These internal indicators are assigned a job to do in the production of bodily movement—they get their hands on the steering wheel (so to

10. It sounds a little odd to say that the indicators are *recruited* for this job if they are, for whatever reason, *already* serving as causes of the appropriate movements. Though this seems improbable for *learned* behaviors, the behaviors we are presently concerned with, the possibility figures in some philosophical thought experiments—e.g., Stich's (1983) Replacement Argument and Davidson's (1987) Swampman. If, however, the *continued* service of an indicator (as a cause of a movement) depends on the occurrence of reinforcement, I shall, for purposes of brevity, speak of this as recruitment. I am grateful to Dugald Owen for discussion on this point.

speak)—in virtue of what they "say" (indicate or mean) about the conditions in which these movements have beneficial or desirable consequences. Since these indicators are recruited for control duties *because* of the information they supply, supplying this information becomes part of their job description—part of what they, once recruited, are *supposed to do*.

Just as our incorporation of a bimetallic strip into a furnace switch *because* of what it indicates about temperature gives this element the function (Type II) of indicating what the temperature is, the reorganization of control circuits occurring during learning, by converting internal elements into "movement switches" in virtue of what they indicate about environmental conditions, confers on these elements the function (Type III) of indicating whatever it is that brought about their conversion to switches. As a result, learning of this sort accomplishes two things: it reorganizes control circuits so as to incorporate indicators into the chain of command, and it does so *because* these indicators indicate what they do. Learning of this sort mobilizes information-carrying structures for control duties *in virtue of the information they carry*. In bringing about this transformation, learning not only confers a function on these indicators, and thereby a *meaning*, but also shapes their causal role, and hence the behavior of the system of which they are a part, in terms of *what they mean*—in terms of the information they now have the function of providing. Such learning *creates* maps at the same time it gives these maps, *qua* maps, a job to do in steering the vehicle.

The kind of learning we are talking about is a special form of *operant* or *instrumental* learning, a kind of learning sometimes called *discrimination* learning. One learns to identify F, or at least to distinguish (discriminate) F from other conditions, by having particular responses to F (or particular responses *in* condition F) rewarded[11] in some special way. The literature on instrumental conditioning, not to mention that on learning theory in general, is enormous. Fortunately, not all this material is relevant to the present point. We need only two facts, both of which are (as facts go in this area) relatively unproblematic.

First, there is Thorndike's Law of Effect, which tells us that successful behavior tends to be repeated (Rachlin 1976, pp. 228–235). More technically, a reward (alternatively, a positive reinforcement) increases the probability that the response that generates it (or with which it co-occurs) will occur again in the same circumstances.

It isn't particularly important for my purposes (though it certainly may be for other purposes) whether we think of rewards as stimuli (e.g., food)

11. Learning theorists typically distinguish between rewards (e.g., the delivery of food) and reinforcement (and effect of the reward on the organism). Unless these differences are important to the point I am making, I shall ignore them and use these terms interchangeably.

or as responses (e.g., *eating* the food). One can even think of them as the *pleasures* (need or tension reduction) that certain stimuli (or responses) bring to an organism.

Neither is it important that we get clear about the exact status of this law. There have been deep (and often legitimate) suspicions about the empirical significance of this law (see, e.g., Postman 1947; Meehl 1950). Unless there is available some *independent* specification of what a reward or reinforcer is—independent, that is, of its effect on the probability of a response—the law seems devoid of empirical content. It becomes a mere tautology: results that tend to increase the probability of behavior tend to increase the probability of that behavior. There is also disagreement about exactly how the reward must be related to the response it strengthens (temporal contiguity? mere correlation?) and about the "associability" of some response-reinforcement pairs (Garcia and Koelling 1966). The latter issue raises questions about the scope of this law—whether, indeed, it is applicable in every situation. Even if cookies reinforce some behavior, they surely will not be equally effective for all behavior. A child might eat her vegetables to get a cookie but refuse to walk on hot coals for the same reward. Finally, Premack (1959, 1965) has argued persuasively for the relative nature of the concept of reinforcement, i.e., that reward and punishment are determined by relations between events in a "value" hierarchy. Any event in this hierarchy (as long as there is a lower event) can be a reward, and any event (as long as their is a higher one) can be a punisher. The critical relationship is the contingency of one event on the other. When a higher event is contingent on the occurrence of a lower event, the higher event serves as a reward and the lower event becomes reinforced. When a lower event is contingent on a higher event, the lower event serves as a punisher and the higher event is punished.

Serious and important as some of these issues are, they are not directly relevant to the way I propose to use this law. What is important is that *something* (call it what you will), *when* it occurs in the right relationship (whatever, exactly, that might be) to behavior performed in certain stimulus conditions, tends (for *some* behavior and *some* stimulus conditions) to increase the chances that that behavior will be repeated in those conditions. There are *some* consequences of *some* behaviors of *some* organisms that are causally relevant to the likelihood that such behaviors will be repeated in similar circumstances.[12]

12. It is especially important to understand that what is changing during learning of this sort is *behavior* (a bringing about of some result or condition), *not* some particular *way* of producing that result (e.g., some particular bodily movement). So, for instance, if going to (or avoiding) place *P* is the behavior reinforced, what is reinforced is (roughly speaking) a process having *occupation* (or *non-occupation*) *of place P* as its product. Since (see chapters 1 and 2) *any* process having this product is the *same* behavior, this behavior can be realized in

Second, we need the fact that such learning requires, on the part of the learner, a sensitivity to specific conditions *F*. Rewards tend to increase the probability that *M* will be produced *in conditions F*. Whether the rewards are administered by a teacher or by nature, making the rewards dependent (in some way) on the existence of special conditions increases the probability of the response in those special conditions. Hence, if learning is to occur, there must be something *in* the animal to "tell" it when conditions *F* exist.

Given these two facts, it follows that when learning of this simple kind occurs, those results (bodily movements or the more remote effects of bodily movements) that are constitutive of the reinforced behavior are gradually brought under the control of internal indicators (*C*), which indicate *when* stimulus conditions are right (*F*) for the production of those results. Making reinforcement of *M* contingent on the presence of *F* is a way of solving The Design Problem. It solves The Design Problem (for those creatures capable of this kind of learning) by promoting *C*, an internal indicator of *F*, into a cause of *M*. *C* is recruited as a cause of *M because* of what it indicates about *F*, the conditions on which the success of *M* depends. Learning of this sort is a way of shaping a structure's causal properties in accordance with its indicator properties. *C* is, so to speak, *selected* as a cause of *M* because of what it indicates about *F*. Unless this is done, The Design Problem cannot be solved. Learning cannot take place. An animal cannot learn to behave in the prescribed way—it cannot learn to coordinate its output (*M*) with condition *F*—unless an internal indicator of *F* is made into a cause of, a switch for, *M*. This is why learning of this sort must recruit indicators of *F* as causes of *M*.

During this process, *C* becomes a cause of *M*. It gets its hand on the steering wheel (if not for the first time, at least in a new way[13]) *because* of what it indicates about *F*. *C* thereby becomes a representation of *F*. After learning of this sort, the bird pecks the target because it *thinks* (whether

many different bodily movements (e.g., in the case of avoidance learning, flight *from* place *P* during learning or *avoidance of place P* after learning).

I think it was Taylor's (1964) failure to appreciate this point about the structure of behavior, about *what* was being reinforced, that led him to criticize (pp. 250ff.) the possibility of avoidance behavior as an operantly conditioned response. I shall return to this important point, and to a fuller discussion of the plasticity of behavior, in chapter 5.

13. I postpone until the last chapter (section 6.4) a discussion of the possibly multiple indicator functions an element might acquire in learning. That is, an element originally recruited to do one thing because of what it indicated about *F* might be recruited to do other things because of this same fact, or recruited to do other things because of what it indicated about some associated conditions *G*. Such developments require at least a preliminary understanding of the way motivational factors contribute to the explanation of behavior, a matter to be discussed in chapter 5.

rightly or not) that the light is red. Or, if one is skittish about giving beliefs to birds, if one thinks that the word "belief" should be reserved for the elements in larger representational networks, the bird pecks the target because it *represents* (whether rightly or not) there being a red light. This explanatory relation, the fact that the bird's behavior is explained (in part at least) by the way it represents the stimulus, derives from the role this internal indicator, and *what* it indicates, played in structuring the process ($C \rightarrow M$) which *is* the behavior. C now causes M; but what explains why it causes M, and therefore explains why the bird *behaves* the way it does, is the fact that C indicated F—the fact that C did what it now has the function of doing. If, before learning, C happened to cause M, or if M was merely produced when C happened to be registering positive, then the bird pecked the target *when* the light was red, but it did not peck the target *because* the light was red. The fact that the light was red does not explain the earlier (prior to learning) behavior of the bird because, prior to learning, even if C happened to cause M, the fact that C indicated that the light was red did not *explain why* it caused M. This was, rather, a chance or random connection between C and M. The bird was just poking around. It is only after learning takes place that facts about the color of the light figure in the explanation of the bird's behavior, and this is so because, after learning, an internal element produces M precisely *because* it indicates something about the light's color.

If we have a system that lacks an internal indicator for condition F, a temporary solution to The Design Problem can nonetheless be reached *if* there is an internal indicator of some condition which, through coincidence, temporary arrangement (by an experimenter, say), or circumstances of habitat, is correlated with F. Suppose, for instance, that the animal has no detector for F (the condition on which the arrival of food is actually dependent) but does have a detector for G. If the animal is placed in circumstances in which all, most, or many G's are F, then the internal indicator of G will naturally be recruited as a cause of M (the movements that are rewarded by food in condition F). The animal will learn to produce M when it senses G. Its G indicator will be converted into a cause of M, and the explanation of this conversion will be the fact that it indicates G (and, of course, the fact that, for whatever reason, G is temporarily correlated with F). An internal representation of G develops because the internal indicator of G is given its job in the production of output because of what it indicates about external affairs. Depending on the degree of correlation between F and G, this will be a more or less effective solution to The Design Problem. The better the correlation, the more successful the animal will be in producing M in conditions F (and, therefore, in getting whatever reward it is that promotes that response).

If the correlation (however temporary) between F and G is perfect, this

solution to The Design Problem will (for however long the correlation persists) be indistinguishable from the original solution, the solution by a system that has an F indicator. But the explanation of the resultant behavior of these two systems will be different. Using the intentional idiom to describe this case, we say that the second animal produces M in conditions F, not because it thinks that F exists, but because it thinks G exists (and, of course, thinks that doing M in conditions G will get it food—more of this in chapter 5). The second animal has a set of beliefs that are temporarily effective in securing food, but whose effectiveness depends on the continuation of an external correlation between F and G, a correlation which the animal itself (having no way of representing F) has no way of representing. This is the situation of rats and pigeons subjected to experiments in discrimination learning. Their internal indicators for rather simple stimuli—the patterns of color and sound they are being taught to discriminate—are enlisted as causes of movement because of a temporary contingency, instituted and maintained by the investigator, between these discriminable stimuli and rewards. Once the training is over, the correlations are suspended (or reversed) and the animal's "expectations" (that doing M in conditions G will get it food) are disappointed.

If the correlations between F and G are reasonably secure, as they often are in an animal's natural habitat, it may be more economical to solve The Design Problem by exploiting a simpler and less costly G indicator than to waste resources on a more complicated F indicator. Engineers do it in their design of machines, nature does it in the design of sensory systems and instinctive patterns of behavior, and individuals do it in developing, through learning, the cognitive rules of thumb for negotiating their way through complex situations. In the case of nature, we know from Tinbergen's (1952) studies that stickleback rely on what Tinbergen calls "sign stimuli." The fish exploit rather crude indicators (a bright red underside, for instance) to recognize one another. Males use the bright red underside to recognize male intruders, and females use it to identify interested males. The fish react similarly to a variety of objects of similar coloration: painted pieces of wood elicit aggressive behavior in the males and sexual interest in the females. But in the fish's natural habitat the correlation is good enough. By and large, *only* stickleback have this coloration. So why develop more expensive receptor hardware for representing conspecifics *as* conspecifics (i.e., as stickleback) when representing them *as* objects with a red underside works well enough? The same economy of effort is evident, as it should be, in individual learning. The Design Problem is solved with whatever resources are available for its solution. If there is no F indicator to convert into a cause of M, there are less optimal solutions. A G indicator will be enlisted *if* G exhibits *enough* correlation with F to make it a useful switch for

M. How much is "enough" depends on the energy required to produce *M* and the consequences of producing *M* when *F* does not exist.

Some animals exhibit a plasticity, a susceptibility, a disposition to have their control processes reconfigured by their experience of the world. As we move up the phylogenetic scale, we find that the behavior of an animal is shaped, not primarily by its genes, but, in larger and larger measure, by the contingencies that dominate the environment in which it lives. Staddon (1983, p. 395) writes:

> Most animals are small and do not live long; flies, fleas, bugs, nematodes, and similar modest creatures comprise most of the fauna of the planet. A small, brief animal has little reason to evolve much learning ability. Because it is small, it can have little of the complex neural apparatus needed; because it is short-lived, it has little time to *exploit* what it learns. Life is a tradeoff between spending time and energy learning new things, and exploiting things already known. The longer an animal's life span, and the more varied its niche, the more worthwhile it is to spend time learning.... It is no surpise, therefore, that learning plays a rather small part in the lives of most animals.... Learning is interesting for other reasons: It is involved in most behavior we would call intelligent, and it is central to the behavior of people.

The reason learning is so central to *intelligent* behavior, to the behavior of *people*, is that learning is the process in which internal indicators (and also, as we shall see in the next chapter, various motivational factors) are harnessed to output and thus become relevant—as representations, as *reasons*—to the explanation of the behavior of which they are a part. It is in the learning process that information-carrying elements get a job to do *because* of the information they carry and hence acquire, by means of their *content*, a role in the explanation of behavior.

It should be apparent that *C*, the internal indicator that is recruited as a cause of *M* during this kind of learning, could have any shape, form, or physical realization. As long as it is the sort of structure that *could* affect *M* (and hence could be recruited as a cause of *M*), what is important about it is not its neurophysiological character, its *form* or *shape*, but the fact that it stands in certain *relations* to those external affairs (*F*) on which the beneficial consequences of *M* depend. It is *what* information *C* carries, not *how* it carries it, that explains its newly acquired causal powers and, hence, the altered behavior of the system of which it is a part. This system's control circuits were reconfigured—*C* was given command duties (or at least given access to those mechanisms having command functions)—*because* it *told* the system what it needed to know. In the business of espionage, informants are recruited because of what they know or are capable of finding out. As long as the way they talk, look, or dress doesn't interfere with their information-gathering and communication functions, details about *how* they do their job are irrelevant. The same is true of an animal's behavior-guidance systems. It is the *semantic*, not the syntactic properties of these internal elements that explain their impact on behavior, and it is for basically this reason that a syntactic theory of the mind (Stich 1983) is unsatisfactory.[14]

As we shall see more fully in chapter 5, it would be wrong to say that, as a result of this kind of learning, *C*'s function is to produce *M*, or even to produce *M when F* obtains. What this kind of learning confers on *C* is an indicator function: the function of indicating when *F* exists. *C*'s function is not to produce *M*. The production of *M* depends not only on *C*, not only on a certain positive *cognitive* state, but also on the right *motivational* or *conative* conditions. The animal must have a *desire* for whatever reward or reinforcement promoted *C* into a cause of *M*. If a rat isn't hungry, it isn't going to behave in the way it was trained to behave on the appearance of the discriminative stimulus. If it isn't hungry, *C* won't cause *M*. The rat won't press the bar. So the function of *C* is not to cause *M*, but to indicate the presence of those conditions that, if the right motivational state is present, will lead, other things being equal, to *M*. In this respect the function of *C* can be usefully compared to the function of the bimetallic strip in a thermostat. The function of this strip is *not* to turn the furnace on. Whether the furnace is turned on depends on *two* factors: the temperature (which the curvature of the strip supplies information about) *and* the position of the adjustable contact (representing what we desire the temperature to be). That is why the strip is only *part* of the furnace switch. Its duties are purely cognitive.

But even this is too strong. The effects of *C* do not depend simply on what I am here calling the motivational state of the organism. The thermostat is too simple an analogy to capture the way *C* may interact with *other* cognitive structures. Even if we suppose that the drive or desire is the same as that existing during learning, once *C* has acquired an indicator function it may produce quite different effects on motor output (quite different, that is,

14. It should also be clear why I reject Stich's autonomy principle and his replacement argument (1983, p. 165) against the relevance of intentional explanations of behavior. A physical duplicate of an intentional agent, though it behaves the same, does not *yet* (not until it acquires sufficient experience to give the internal indicators the requisite functions) behave that way for the same reasons. Although physically indistinguishable systems will behave the same way (*C* will cause *M* in both), there is no reason to suppose—and if they have had different *histories* every reason *not* to suppose—that the explanation of *why C* causes *M*, of *why* they behave that way, will be the same for both. The only reason one might think the explanations must be the same is if one mistakenly identifies the bodily movements, *M*, with the behavior, *C*'s causing *M*, of which they are a part.

from those it had during learning), depending on what other indicator states are registering positive and depending on what other sorts of associative learning may have taken place between *C* and these other structures. A consistent pairing of conditions *F* and *G* (and, hence, a consistent pairing of the internal indicators of *F* and *G*), for instance, or a change in the kind of consequences (from rewarding to punishing) associated with *M*, may cause a change in the sort of movements (or nonmovements) that *C* (the internal indicator of *F*) produces. What the original learning situation did was to give *C*, not the job of producing *M*, but instead the job of supplying intelligence relevant to the production of *M* and whatever other movements might secure results of the kind that happens to be desired at the time. *C* retains this information-supplying job even when the *use* to which that intelligence is put changes as *C* becomes integrated into a larger and more complex control system.

I do not greatly care whether, in the case of very simple creatures, one chooses to call the products of this learning process—the representational structures described above—*beliefs*. Perhaps this is premature. Perhaps, as was suggested above and as some philosophers have argued (see, e.g., C. Wright 1986; Davidson 1987; Evans 1981), the ascription of belief requires a *system* of beliefs—a representional *manifold* in which the elements not only interact with one another to produce (via inference) new beliefs, but also interact with desires, emotions, intentions and attitudes to yield novel forms of behavior. If sea snails are capable of the kind of associative learning described here (and it seems they are capable of a rather primitive version of it[15]), then surely, some will say, this type of learning is too humble to be the source of genuine beliefs. Snails don't have minds. Their behavior isn't to be explained by what they *believe* and *desire*. Dogs, cats, and chimps may have reasons for some of the things they do, but not bugs and snails.

We will explore the way simple representations interact to generate more complex representational structures in chapter 6, and we will explore the way desires figure in this explanatory scheme in chapter 5. If it turns out that one feels more comfortable in reserving the intentionalistic

15. *Hermissenda crassicornis*, a marine snail, can be conditioned by pairing stimuli (light and turbulence) to which the snail is sensitive. Daniel Alkon and his associates (1983) have not only taught these snails something; they have also traced, at the neuroanatomical and the chemical level, the level *at which* one can trace the *change* in the efficacy of internal indicators (of light and turbulence) on the motor control system.

Though this type of learning is naturally thought of as a form of classical (Pavlovian) conditioning, the learning can also be regarded as a form of operant conditioning. The snail has its response to light (forward movement) punished by turbulence and thereby changes the way it responds to light. I am grateful fo Ruth Saunders, Naomi Reshotko, and Rob Cummins for helpful discussions on this point.

idiom—the language of *desire, belief, knowledge*, and *intention*—for creatures exhibiting a certain minimum level of organization, a certain critical mass of representational complexity, well and good. I have, as I say, no great interest in what seems to me to be a terminological boundary dispute of negligible philosophical interest. The important fact, or so it seems to me, is that even at this simple level we can find organisms that not only have a system of internal indicators on which they depend to guide them through their environment (this itself is nothing very special; it occurs at almost every biological level) but also have internal representations that acquire their status and function *as guides* (thereby getting their hands on the steering wheel) *because* of what they *tell* the organism about the environment in which guidance is necessary. Even at this level, then, we have internal structures whose relevance to the explanation of behavior resides in *what* they say (mean, indicate) about the conditions on which the success of behavior depends. Even at this level, then, we have internal structures that not only mean something but also mean something *to* the organism in which they occur.

If such behavior to which these structures give rise is still too simple and stereotyped to qualify as intelligent, and if, therefore, the internal determinants of such behavior are not to be classified as *reasons*, then some other name must be found. Perhaps we can think of these simple and comparatively isolated representations as proto-beliefs, and of the behavior they give rise to as (in some way) goal-directed but not goal-*intended* (for more on this distinction, see chapter 5). Proto-beliefs may then *become* beliefs by becoming integrated into a larger constellation of representational elements or by acquiring whatever other external trappings may be required of genuine belief. Whatever we choose to call them, though, the individual elements described here exhibit the essential properties of genuine beliefs: they *have* a propositional content, and their possession of this content helps explain why the system in which they occur behaves the way it does.