MEANING, FUNCTIONS, AND THE PROMISE OF INDICATOR SEMANTICS

Jason Richardson

Thesis submitted to the Faculty of theVirginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Arts

in

Philosophy

Valerie Hardcastle, Chair
James Klagge
Joseph Pitt

August 8, 1996
Blacksburg, Virginia

Keywords: Dretske, Meaning, Functions, Indication, Audio Localization

MEANING, FUNCTIONS, AND THE PROMISE OF INDICATOR SEMANTICS

by

Jason Richardson

Valerie Hardcastle, Chair

Department of Philosophy

(ABSTRACT)

In this thesis, I first present Fred Dretske's theory of mental representations, which purports to show how a physical thing could have (non-derived) meaning. In order to illustrate the applicability of the theory to an actual physical system, I discuss the theory in relation to two theories of audio localization (i.e., the capacity to locate the source of sounds in one's environment).

Having clarified the theory, I examine two charges laid against it. Lynne Rudder Baker charges the theory with circularity. Her charge is refuted by appealing to the concept of a "standby function." Stephen Stich charges the theory with vagueness. His charge is refuted by appealing to a general analysis of functions. I conclude that a careful use and analysis of the previously unanalyzed term “function” makes possible the refutation of these two charges.

To Valerie Gray Hardcastle,
who knew just how
to temper the demand for excellence
with the need for patience

and

To all the professors
in the department of Philosophy,
who rarely faltered
in their affirmation of intellectual diversity or
in their disapproval of intellectual pretentiousness

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

INTRODUCTION

Contemporary philosophers of mind have expended a great deal of effort in developing naturalistic theories of mental representation.[1] This effort reflects a desire to demonstrate that the mind is continuous with the natural world. Inasmuch as the mind's capacity to represent is one of the basic characteristics which purportedly sets it apart from natural phenomena, this capacity has been a primary target for efforts at "naturalizing the mind."[2] In accordance with these trends, Fred Dretske's theory of mental representations, known as "indicator semantics," attempts to show how a purely physical thing could have meaning.[3]

In Chapter 1 of my thesis, I present some essential aspects of Dretske's theory. I discuss several standard criticisms of Dretske's theory, as well as his replies to those criticisms. Lastly, I use psychological and physiological literature dealing with the phenomenon of audio localization to illustrate the applicability of Dretske's theory to an actual physical system.

In Chapter 2, I defend Dretske's theory against an important criticism offered by Lynne Rudder Baker. Baker argues that Dretske's theory is circular. I show that Baker's criticism is correct, but that Dretske's theory can be modified in a way that rescues it from the criticism, while leaving the revised theory substantially the same. My interpretation makes use of the concept of a "standby function."

---

[1] For some representative efforts, see Stich and Warfield 1994.
[2] The title to Fred Dretske's most recent book. See Dretske 1995.
[3] The meaning that attaches to some physical things (books, works of art, etc.) is derived from the meaning that attaches to their users. Throughout this essay, I am interested only in meaning that is non-derived, like the meaning that attaches to brain states.

In Chapter 3, I present a puzzle used by Stephen Stich to show a weakness in Dretske's theory. I argue that the weakness can be overcome by analyzing in detail a basic concept in Dretske's theory: the concept of a function. A deeper understanding of functions provides the necessary resources for solving Stich's puzzle.

CHAPTER 1

LISTENING TO DRETSKE: AUDIO LOCALIZATION AND THE PROMISE OF INDICATOR SEMANTICS

In this chapter, I sort through empirical studies of audio localization in search of support for Fred Dretske's naturalistic theory of mental representations. In the first major section, I present a brief account of the most crucial aspects of Dretske's theory, as well as some philosophical criticisms of it. The second major section discusses Dretske's account in light of two different theories of audio localization. I argue that evidence for the first theory does not support Dretske's account. However, the first theory is important in clarifying precisely what is needed from the second theory. Finally, I contend that the second theory contains all of the necessary elements, thereby lending some empirical credence to Dretske's account.

1.1 DRETSKE'S THEORY AND SOME CRITICISMS

1.1.1 Indicator Semantics

Dretske's theory of mental representations, known generally as "indicator semantics," attempts to show how physical things acquire the capacity to represent. The theory centers on a class of physical entities that are able to occupy states which covary reliably with conditions of the world. These entities are called "indicators" (Dretske 1988, p. 58).[4] Good examples of indicators are gas gauges, which indicate the presence of a certain amount of gas in a gas tank; doorbells, which indicate the presence of person(s) on one's doorstep; and compasses, which indicate the presence of magnetic

[4] Unless otherwise noted, numbers appearing in the text refer to pages in Dretske 1988, Explaining Behavior.

north. In all of these examples, certain states of the indicator can be considered as signs that a particular condition of the world is present (54).

Biological organisms possess a representational system whenever indicators internal to an organism acquire the function of indicating whatever they indicate (66-67). For example, my cat Zoey has an internal state that covaries reliably with the sound of food being poured into her food dish. If some mechanism gives that internal state the function (or job) of indicating this sound, then Zoey has a representation of the sound.

Furthermore, *only* indicators which have been given the function of indicating some condition of the world count as representations, because mere indicators do not have the capacity to misrepresent. Misrepresentation is the capacity to represent one condition of the world as present when, in fact, something else (or nothing at all) is present (64). Suppose that my cat Zoey's representation for the sound of food-being-poured can also be set off by the sound of marbles-being-poured. Since Zoey's indicator of the food-being-poured sound only has the function of indicating this particular sound, the indicator is not performing the correct function when it indicates the marbles-being-poured sound. A mere indicator of the sound of food-being-poured could not exhibit this capacity for misrepresentation, because it would not have a function which could be performed incorrectly.

An indicator is given its indicator function by being converted into a switch for behavior (88). As a young kitten, Zoey was rewarded with food whenever she happened to investigate her food dish after hearing the food-being-poured sound. Consequently, a mechanism inside Zoey recruited the internal state which registered the sound, converting the state into a switch that causes Zoey to come running every time she hears the sound. On Dretske's account, this recruitment process is responsible for turning the indicator state into a representation.

1.1.2 Some Philosophical Criticisms of Indicator Semantics

In an article entitled "Misrepresentation," Dretske defends indicator semantics from two charges which threaten it (Dretske, 1994). The two charges to which he replies are the distality and the disjunction problems.

The distality problem asks how we are to know *which* condition of the world is identified by an indicator. Dretske uses the example of magnetosomes to illustrate this point. A magnetosome is a sensory system in certain bacteria living in northern ocean waters. Bacteria possessing this system have the capacity to align their direction of motion with the earth's magnetic north. As a result of their motion towards magnetic north, they are drawn into deeper waters and away from deadly oxygen-rich surface waters.

Dretske recognizes that we can pose the following question: does the magnetosome's magnetic detector have the function of indicating the presence of magnetic north or of oxygen-rich water? Presumably, the bacteria evolved the magnetic field detector as a means for avoiding oxygen-rich water. However, if we were to bring a magnet into close proximity to the magnetosome, we could easily fool the bacterium into following the local field up into the oxygen-rich surface water. It seems wrong to say that the magnetosome is not working properly, because it *did* detect the direction of magnetic north and orient the bacterium's motion. That the bacterium was lead to its demise does not alter the fact that the magnetosome was functioning as it was supposed to. Hence, it may appear that the magnetosome's function is to indicate the presence of magnetic north. But this result runs counter to the observation that the magnetosome's indicator properties are useful for helping the bacterium to avoid oxygen-rich surface water.

To resolve this problem, Dretske points to the capacity of more complex organisms to detect conditions of the world in several different sensory modalities. In such organisms, an indicator state may have the job of producing a behavior upon the presence of one of several different stimuli. However, since any one of the stimuli can generate the behavior that is tied to the state, no single stimulus is *necessary* for the indicator to assume its function. If no particular stimulus is necessary for the state to assume its indicator function, it must be the state's function to indicate, not the stimulus, but the condition of the world associated with the stimulus.

For example, imagine equipping one of the bacteria from the previous example with a photosensor to detect the presence of greater (i.e., more intense) light whenever in dangerously oxygen-rich surface waters. Should we identify the thing indicated with oxygen-rich water, or with one of these two stimuli -- magnetic north or greater light -- associated with the oxygen-rich water? Since both of these stimuli are *alone* sufficient, under normal conditions, to send the magnetosome in the correct direction, neither one is *necessary*. Therefore, it must be in virtue of the oxygen-rich water that the internal state takes on and maintains its indicator function.

In anticipating criticisms of this answer to the distality problem, Dretske discusses the disjunction problem. Suppose a critic argued that the indicator consists of the *disjunction* of these stimuli. The modified bacterium's sensors do not indicate the presence of "greater light" or of "magnetic north" but do indicate the presence of "greater light or magnetic north." The internal state guiding the bacterium is the state which has the function of indicating *that* one of these stimuli occurred. Hence, Dretske is revisited with the problem of deciding where to locate the thing indicated: is it oxygen-rich water, or the disjunctive set of stimuli which indicate the presence of oxygen-rich water?

Dretske fashions an answer to the disjunction problem by invoking yet another capacity of complex organisms. Imagine an organism capable of associative learning. It can learn to react to some stimuli by their association with other stimuli. In the presence of a stimulus s1 which normally indicates the presence of some adverse environmental condition, an organism can learn to produce avoidance behavior in the presence of a stimulus s2. Association of the behavior appropriate to s1 with the second stimulus s2 occurs because of the regular association of s2 with s1.

Of course, the disjunctive set of these stimuli would vary in content over time. The association of new stimuli with the original set of stimuli would substantially alter the disjunctive set of all such stimuli. Dretske points out, however, that we tend to think of the content of our representations as time-invariant. For instance, after learning to associate "cow" with cows, my use of the word "cow" functions in the same way ever after. It presumably picks out the same thing each time I use it (correctly) to refer to cows.

If the disjunctive set of stimuli does not pick out a time-invariant feature of the world, yet time-invariance is required in order for something to count as the content of a representation, then the disjunctive set of stimuli cannot count as the content of representations. Thus, an internal state which produces avoidance behavior in the presence of any one of several stimuli has the function of indicating, not that one of these stimuli occurred, but that some adverse thing is present in the organism's environment. My "cow" indicator has the function of picking out cows, not the disjunctive set of stimuli associated with cows.

Dretske's solution to the disjunction problem is given additional importance by the fact that his solution to the distality problem relies on it. Also, since Dretske's solution to the disjunction problem relies on the use of associative learning, the importance of learning is underscored. Any living system alleged to have meaningful internal states must have the capacity to learn.

In the next major section, I examine empirical studies concerned with the phenomenon of audio localization. These studies should enliven the abstract picture that I have just sketched, by showing how the theoretical requirements laid out by Dretske are met by actual biological systems.

1.2 TWO THEORIES OF AUDIO LOCALIZATION

1.2.1 The Duplex Theory and the Binaural System of Audio Localization

According to the duplex theory of audio localization, organisms locate sounds along horizontal and vertical axes through two different but related mechanisms.[5] Azimuth, the horizontal component, is resolved by appealing to differences in the arrival time of sounds to the ears (called interaural time differences, or ITDs). Elevation, the vertical component, is resolved by appealing to differences in pressure (called interaural level differences, or ILDs) (Middlebrooks and Green, 1991).

How are these differences at the two ears created from a single sound source? Differences of *pressure* result from "shadows" cast over one ear by sound coming from the opposite side of the head (Middlebrooks et al., 1991). Differences in *time* are constructed in low level neural circuitry from differences in the phase of sound waves, caused by the distance required to reach the ear on the

[5] Distance from the organism is ascertained via a different aural mechanism.

far side of the head (Konishi, 1993; Middlebrooks et al., 1991). Hence, the duplex theory postulates a binaural system of audio localization.

The duplex theory of sound localization makes use of the concept of a "space-specific neuron." A space-specific neuron is an individual neuron which codes information about the spatial origins of sounds. In owls, ITD and ILD data is processed in parallel (Konishi, 1993). This processing culminates in neurons whose firing correlates with sounds originating from well-specified regions of space.

How does this information processing occur? Neurons in the central inferior colliculus (ICc), tuned broadly to frequency, each specify a unique ITD. That is, their firing rates correlate significantly with ITDs. The azimuth of a sound stimulus is stored in ensembles of these neurons. These ensembles in turn project onto space-specific neurons in the external inferior colliculus (ICx) (Takahashi, 1989). Empirical results have confirmed that some neurons in the ICx fired only when sounds originated from certain locations (Konishi, 1993).

A similar mapping in the ICc for elevation has not been established, but probably exists. Takahashi speculates that neurons which select for ITDs may select for ILDs, as well, except that the neural map associated with ILDs would vary along a dimension orthogonal to that of stimulus azimuth (Takahashi, 1989).

Due to the comprehensiveness of the research performed on owls, it is nearly certain that an auditory system like that described by the duplex theory is the system used by all owl species. Research in cats suggests that this extends, in all likelihood, to mammals (Konishi, 1993). One researcher is confident enough in these findings to suggest that attention may now be turned to how this information is used in higher auditory processes (e.g., speech, music, etc.) (Konishi, 1993).

The duplex theory describes a physical system instantiating something very similar to the systems described in Dretske's theory of mental representations. Konishi, a duplex theorist, thinks of the auditory system as a kind of algorithm, involving step-by-step transformations of maps which serve as indicators of events taking place outside the organism (Konishi, 1993). In the ICx, particular neurons can be conceived as indicators of sounds originating from well-defined spaces surrounding the organism.

The space-specific neurons found in the ICx of owls project to a portion of the superior colliculus known as the optic tectum. In the optic tectum, the spatial information is integrated with data provided by other sense modalities. This integrated data is used in head orientation responses (Samson, Clarey, Barone, & Imig, 1993). Hence, we have here an example of an internal indicator which has been harnessed to the production of behavior, namely, orientation of the head to sound source locations.

Of course, the binaural system of audio localization does suffer from a few problems. Information derived from binaural cues is troubled by considerations of ambiguity. Ambiguity arises due to several circumstances. In front-back confusions, a sound whose source is to the front of a subject is reported as coming from behind her. As it turns out, front-back confusions are merely a species of the cone of confusion. The cone of confusion derives from the observation that the interaural difference cue generated by a sound source at a given angle is always ambiguous with respect to a sound source at another angle. For any given axis, there are at least two different sound source angles which will produce the same interaural difference (Middlebrooks et al., 1991).

But perhaps these apparent liabilities are assets in our attempt to regard the binaural system as an instantiation of a Dretskean representational system. Perhaps an account of misrepresentation is possible precisely *because* the space-specific neuron can be fed ambiguous information. Although a sound issues from only one direction, the interaural cue it generates can be associated with *two* directions. We may speculate that transformations performed upon this ambiguous information, through random fluctuations in the auditory nerves, will sometimes cause the auditory system to settle upon the wrong direction. A case of misrepresentation occurs, and the wrong behavior (e.g., head orientation to the wrong direction) is triggered.

This story about audio localization appears *prima facie* to provide an example of a system that has the properties which Dretske says would be sufficient to make a system representational. The space-specific neuron is an internal state that is tied to a specific condition of the world. My orienting response to a noise in my environment is produced because I have an internal state whose function is to indicate sound source locations to me. Additionally, the binaural system is capable of misrepresenting the location of sound sources. The space-specific neuron can sometimes represent a sound as coming from one direction, when the sound actually originated from another.

1.2.2 The Binaural System: a Sophisticated Representational System?

Upon closer examination, however, there are reasons to doubt that the binaural system of audio localization alone yields a representational system of the sort that interests Dretske. Dretske intends for his theory to explain the representational capacities of a relatively sophisticated sort of representational system, namely, those systems which exhibit intelligent behavior. And the binaural system alone is not capable of resulting in genuinely intelligent behavior, or so I will argue.

Dretske distinguishes sophisticated representational systems from simple ones (at least in part) by distinguishing between tropistic and intelligent behaviors (93ff). A tropistic behavior is essentially a reflexive behavior. Such behaviors, when initiated by the proper stimulus, occur automatically, without regard to any special circumstances that obtain when the behavior is initiated. By contrast, an organism exhibits intelligent behavior when the organism demonstrates the ability to adjust its behavior in light of changing conditions.

Consider the following example of a tropistic behavior (93-94). The eyes of a certain species of caterpillar are genetically hardwired to cause members of the species to crawl in the direction which creates equal intensities of light on their two eyes. If one eye is blinded, a caterpillar will steer in circles until it starves to death. Because its representation of light is hardwired, the caterpillar cannot adjust its behavior to suit new circumstances.

Unlike these caterpillars, human beings (and other organisms with sophisticated representational systems) exhibit relatively intelligent behavior, because they can alter their behavior to suit changing circumstances. If a human being is moved to a new environment or is divested of some representational ability that she previously possessed, she can nonetheless learn to adapt her behavior to suit the new circumstances. Caterpillars are limited by specialized hardware which is suited only to certain, very narrow ranges of circumstances. If we alter these circumstances, by rewiring the caterpillar's hardware or by placing the caterpillar in a new environment, we can demonstrate the caterpillar's inability to exhibit genuinely intelligent behavior.

According to Dretske's theory, only certain kinds of recruitment processes are able to create sophisticated representational systems of the sort possessed by human beings. Dretske distinguishes among (at least) three kinds of recruitment processes, in which indicator states are given their indicator functions. Each of these recruitment processes can be considered a form of learning. Only two of these forms of learning can be used to create sophisticated representational systems. These forms of learning are "discrimination" learning (99) and "observational" learning (119).

A third form of learning cannot by itself be used to create sophisticated representational systems. What I will call "developmental" learning is the kind of learning which takes place as the result of native capacities which come "on-line" at various points in the developmental process. That is, an organism acquires, or "learns," new capacities as biological systems mature (92-93). Since these capacities cannot be substantially altered once they are acquired, this form of learning cannot be used (exclusively) to create a sophisticated representational system. The representations acquired cannot be modified in the light of an organism's experience.

If these points are granted, then certain empirical findings arguably show that the binaural system alone should not qualify as a sophisticated representational system. Studies show that, without the presence of biological structures lying *outside* the binaural system, space-specific neurons of the Icx cannot be recruited for behaviors other than head orientation. If structures outside the binaural system are necessary for the harnessing of representations of sound source locations to behavior, then the binaural system alone cannot be sufficient for the task.

Some researchers studying audio localization in animals employ learning paradigms in which the animals being studied are trained (via discrimination learning) to respond in certain ways to sound sources in the experimental environment. Typically, the animals are trained to approach the sources of sounds, where they receive rewards if they accurately localize the sources. Researchers have explored the neurological substrates of this conditioned behavior by training two groups of animals on the localization task, then lesioning portions of the brains of one group prior to assessing performance on the task.

In several studies, two groups of cats were trained to approach sound sources for rewards (Jenkins and Masterton 1984, 1982). The auditory cortices of cats in one group were ablated, and both groups of cats were subsequently tested for their performance on the localization task. Cats who suffered the cortical lesioning performed significantly worse than cats who did not. The research shows that the auditory cortex is importantly involved in learned or conditioned responses to the localization of sound sources (1984).

These studies show that the binaural system alone is not sufficient to support conditioned responses to the localization of sound sources. The only learning that takes place in the binaural system is the developmental learning which is not sufficient to yield genuinely intelligent behavior. Since the auditory cortex appears to be implicated in discrimination learning, we need to find a system of audio localization which extends into the auditory cortex.

Thus, we have incentive to look beyond the binaural system of audio localization for a complete illustration of Dretske's theory. Fortunately, this century has seen the development of *two* theories of audio localization. This second theory can provide what the first theory does not -- namely, evidence that information about sound source location is accessible to the auditory cortex, where it can be harnessed to a variety of behaviors through discrimination learning.

1.2.3 The Spectral Cue Theory and the Monaural System of Audio Localization

A rival to the duplex theory has been under appraisal since at least 1967, when Batteau observed that information about sound source locations could result from sounds striking the head, shoulders, chest, and (especially) the pinna (Middlebrooks et al., 1991). Sound interfering with the body creates characteristic spectra. A spectrum is the sum of the harmonic variations of the fundamental frequency of a sound energy, where each harmonic is a multiple of the fundamental frequency (Handel, 1988). Sound spectrums can be heard and analyzed in a single ear; hence, the rival to the duplex theory postulates a monaural system of audio localization.

Spectral cues offer information about the spatial origins of sounds by producing a pattern of troughs and peaks that are characteristic of a particular region of space. Sound that gets filtered through the pinna is an especially rich source of information. Waves propagating along direct paths to the eardrum are augmented by waves propagating along reflected paths. The length of the reflected path varies with the elevation of the source. It has also been shown that monaural cues carry information about azimuth (Middlebrooks et al., 1991).

Only recently, however, has evidence emerged for the important role that spectral shape cues play in audio localization. Some researchers have developed mathematical models which show how the brain could be deriving useful localization information from the patterns of sounds in pinna-based spectra (Zakarauskas and Cynader, 1993). Other researchers have directly elicited orientation responses from the use of spectral cues (Middlebrooks, 1992). Still others have used connectionist methodologies to suggest that spectral cues are importantly related to audio localization

(Middlebrooks et al., 1994; Neti et al., 1992). I shall discuss four studies which provide evidence for the importance of spectral cues.

Poon and Brugge (1993) studied the characteristics of auditory nerve fibers, which convey information from tiny hairs in the cochlea of the inner ear. Their study shows that auditory nerve fiber are sensitive to spectral notches, which in turn are correlated with the fiber's frequency selectivity (Poon and Brugge, 1993). In humans and other organisms, the auditory nerve fibers carry their information to the cochlear nucleus. Hence, we know that the auditory sense mechanisms convey important information about the sound sources to early stages in the auditory nervous system.

In a second study, a three-layer, back-propagation neural network was trained to recognize spatial locations based on spectral data recorded from a neuron in the cochlear nucleus of a domestic cat. Having been trained, the network created by Neti et al. (1992) successfully localized sound on the basis of purely monaural, spectral information.

For the purposes of this paper, three things should be noted about the Neti study. First, the study must be qualified by the observation that the original sounds from which the spectra were derived were artificially "flat" (Neti et al., 1992).[6] Spectral cues have a peculiar limitation -- the patterns in the spectrum may be present as a result of interference with the external ear, or as a result of patterns intrinsic to the sounds themselves. The pattern processor has no way to tell these situations apart (Middlebrooks et al., 1991). To avoid complications introduced by this problem, the pinna-based spectrum was developed from a flat sound which would not introduce uncontrolled variability into the

[6] A "flat" sound is a sound composed primarily of white noise, without any noticeable peaks in amplitude or changes in frequency.

spectrum. Hence, the ability of organisms to use spectral cues in a *realistic* setting must be affirmed only cautiously.

Second, the network was especially sensitive to certain prominent features often present in spectral data. These are notches, or sharp dips in the intensity of sounds, associated with a particular frequency. Neurons in the hidden layers of the neural net detected spectral notches, especially the first spectral notch (FN) between 5-18 Khz. According to Neti, the network resorted heavily to the FN spectral notch in determining spatial locations. Moreover, behavioral studies have indicated that organisms are sensitive to these notches. These studies provide independent evidence for the importance of Neti's finding that spectral notch information can be used by connectionist nets (Neti et al., 1992).

Third and finally, a word about the hidden-layer units in Neti's study. An analysis of these units revealed that they are spatially selective, in a way which is highly influenced by the first frequency notch (Neti et al., 1992). However, since the back-propagation method used in this study is instantiated nowhere in the brain, it can only be speculated that some other method of training would result in similar sensitivities to spatial features (Bechtel and Abrahamsen, 1991). Neti's study is important, nonetheless, in demonstrating that connectionist nets can use real spectral information to develop spatial sensitivities.

A third study indicates that the information from which spatial sensitivities were derived may be available, not only in the cochlear nucleus, but also in the auditory cortex. In an interview conducted by a reporter from Science, Middlebrooks recalls making an interesting observation concerning cells in one region of the auditory cortex of the cat: "the temporal pattern of firing of neurons changes as a function of sound location..." (Barinaga, 1994).

With this thought in mind, Middlebrooks et al. trained a 40 unit, single layer linear perceptron to recognize spatial locations based on data gathered from a neuron in the auditory cortex of a cat (the ectosylvian sulcus cortical region). The network generated the correct solutions using data compiled from the joint activity of many different neurons (Middlebrooks et al., 1994). In other words, spatial location is indicated in a distributed fashion by the recognition of temporal patterns. It is not unreasonable to believe that these temporal patterns derive from spectral cues.

The last clue needed to confirm the relevance of spectral shape cues is the direct correlation of neural firing with monaural cues in the auditory cortex of mammals. Such a discovery has recently been made. In 1993, Samson et al. observed cells in the auditory cortex of a cat which were sensitive to azimuth when the subject had been exposed to monaural sounds only. In the accompanying article, the researchers took pains to point out that these monaural directional (MD) cells did not exhaust the cells in the auditory cortex. In other words, the MD cells do not form a map of auditory space in a manner equivalent to the somatosensory cortex or the visual cortex.

The majority of researchers in the field believe that spectral cues figure prominently in the localization of sounds (*cf.* Middlebrooks et al., 1991). The fact that some unilaterally deaf persons can make correct judgments about horizontal sound source location provides near definitive assurance that monaural spectral cues provide at least some assistance in determining azimuth (Middlebrooks, 1992). Hence, spectral cues provide important information regarding the location of sounds.

However, there is an important study which restricts this conclusion. Wightman and Kistler set up a series of experiments in which they pitted ITD cues against ILD cues and spectral cues (Wightman and Kistler, 1992). They used headphones and a digital computer system to simulate a free-field, anechoic (i.e., non-echoing) sound chamber. This newly developed technology permitted complete control of the audio input. (Additional experiments confirmed that the use of headphones did not jeopardize the results.)

These results indicated that low-frequency ITD cues dominate the perception of sound source location. ILD cues and spectral cues regain influence only when low-frequency cues are eliminated. Note the experimenters:

> The clearest result is that when a large ITD in the low frequencies signals a source far to the right and close to the horizontal plane..., all judgments of apparent source direction cluster around these values. This occurs in spite of the presence in many stimuli of large interaural intensity and spectral shape cues that indicate source directions on the opposite side and at higher or lower elevations (Wightman et al., 1992).

Fortunately, these results do not demonstrate that spectral shape cues have *no* influence on audio localization. Neti's results suggest that spectral shape cues may not be used below 5 Khz. This frequency is also the highest frequency used in the Wightman and Kistler study (Neti et al., 1992; Wightman et al., 1992). Hence, spectral shape cues probably play an especially important role in higher frequencies.

1.2.4 The Representation of Spectral Cues

How is the source of a sound represented in the monaural system? Is the information present in the form of indicator states which are given the function of indicating the sources of sounds in external space? Moreover, are these indicators recruited to cause behavior *because* of what they indicate? In order to answer these questions, I will discuss the nature of the biological mechanisms that underlie or implement sound source representations in the monaural system.

I reiterate, first of all, that spectral cues make prominent use of features whose presence indicates the source of a sound in external space. The features to which I am referring are the frequency notches discussed in the Neti study and the Poon and Bogg study. The frequency notches are local drops in the pressure of an acoustic signal, created by the interaction of sound waves with the head, chest and pinna. In Neti's study, frequency notches were used by a connectionist network to localize sound sources. Poon and Bogg showed that neurons in the cochlear nucleus are sensitive to these notches. I suggest that these features can be considered as examples of the indicators in Dretske's theory of mental representations, because these notches covary reliably with the presence of sounds in certain locations in external space.

A *prima facie* concern is raised, however, by one characteristic of the representation of sound source locations in the monaural system. My discussion of the monaural system suggests that the representation of sound source locations is "time-dependent." An acoustic spectrum is built from the sum of the harmonic variations of the fundamental frequency of an acoustic signal. This spectrum varies as a function of time. Researchers speculate that these time-dependent spectra are preserved by the monaural system and that the same information is conveyed to neurons in the auditory cortex. Middlebrook's study using a connectionist network showed that time-dependent firing patterns of neurons in the auditory cortex possess the information necessary to localize sound sources.

Can time-dependent representations serve as indicators, as Dretske's theory requires? The characteristic of time-dependency might appear to count against the possibility that something could serve as an indicator, because the indicator state cannot be picked out at any particular instant in time. Hence, there does not seem to be a specifiable state to serve as the indicator. Because the indicator state cannot be identified with a single state, it appears that no single thing can be picked out as the indicator, as required by Dretske's theory.

However, this worry should disappear upon further consideration. The mere fact that a representation cannot be located at a particular moment in time does not mean that nothing determinate is covarying reliably with a condition of the world. Sometimes an indicator exists only as a function computed over a temporally extended domain. Imagine a sailor who, while on a ship at sea, is attempting to straighten a plank with his adze. The sailor uses a level to determine if his plank is true. Since the ship is rocking, he cannot simply wait for the bubble to rest in the middle of the level. Instead, he must wait until the bubble is traveling, on average, the same distance on either side of the middle. When he is satisfied that this is happening, he considers the plank to be level. Furthermore, the mere fact that the event is temporally extended does not mean that there is nothing which serves as the indicator. The plank's being straight covaries reliably with a certain *pattern of activity* within the level.

These points ought to allay concern over the indicator itself, but Dretske's theory also requires the indicator to have the *function* of indicating some condition of the world. Somehow, we must be assured that groups of neurons in the auditory cortex have the job of indicating the presence of sound source locations. Two arguments give us reason to think that groups of neurons in the auditory cortex do have this function.

First, I argue that groups of neural units have the function of indicating the locations of sound sources because these groups *lose* their indicator properties when those indicator properties are not exercised. In other words, if a group of neural units is given different indicator properties when its previous properties were not regularly exercised, we can infer that the group had the function of indicating whatever it indicated. The central nervous system makes efficient use of its resources. If a group of neural units are not being used, those units are co-opted for other purposes.

Studies of plasticity in the auditory cortex of mammals show that groups of neural units in the auditory cortex get co-opted in just this manner when their indicator properties are no longer being exercised. In one review of recent findings, a researcher discusses studies of cats that were given restricted cochlear lesions. Lesioning of the cochlear nucleus results in deprivation of input to select portions of the auditory cortex. The deprivation of input to some portions of the cortex resulted in the expansion of adjacent representations. Moreover, the new and expanded representations became normal functioning parts of the cortex (Irvine, 1992).

A second argument invokes the ability of groups of neural units in the auditory cortex to misrepresent. Mere indicators cannot be representations, because an indicator cannot represent a condition of the world as present when in fact the condition is not present. Indicators are "hardwired," so to speak; they carry information about that condition with which they reliably covary, *whatever* that condition may be. However, if an indicator is given the function of indicating some condition of the world, then the resulting representation is capable of "fooling" an organism. The organism is capable of making use of the indicator, whether or not the indicator is signaling the presence of its assigned condition of the external world.

I mentioned an example of this phenomenon in describing the properties of the monaural system. Spectral cues can arise from the interaction of sound waves with the body, or they can arise from properties intrinsic to the spectra of a given sound. Despite the fact that the latter, spurious spectral cues can sometimes fool organisms, the neural representation in the auditory cortex remains stable. It maintains a "lock," as it were, on the spectral cues that indicate the location of sounds in external space. The maintenance of this lock shows that the organism is *relying* on its indicators; the indicators are *supposed* to operate in a certain way, even when they are not actually doing so. Hence, the ability to misrepresent demonstrates the presence of an indicator function.

These arguments support the view that the monaural system produces indicators which have indicator *functions*. However, I introduced the monaural system because I wanted to show that some system of audio localization yields representations which could exhibit plasticity because of recurring experiences with misrepresentation. A system of audio localization must be modifiable in light of experience, in order to qualify as a sophisticated representational system. If the factors that give rise to misrepresentation are hardwired in the monaural system, as in the binaural system, then the monaural system will not meet the final requirement of plasticity.

My discussion of the monaural system gives us reason to believe that the monaural system at least *could* be involved in discrimination learning. Previously, I argued that the auditory cortex is implicated in the biological mechanisms underlying discrimination learning for auditory behaviors. I also showed that some representations produced by the monaural system are located in the auditory cortex. Because the monaural system generates representations in the neural site that is necessary for discrimination learning in the auditory system, I conclude that the monaural system makes intelligent behavior possible. Hence, the monaural system ought to qualify as a sophisticated representational system. Finally, if the monaural system qualifies as a sophisticated representational system, then we have picked out a living system which illustrates Fred Dretske's theory of mental representations.

CHAPTER 2

STANDBY FUNCTIONS AND A CHARGE OF CIRCULARITY

In this chapter, I present three desiderata of any naturalistic theory of mental representations and show how Dretske's theory of mental representations satisfies these desiderata. After discussing a fourth requirement that is sometimes imposed on theories of representation, I discuss an important criticism developed by Lynne Rudder Baker. Baker contends that Dretske's theory is circular, because it explains meaning in terms of structuring causal roles, while explaining structuring causal roles in terms of meaning.

In defending Dretske's theory from this charge of internal circularity, I argue that Baker's criticism is sound but that Dretske's theory can still succeed if a state's meaning can be separated from a state's structuring causal roles. I effect this separation through the notion of a "standby function." Having clarified this notion, I anticipate several criticisms.

2.1 DRETSKE'S THEORY AND THE THREE DESIDERATA

In the recent literature on mental representations, philosophers have proposed three desiderata of any naturalistic theory of mental representations (Warfield and Stich 1994, pp. 5-6). No theory which fails to satisfy these desiderata is generally considered a plausible candidate for a naturalistic theory of mental representations. Fred Dretske, among other philosophers, has developed a theory which purports to satisfy these desiderata.

The first two desiderata stipulate that meaningful states must have (1) content and (2) the capacity to misrepresent. A third desideratum, which I call the "constraint of physicalism," requires an adequate theory to show how *physical* things could satisfy (1) and (2).[7] I will explain how Dretske's theory purports to satisfy these three desiderata.

2.1.1 Content and the Capacity to Misrepresent

We can think of the first two desiderata as specifying necessary conditions for a thing's having the capacity to represent.[8] Dretske believes that these necessary conditions are satisfied by his theory of mental representations (p. 70). According to his theory, C has meaning if C has the "function of indicating" some condition of the world F (p. 52). This account has two distinguishable components, which correspond to the first two desiderata. I will explain what Dretske means by each of these two components and show how each satisfies its respective desideratum.

For Dretske, to "indicate" means to covary reliably (p. 53). For example, the buzzing of a doorbell "indicates" that someone is at the door, because the buzzing correlates reliably with the presence of a person at the door. Put another way, the buzzer "carries the information" that someone is at the door. If we wished, we could even say that the buzzer is "about" persons at the door.

This capacity of one thing to be "about" another is an important attribute of mental representations. A thing has "content" in virtue of possessing this capacity. According to Dretske's theory, a thing C has content in virtue of indicating the presence of some condition of the world F, thereby being "about" F.[9] Hence, Dretske's theory seems (at least intuitively) to satisfy one of the two necessary conditions for a thing's having meaning, namely, the requirement of having content.

---

[7] Philosophers who think of intentional states as theoretical primitives will generally regard this constraint as unnecessary.

[8] Throughout this essay, the notions of "meaning" and "representation" are used interchangeably.

[9] Dretske's theory should not be dismissed because of its failure to provide an intuitively satisfacto-
ry account of reference for natural language terms. It is not immediately apparent, for instance, that the

However, C's capacity to indicate F does not by itself give C meaning. In order to see this, suppose that C is "hardwired" in such a way that C always indicates the presence of condition F. C "lights up," we might say, whenever placed in the presence of condition F. Examples include compasses which always indicate the presence of magnetic north and "dummy" lights which always indicate the presence of an overheated engine.

Now, although C reliably indicates the presence of F, we cannot say that C means, or represents, F. Why? Because mental representations are characterized by an ability to get things wrong, that is, to "misrepresent" (p. 65). The species of mental representations that we call "beliefs," for instance, are generally said to be "true" or "false."[10] But C is never "false," never misrepresents, if it invariably indicates the presence of F. Moreover, if C cannot misrepresent, then C lacks an ability which a thing must possess in order to have meaning.

Of course, C can always break down and fail to indicate F, or indicate something other than F (e.g., G). But C has not thereby acquired the ability to misrepresent. It simply indicates something different; it is "about" something other than F (p. 65). The problem, then, is to figure out how

---

term "Alpha Centauri" has content because it "indicates the presence of" Sol's closest stellar neighbor. Nonetheless, it may (in principle) be possible to build up an account of reference for natural language terms vis-a-vis the content possessed by lower-level mental states. See Fodor 1994 for an insightful discussion of these issues.

[10] It is important to realize that Dretske's theory does not give a full-blooded account of beliefs. Dretske notes that tokenings of beliefs may need to be appropriately hooked up to one another before any of them count as genuine beliefs (pp. 106-107). Furthermore, a satisfactory account of beliefs should explain the attitudinal aspect of beliefs, inasmuch as beliefs are typically conceived to be "propositional attitudes." Hence, Dretske's theory should be understood as giving an account of a relatively simple variety of representations, which we might call (following Dretske) "proto-beliefs."

something false "truly characterizes a physical system."[11] How can C be about F even when it *fails* to indicate F?

Dretske's solution to the problem of misrepresentation is to stipulate that C must also have the "function" of indicating F.[12] This move appears to do the trick, because C can have the function of indicating F, and so represent F, even though C actually indicates G. Hence, this account shows how one thing could be "about" some condition of the world yet have the capacity to misrepresent that condition.

2.1.2 The Constraint of Physicalism

Many philosophers believe that an adequate theory of mental representations must show how physical things could have meaning. Jerry Fodor, for example, argues that scientific psychology is baseless unless it can show that the entities it investigates are part of the physical world (Fodor 1994, p. 5). Since mental phenomena are the kinds of things investigated by a scientific psychology, mental phenomena must be understood as arising from, or reducible to, or in some way intimately related to the physical world.

Dretske's theory appears to satisfy the constraint of physicalism, because it assigns meaning to states of the brain.[13] Since states of the brain are physical things, Dretske's account (if successful) shows that a physical thing can have meaning. Moreover, if actual states of the brain possess the properties which Dretske thinks are sufficient to impart meaning to a physical thing, then some

---

[11] See Dretske's comments in the "Roundtable Discussion" in Hanson 1990, p. 207.

[12] For Dretske, to say that C has the function of indicating F is (roughly) to say that C has been given a job to do because of what it indicates (p. 69). To my knowledge, Dretske nowhere provides an explicit, and non-intentional, analysis of "function." I attempt to supply this analysis in Chapter 3.

physical things actually exhibit one of the basic characteristics of the mental, viz. the capacity to represent.

2.2 PROBLEMS WITH THE DRETSKEAN ACCOUNT

2.2.1 A Fourth Requirement (?)

If the foregoing account is correct, then Dretske's theory satisfies the three desiderata of any naturalistic theory of mental representations. However, Dretske's theory may not be out of the woods yet. Many philosophers place a fourth requirement on naturalistic theories of mental representations.[14] They argue that an organism's representational capacities must *play an explanatory role in the organism's behavior*. If a state C has meaning, but its having meaning does not make any difference to the behavior of the organism of which it is a part, then C's having meaning is irrelevant to *explanations* of that organism's behavior (Dretske 1988, pp. 79-80).

Suppose, for instance, that I wire my microwave to radiate electromagnetic energy in discrete pulses, and I arrange for these pulses to say things in Morse code ("cook, cook, cook..."). No one is going to think that my lunch is heated because of the *message* that is being carried in the pulses. Rather, the food is heated because the electromagnetic energy is causing an increase in the food's mean kinetic energy. If a theory of mental representations creates an analogous role for meaningful states of the brain, then the theory gives meaning no explanatory role. It explains meaning in terms of physical states but fails to explain any physical states in terms of meaning. Meaning becomes a sort of epiphenomenal property, clinging to a substratum without affecting the substratum.

---

[13] Dretske uses the generic term "structures" in Explaining Behavior, but the Preface makes it clear that the structures that Dretske is thinking about are neural structures. See pp. ix-x.

Dretske is committed to the view that meaning, or a thing's having meaning, does have a role to play in explaining the behavior of organisms with representational capacities (Dretske 1991, p. 113). To show how Dretske gives meaning an explanatory role, I must review his distinction between "triggering" and "structuring" causes.

Consider the example of a terrorist who plants a bomb in a general's car (Dretske 1993, pp. 122ff); the bomb is designed to explode when someone turns a key in the ignition. Suppose the plan works -- the general turns the key and the bomb explodes. On the one hand, we might argue that the general was the cause of the explosion. After all, he turned the key, thus initiating the chain of events which led immediately to the explosion. The general acted as the "triggering" cause of the explosion.

On the other hand, we might prefer to say that the terrorist caused the explosion. The terrorist planted the bomb; the terrorist structured the car in such a way that it would explode when the general turned the ignition. Thus, we could say that the terrorist acted as the "structuring" cause.[15]

Triggering causes are useful for explaining why an event E happened at one time rather than another (p. 42). Returning to our example, the general's car exploded when he turned the key in the ignition, not when he turned off the car alarm or opened the car door. Structuring causes are useful for explaining why an event E happened, rather than E' or E'' (p. 42). Thus, the terrorist's action explains why the general's car did not start, but instead exploded, when the key was turned in the ignition.

---

[14] I say "many" philosophers, rather than "all," because some philosophers are not convinced that the requirement is crucial. See, e.g., Segal and Sober 1991, p. 28.

[15] These modes of causal explanation are considered by Dretske to be complementary rather than mutually exclusive.

Dretske argues that meaning has an explanatory role, not as a triggering cause, but as a structuring cause of behavior (p. 50). A state's meaning does not explain why a causal chain gets initiated; rather, it explains why one causal chain occurs instead of another. A state C's meaning is explanatorily relevant to behavior by explaining why the organism is structured to perform a behavior B (rather than B', B'', etc.) upon C's indication of F.

The structuring of an organism occurs through a recruitment process.[16] Simple learning mechanisms provide good examples of the recruitment process (pp. 95-102). Imagine, for instance, that a state C's property of indicating some condition of the world F has made it valuable to an organism O. That is, every time O performs a behavior B in the presence of F, O gets rewarded. Hence, O recruits C to do a job, namely, to initiate B upon C's indication of F. The learning mechanism does the physical structuring; however, since the mechanism structures O to perform B in virtue of C's indicating F, we can say that C's indicating F is the structuring cause of O's performing B.

2.2.2 Baker's Criticism of Dretske's Theory

Dretske appears to have succeeded in showing how a physical state can be a cause in virtue of that state's representational capacities. However, recent criticism of Dretske's account shows that the account is actually not successful, because the theory proposed by Dretske is circular (Baker 1991,

---

[16] Note that, on my view, the term "recruitment" can be used to designate two different kinds of processes: (1) the process by which an indicator state acquires the function of indicating whatever it indicates and (2) the process by which a physical state that *already* has the function of indicating some condition of the world becomes recruited as a structuring cause of behavior. Dretske does not distinguish between these two kinds of processes. I generally distinguish between them by referring to (2) as a "secondary recruitment process" (where "secondary" indicates, not importance, but temporal order).

1995). Lynne Rudder Baker has argued that Fred Dretske is committed to the following propositions:

(a) A mental state C has meaning in virtue of its having a structuring causal role in a C->M behavioral process.[17]

(b) A mental state C has a structuring causal role in a C->M behavioral process in virtue of its having meaning.

Together, "(a) and (b) form a tight circle" (Baker 1995, p. 59). Moreover, Baker thinks that Dretske cannot give up either of these propositions without admitting that his theory has failed.

Dretske is committed to (b), Baker argues, because of his commitment to the view that a physical thing's having meaning can explain that thing's causal efficacy. He cannot give up this view without admitting that his theory merely shows how meaning can supervene on physical states while failing to influence those states. The account will have failed, inasmuch as Dretske's explicit goal has been to make room for "reasons in a world of causes" (the subtitle to Dretske's Explaining Behavior).

Dretske's commitment to (a) is perhaps less apparent, especially since he says expressly, not that C has meaning in virtue of having a structuring causal role, but that C has meaning in virtue of having the function of indicating F. Baker presents careful arguments which purportedly show that an internal state C gains the function of indicating F only because C acquires a structuring causal role (Baker 1991, pp. 103-107). But, if C has meaning in virtue of C's having the function of indicating F and C has the function of indicating F in virtue of C's having a structuring causal role, then C has

[17] Baker's use of the expression 'a C->M behavioral process' to represent a behavior reflects Dretske's considered view. I have simplified my summary of Dretske's theory by using the single variable 'B' to represent a behavior.

meaning in virtue of having a structuring causal role. Thus, if Baker's arguments are sound, Dretske is committed to (a).

Baker notes that the circle in Dretske's theory is often signaled by "slippage" in the use of the expressions "indicating" and "having the function of indicating" (Baker 1991, p. 108). When Dretske is being careful, he says that an internal state C means F because C has the function of indicating F.[18] Sometimes, however, Dretske appears to say that C means F (merely) because C indicates F.[19] In slipping from the former to the latter, Dretske reveals the circle in his theory.

I propose tracing the slippage to its source by focusing on the process of recruitment. In various places, Dretske says that meaning has an explanatory role in behavior because a meaningful state C is recruited in virtue of what it means. That is, C's having meaning explains C's being recruited (Dretske 1990a, p. 785; Dretske 1988, p. 99). However, Dretske also claims that a state acquires the function of indicating whatever it indicates when, in fact, C gets recruited (Dretske 1990a, p. 785; Dretske 1990b, p. 121; Dretske 1988, pp. 84-85, 87). So, C's being recruited explains C's having one of the properties necessary for it to have meaning. But if C's being recruited explains C's having meaning, how can C's having meaning explain why it was recruited?

This problem is manifest in Dretske's insistence that these things happen "at the same time" (Dretske 1992, p. 197; Dretske 1990a, p. 785; Dretske 1988, p. 84). Somehow, at the very moment that C's having the function of indicating F causes C to be recruited, C's being recruited causes C to acquire its function as an indicator. But is it really plausible to say that C is recruited in virtue of a property that C does not acquire until the very moment of its recruitment? Besides being

[18] See, e.g., Dretske 1988, p. 52, and Dretske 1990a, p. 784.

ontologically suspect, this claim reveals the circularity in the heart of Dretske's account: recruitment explains meaning, and meaning explains recruitment.

Like Dretske, defenders of Dretske's theory sometimes say that a state C is recruited in virtue of C's *indicating* F, rather than in virtue of C's *having the function of* indicating F.[20] However, these defenders are either hedging or have given up on Baker's proposition (b), viz. meaning is explanatorily relevant to behavior. By stipulation, C does not mean F until C has the function of indicating F. Hence, Dretske's defenders do not assert that C's having *meaning* causes C to be given a structuring causal role; they merely assert that C's having *certain indicator properties* causes C to be given a structuring causal role. But, in that case, an organism O's behavior cannot be explained in terms of states that have meaning -- not without illicitly taking (mere) indication as sufficient for meaning.

Thus, we are forced to admit the cogency of Baker's criticism. However, we are not forced to accept Baker's conclusion, namely, that Dretske's account has failed (Baker 1995, p. 59; Baker 1991, p. 103). As I will show, we can adopt a modified version of Dretske's theory which rejects (a). Moreover, if (a) does not hold, then Dretske's theory is no longer circular in the manner pointed out by Baker.

2.3 A WAY OUT OF THE CIRCLE

Recall that we are forced to accept Baker's proposition (a) only if we agree that C takes on its function of indicating F in virtue of C's acquiring a structuring causal role in producing behavior. I will now propose an account in which C does *not* acquire its function of indicating F in virtue of C's

[19] See, e.g., Dretske 1988, p. 88 (second full paragraph).

acquiring a structuring causal role. In other words, while C's acquiring a structuring causal role is generally sufficient to show that C has meaning, it is not a necessary condition for C's having meaning.

Imagine an organism O that possesses a biological mechanism M which embodies the following heuristic: recruit all internal states which indicate (or "carry information about") conditions of the world. Importantly, M does *not* recruit a state C because C's indicator properties have already made C valuable to the organism. Imagine, rather, that O uses M to acquire indicators because the information carried by indicators is valuable, in general.

States recruited by M are given a job to do: they are supposed to "stand by" until they are needed. Let us stipulate, then, that a state C which has the property of indicating F and the function of standing by until the information carried about F is needed has the "standby function" of indicating F. In such cases, I will speak of C as having the "s-function" of indicating F.

For my purposes, the most important property of internal states with s-functions is their disengagement from systems of motor output. Since these states are on standby, they are not tied to any mechanism which produces motor output. If a state C with the s-function of indicating F "lights up" while on standby, nothing happens. State C has not yet acquired a structuring causal role. Hence, C does not have its s-function in virtue of having a structuring causal role.

This account produces the required separation between an internal state's meaning and its role in initiating behavior. However, it simultaneously generates a problem. In the standard Dretskean account, indicators are distinguished from indicators that have indicator functions by the fact that the

---

[20] See, e.g., Godfrey-Smith 1990 and Hardcastle 1994.

latter are able to produce motor output. However, since neither indicators nor indicators with standby functions are connected to systems of motor output, it is no longer clear how to distinguish between these two kinds of internal states. Plainly, there must be *some* real difference between the two kinds of states; otherwise, they could not have different functional properties. But if they cannot be distinguished by their effects on motor output, how can they be distinguished?

This problem can be resolved by considering the possibility of (what I call) "spurious" indicators.[21] A spurious indicator is an internal state which bears a counterfactual relation to some condition of the external world that will never actually occur in the presence of an organism.[22] If the condition *were* to obtain, then the indicator would produce an indicator token which indicates the presence of the condition. But since the condition never *actually* obtains, the indicator is spurious.

Spurious indicators are noteworthy because of the nervous system's tendency to weed them out. In the nervous system, synaptic connections can be regarded as indicators, inasmuch as their activity is correlated with the presence of conditions of the external world. Now, studies have shown that human neonates possess an extraordinary abundance of synaptic connections. These connections are gradually weeded out over time, and the weeding out process occurs on the basis of experience. If a synaptic connection is part of a path along which sensory input frequently flows, then the synaptic connection is stabilized. The connection may even remain intact long enough to become a permanent

[21] I am indebted to Valerie Hardcastle for proposing this concept to me and to Andrew Baxter for convincing me that the concept solves the associated problem.

[22] For example, cats might have a spurious indicator for perfectly square mice; human beings might have a spurious indicator for golden spheres a mile in diameter; etc. Any indicator which signals the presence of something that could not exist in the actual world is a spurious indicator.

part of the nervous system. On the other hand, if a synaptic connection is part of a path along which sensory information never flows, then the synaptic connection is eventually eliminated.

The tendency of the nervous system to eliminate spurious indicators reveals something important about indicators in the nervous system -- namely, *all* indicators in the nervous system have the s-function of indicating some condition of the world. How do we know that they all have s-functions? Because the nervous system employs pervasive mechanisms which stabilize indicators that get used and eliminates indicators that do not get used. The elimination of unused indicators shows that the indicators acquire a function merely by being incorporated into the nervous system. When that function is not being performed, the indicator is eliminated.

An indicator does not have to be incorporated into such a system in order to have an s-function. The presence of a pervasive mechanism which eliminates indicators that are not being used is a sufficient, but not a necessary, condition for an indicator's having an s-function. The example of synaptic connections merely shows one way in which it is possible to distinguish mere indicators from indicators with s-functions, namely, by their incorporation into a certain kind of system. Indicators existing in systems that are capable of weeding out spurious indicators have the function of indicating whatever they indicate, as opposed to indicators which do not exist in such systems.[23]

Suppose that we have correctly deduced that all indicators in the nervous system initially have the s-function of indicating some condition of the world. These indicators are recruited simply by being stabilized in a system like the nervous system. Their function is an s-function because it is not

---

[23] In Chapter 3, I argue that the very concept of a function entails the existence of a "control structure" which controls the proliferation of tokens of some type, much as spurious indicators are

yet connected to any mechanisms of motor output. Finally, an indicator C with the s-function of indicating F counts as a representation of F, despite not having a structuring causal role in producing behavior.

Thus, modifying Dretske's account slightly, by permitting the existence of indicators which have s-functions, rescues Dretske's account from circularity. Dretske's account is circular only if all meaningful internal states have their meaning in virtue of possessing a structuring causal role. But the possibility of indicators with s-functions shows that some internal states may have meaning without ever having a structuring causal role.

2.4 REPLY TO SOME CRITICISMS OF THE MODIFIED DRETSKEAN ACCOUNT

2.4.1 The Problem of Unspecified Structuring Causal Roles

In defending Lynne Rudder Baker's criticism, a critic might object that even states with the s-function of indicating some condition of the world have their functions in virtue of *some* structuring causal role or another. After all, an organism O recruits a state C via a mechanism M because the indicator properties associated with C are at least potentially valuable. Surely O anticipates (perhaps implicitly) that C will be tied to the production of behavior at some point; otherwise, O would not use M to recruit all those states. Therefore, C really does have meaning in virtue of some structuring causal role or another, even if that role is presently unspecified. And if C has its meaning in virtue of a structuring causal role, then Baker's circle is still in force.

However, I reply that the structuring causal role may not only be unspecified -- it may never exist. There is no reason to think that every state that gets recruited will actually be used. Some states

---

"weeded out" by the nervous system. Hence, the concept of a standby function fits in well with a may possess the s-function of indicating F for as long as those states exist without ever being harnessed to the production of behavior. If someone insists that C has meaning in virtue of a structuring causal role, then she is insisting that C has its meaning in virtue of a state of affairs that may never obtain.

2.4.2 The Problem of Potential Structuring Causal Roles

Suppose the critic continues to push the point. Even if she grants that an s-functional state may never have a structuring causal role, she can still argue that the meaningful state possesses meaning in virtue of a *potential* structuring causal role. The state C is recruited by an organism O because there are some circumstances under which C *would* become useful to O, even if those circumstances never actually arise. Hence, C has meaning in virtue of its potential to occupy some needed structuring causal role for O.

Against the critic, I argue that a state C may have its s-function though C is not even potentially available to be recruited as the structuring cause of some behavior. I propose two examples to establish this point. Imagine a small, furry mammal living in prehistoric times, indeed, the opening days of mammalian life. On several recent occasions, while busily foraging for food, the mammal has seen a large shape glide overhead. The mammal's internal indicator for this shape was stabilized as a result of these incidents, and the mammal consequently possessed an internal indicator with the s-function of indicating the presence of this large shape.

It turns out that the large shape is a pterydactyl (large avian reptile predator of small, furry mammals). Furthermore, the pterydactyl is the very last member of its species. Suppose further that,

---

general analysis of functions. Thanks to Pat Croskery for pointing this out to me.

a few minutes after the small mammal's indicator for the pterydactyl has finally been stabilized, the pterydactyl is suddenly devoured by a ferocious Tyrannosaurus Rex. In this case, the small mammal possesses an indicator which is not even potentially valuable, because there is nothing capable of causing the indicator to produce the indicator token which represents the presence of pterodactyls.[24]

Consider another example. A three-toed sloth is surprised on several occasions by tree frogs. The frogs jump directly in front of the sloth's nose, then hop immediately away. In the customary manner, the sloth stabilizes an indicator with the s-function of indicating the presence of tree frogs. This indicator enables the sloth to represent tree frogs.

Now, suppose that the sloth has absolutely no use for a tree frog representation. The presence of tree frogs is not correlated with the presence of food or predators or other sloths, nor are tree frogs themselves particularly valuable to sloths. Here, we have another case of an s-functional indicator which has not been used to structure any of an organism's repertoire of behaviors.

These two examples illustrate that s-functional indicators may not even have the potential to structure behavior, for two general reasons: first, the environment may alter in such a way that the indicators are rendered useless; second, the indicators acquired may be intrinsically useless to the organism which acquires them. Or, we can imagine cases in which both of these circumstances obtain.

---

[24] The critic may be skeptical. Perhaps a flock of small birds flying in just the right formation could pass for a pterodactyl. However, the example could be amended as necessary to establish this point: if there is even one possible world (nomologically close to our own) in which no existing thing has the ability to induce the indicator to "light up," then the indicator's potential as a structuring cause cannot be a necessary condition for its having meaning.

We can easily multiply cases in which internal states acquire the s-function of indicating some condition F despite the fact that F could never be recruited to structure an organism's behavior. Since these cases show that an indicator can be given an s-function without the indicator's even having the potential to structure behavior, then having the potential for structuring behavior is not a necessary condition for an s-functional indicator's having meaning.

2.5 A FINAL EXTENSION OF THE MODIFIED DRETSKEAN ACCOUNT

If these arguments are correct, then it is possible that *some* internal states have meaning in virtue of having the s-function of indicating F -- full stop.[25] Internal states of this sort fall out of the circle described by Baker's propositions (a) and (b). Against (a), some internal states have meaning solely in virtue of having the s-function of indicating some condition F, rather than in virtue of having structuring causal roles. On the supposition that these internal states may later be recruited as structuring causes of behavior because of what they mean,[26] I conclude that internal states recruited by an M-type mechanism are available for (non-vacuous) explanations of an organism's behavior in terms of the meaning of those internal states.

My arguments to this point have merely vindicated the explanatory relevance of the meaning of internal states recruited by an M-type mechanism. I will now vindicate the explanatory relevance of meaningful states recruited by other sorts of recruitment mechanisms. Consider, for instance, a mechanism L which (unlike M) recruits C because C's indicator properties have resulted in immediate reward for behavior structured by C. Internal state C has the function of indicating F (and, thus,

[25] Where s-functions are conceived as a special, limiting case of functions generally.
[26] Here is the "secondary recruitment process" discussed in footnote 16.

represents F) in virtue of the structuring causal role for which it was recruited by L.[27] But now the circle pointed out by Baker is apparent once again. How does L recruit C because of C's meaning, when C takes on meaning only because it was recruited by L?

I think that Baker's worry can be alleviated by thinking in terms of a hierarchy of recruitment mechanisms.[28] At one level in the hierarchy, recruitment mechanisms are responsible for recruiting internal states simply because of the states' indicator properties. Internal states recruited at this level are merely being given jobs by recruitment mechanisms. The internal states acquire meaning at this level, but the meaning that is acquired cannot, on pain of circularity, explain why the states were recruited.

At a secondary level in the hierarchy, recruitment mechanisms are responsible for recruiting states because these states have the function of indicating some condition of the world. We can think of these recruitment mechanisms as working for relatively sophisticated cognitive systems which direct and integrate the efforts of less sophisticated systems. The more sophisticated systems are "interested" not only in a state's indicator properties but in the fact that these properties already perform some job for the organism. Only in these cases does meaning have an explanatory role. A state must first have acquired meaning before that meaning can cause an organism's behavior to be structured in a particular way.

---

[27] As I indicated earlier, C's acquisition of a structuring causal role is generally sufficient to show that C has meaning.

[28] See footnote 16 above. Also, note that Dretske proposes a similar two-level approach in his response to Baker (Dretske 1991, p. 115). However, since Dretske *accepts* Baker's proposition (a), his two-level approach is subverted. As Baker shows (Baker 1995, pp. 60-61), Dretske cannot *both* accept (a) *and* argue that the structuring causal role explained by C's meaning is not the same causal role which gave C its meaning.

The value of thinking about M-type mechanisms, as opposed to L-type mechanisms, lies in the former's explicit breaking of the tie between meaning and structuring cause. As we consider M-type mechanisms, we are forced to realize that meaning can be acquired by a state independently of the use to which meaning is put. Thus, a Dretskean account of meaning can be vindicated if we posit a hierarchy of cognitive systems, in which some systems give states meaning, while others use that meaning to the benefit of the organism.

CHAPTER 3

FUNCTIONS AND A PUZZLE FOR DRETSKE

In my final chapter, I describe and respond to several problems at the foundations of Dretske's theory of mental representations. I claim that the definitions of two crucial concepts -- "indication" and "function" -- must be reconsidered. First, I present a puzzle used by Stephen Stich to show that Dretske's definition of indication suffers from vagueness. Second, I argue that the problem with Dretske's definition of indication can be solved by thinking more carefully about the concept of a function.

3.1 A PUZZLE FOR DRETSKE

3.1.1 Stich's Puzzle

Dretske defines indication as a reliable covariation between physical tokens of two physical types (Dretske 1988, p. 52). The notion of reliability has both a descriptive and a normative component. In order to determine whether the covariation between two physical types X and Y is reliable, we need to know two things: (1) the objective probability a that Y-tokens will be found in the presence of X-tokens (i.e., the descriptive component) (2) a standard of reliability ß by which to evaluate the objective probability a (i.e., the normative component). By these conventions, X indicates Y only if the objective probability a exceeds the standard of reliability ß.

Stephen Stich has argued, in effect, that there is no principled way of providing values for the standard of reliability ß (Stich 1990, pp. 803-804). Suppose that we want to evaluate the covariation between two physical types X and Y, in order to see whether the covariation is reliable enough to make X an indicator of Y. How accurately must X-tokens predict the presence of Y-tokens?

Additionally, how would we justify the standard that we use to make such evaluations? Unless these questions can be answered, it will be nearly impossible to use Dretske's definition to determine the actual semantic content of any physical state.

In order to show how essentially arbitrary the task of assigning standards of reliability would be, Stich proposes an example and a related puzzle. Imagine a geographical region that is inhabited by a species of bird and two species of butterflies. One species of butterfly, the monarch, has developed a defense against predation by the bird species. If one of the birds eats one of the monarchs, the bird immediately gets sick. Thus, the birds have learned to avoid eating monarchs. However, a second species of butterfly, the viceroy, has taken advantage of the monarchs' defenses; the average viceroy looks almost indistinguishable from the average monarch. Thus, the birds also avoid the viceroys, despite the fact that eating a viceroy does not make any of the birds sick. Finally, imagine that the population of monarchs and viceroys varies systematically along an east-west axis. The east end of the region is populated only by monarchs, the west end is populated only by viceroys, and the population of each butterfly species diminishes as a linear function of the distance from the end that it dominates.

Here is Stich's puzzle: if we imagine that one of the birds undertakes a journey from the eastern to the western edge of the region, when does the bird's indicator cease to be a monarch indicator and become an indicator of something else? Suppose that there are 100,000 butterflies at any given longitude in the region. We can evaluate the reliability of the bird's indicator at any point by (1) assessing the objective probability of generating a false positive (i.e., mistaking a viceroy for a monarch) and (2) comparing this probability with some standard. Does the bird lose a monarch indicator and gain a monarch-or-viceroy indicator when the probability of a false positive is 1:90,000? Why not 1:8,500? or 1:50? The choice of a standard seems to be essentially arbitrary.

3.1.2 Dretske's Reply Is Inadequate

To this criticism, Dretske responds that the only principled way of making such decisions is to require perfect reliability (Dretske 1990, pp. 819-824). Stich's puzzle can be solved by construing reliability as an absolute description. Covariation is either reliable or not; there are no *degrees* of reliability. To count as indication, the presence of tokens of one physical type must *guarantee* the presence of tokens of another type. Thus, in the example above, the bird has a monarch indicator only at the far eastern edge of the region, where the probability of mistaking a viceroy for a monarch is zero.

Of course, Dretske cannot *literally* mean that the reliability must be perfect -- not unless he intends for his definition of indication to apply to nothing whatsoever. Since at least the time of Hume, philosophers have generally conceded that it is impossible to know with certainty (i.e., with a probability of 1) that one physical type will always covary with another. Physical conditions can change. Mechanisms responsible for stabilizing local conditions can break down. Violent winds may blow a viceroy from the west into the previously viceroy-exclusive east. These possibilities cannot be excluded *a priori*.[29]

[29] Dretske might argue that his point is metaphysical, not epistemological. Perhaps he is merely *stipulating* that the reliability must be perfect, even if we cannot know that any particular indicator covaries perfectly with its indicated condition of the world. I do not think that Stich, for one, would be at all bothered by this reply to his puzzle. The proposal immunizes Dretske's theory from Stich's criticism, but at the cost of making the definition of indication empirically useless.

Dretske is obviously aware of these considerations. Hence, he must have in mind something other than genuine certainty. Perhaps he means that an indicator assignment requires the probability of false positives to be negligible. If this is what Dretske means, however, then he has merely shifted the locus of the problem pointed out by Stich. Instead of the old problem of establishing a standard of reliability, Dretske is faced with the new problem of deciding under what conditions his proposed standard obtains. Furthermore, this new problem has precisely the same form as the old problem. Consider again Stich's puzzle. As the bird flies west, when does the probability of mistaking a viceroy for a monarch become high enough to be non-negligible? When is the reliability of the bird's monarch indicator far enough from perfection to undermine its status as a monarch indicator? Requiring perfect reliability has temporarily saved Dretske's definition of indication, but only by hedging it with a question-begging *ceteris paribus* clause. Other things being equal, thinks Dretske, X indicates Y only if X-tokens guarantee the presence of Y-tokens. But under what conditions can the *ceteris paribus* clause be discharged? When is the probability of false positives small enough to be neglected for the purposes of making indicator assignments?

Dretske admits that he has no principled way of answering these questions, but he argues that no one else has been able to answer them, either. He cannot be expected to abandon his position merely because he cannot solve a problem which "bedevils many areas of philosophy" (Dretske 1990, p. 824). Though candid, this response is certainly not going to win over any of Dretske's critics. They might suggest that, in the face of such an intractable problem, the entire project of providing a naturalistic theory of mental representations should be considered misguided.

In the remainder of this chapter, I want to propose a solution to the problem of setting standards of reliability. This solution relies on a particular way of thinking about functions. Hence, I turn to the philosophical analysis of functions.

3.2 THE PHILOSOPHICAL ANALYSIS OF FUNCTIONS

3.2.1 Larry Wright and the Etiological Approach to Functions

As Peter Godfrey-Smith notes in a recent review of the literature on functions, the philosophical analysis of functions has given rise to some important developments, especially in the last twenty years (Godfrey-Smith 1994, p. 344). This recent period is marked by the publication of Larry Wright's seminal paper on functions, in which he recasts functions in etiological terms (Wright 1973).[30] I shall clarify Wright's analysis of functions, examine some of the proposed refinements of his analysis, and pick out some key points that have emerged from this growing dialogue.

Larry Wright's proposed analysis of functions is as follows:

The function of X is F means:

(i) X is there because it does F
(ii) F is a consequence of X's being there.

First, condition (i) is etiological, or historical. The condition appeals to the historical circumstances that are responsible for the stabilization of a trait in a population. In many cases, these circumstances will include some reference to processes of natural selection. To say "X is there" is to say that X is a trait whose existence can be explained through a particular series of historical events.

---

[30] Wright and others use the term "etiology" to emphasize the historical dimension of the causal sequence whereby a certain system comes to occupy a certain state. They are concerned, not just with the causes of X, but with the particular causal sequence that resulted in X. See Appendix A for a more extended discussion of this point.

For example, fish have gills because gills enable fish to breathe underwater. This capacity is explained by appealing to a particular sequence of events that occurred during the evolution of fish.

Second, both conditions (i) and (ii) should be understood as making counterfactual-respecting claims about what a particular trait does. By "counterfactual-respecting," I mean that the trait has its function even if it is not presently exercising its function. For example, it is the function of a muscle to contract, whether or not any given muscle is presently contracting. Without this stipulation, muscles would not have their function unless they were always contracting. Since muscles are often in a relaxed state, awaiting the command to execute their function, any analysis of functions alleged to make sense of the function of a muscle must allow for the counterfactual exercise of functions.

Finally, I want to make clear the primary advantage of Wright's analysis. Wright constructed his analysis with the intent to solve a recurring problem in previous analyses of functions. These analyses inevitably failed to exclude mere "accidental" effects from counting as a thing's function. For example, the heart has the effect of making a "lub-dub" sound, but this is not the heart's function. According to Wright, previous analyses fail to show why this effect could not qualify as a function of hearts. Wright's own analysis excludes the heart's sound-making capacity by requiring us to examine the history of the preservation of hearts in various species. Did the heart's sound-making capacity contribute to the preservation of hearts? If not, then this capacity cannot be a function of hearts. In general, effects that qualify as genuine functions are those effects which tend to contribute to the survival of an organism and, thus, to maintenance of the trait in the population.

Against Wright's account, critics have argued that it is too unrestricted. The generality of his formulation has exposed the analysis to counterexamples. For example, Christopher Boorse argues that, on Wright's analysis, a small rock which holds up a larger rock against the rush of water in a stream bed has the function of holding up the larger rock (Boorse 1976, p. 74). After all, the fact that the small rock is holding up the larger rock explains why the small rock is there (i.e., the small rock is wedged against the stream bed by the larger rock). And holding up the larger rock is a consequence of the small rock's being there. But plainly the small rock does not have the function of holding up the larger rock. Hence, Wright's analysis must be faulty in some respect.

Many philosophers have conceded this and similar counterexamples to Wright's analysis, while retaining Wright's fundamental insight, namely, that an etiological conception of functions can solve the problem of accidental effects. Of the many analyses that retain the spirit of Wright's approach, I will briefly examine (aspects of) the analyses proposed by Ruth Millikan, Peter Godfrey-Smith and Paul Griffiths.

3.2.2 Millikan, Godfrey-Smith and Griffiths on Functions

Consider the detailed proposal made by Ruth Millikan in Language, Thought and Other Biological Categories (Millikan 1984, chaps. 1 & 2). Millikan modifies Wright's analysis by explicitly building an appeal to evolutionary history into her analysis. As in Wright's analysis, appealing to historical categories makes it possible for Millikan to exclude accidental effects. Since these effects do not contribute to the maintenance of a trait in a population of organisms, they are excluded from qualifying as a trait's function.

Because Millikan builds the idea of evolutionary descent directly into her analysis, the account does not apply directly to individual structures in individual organisms.[31] Instead, individual structures have their functions in virtue of being tokens of types over which the variable in the analysis ranges. For instance, a certain organism's heart has the function of pumping blood, because hearts in general have the function of pumping blood. Token structures inherit their functions from the types of which they are tokens.

In addition to solving the problem of accidental effects, Millikan's analysis makes allowances for structures which are not capable of performing their functions. On Millikan's account, a structure retains its function even if it performs its function improperly, because the structure still belongs to the type of which the structure is a token. For example, a broken can opener still has the function of opening cans. A blood cell which cannot transport oxygen still has the function of transporting oxygen. A token structure's inability to perform its function does not deprive the corresponding type of its function.

However, Millikan's analysis is not without problems of its own. For example, Millikan's analysis of functions does not handle organs with vestigial effects. A special case of vestigial effects occurs when the function for which a trait was originally selected is not the function for which the trait was most recently selected. For example, Darwin argued that the swimbladder in blowfish originally developed as a means for flotation and was only later converted to its present use in the blowfish's respiratory system (Prior 1987, p. 315). According to Millikan, any effect is a trait's function if that effect has contributed to the preservation of the trait that exhibits the effect. But this gives to the

[31] Wright seems to have intended that his analysis should apply directly both to individual structures

blowfish's swimbladder the function of aiding flotation, despite the fact that it does not presently have that effect.

In light of this and other problems, Peter Godrey-Smith has proposed a slight modification of Millikan's view. He argues that only effects which have been maintained as a result of the most recent episode in a trait's evolutionary history should count as that trait's function (Godfrey-Smith 1994, pp. 356-358). This stipulation deals effectively with the counterexample to Millikan's analysis. Since flotation is not part of the explanation for the preservation of the swimbladder in the most recent period of this structure's evolutionary history, flotation cannot be the swimbladder's function.

But it is not clear that the analysis that Godfrey-Smith proposes copes adequately with other sorts of vestigial effects. For example, the human appendix presumably performed *some* function in an evolutionarily more ancient episode than the present one. However, since it does not perform *any* function in the most recent episode of its evolutionary history, it is not clear how to classify the appendix. Godfrey-Smith's analysis specifies that a trait's function is that effect which explains the trait's stabilization in the most recent period of its evolutionary history. If there is *no* effect which provides such an explanation, then it is not clear how the trait should be characterized. Should it be said to have no function, or should it be said to have the function that it had in the previous evolutionary episode?

Paul Griffiths has provided an analysis specifically designed to deal with all varieties of vestigial effects (Griffiths 1993, pp. 416-418). Griffiths thinks of his analysis as a hybrid account. He incorporates insights from Wright's paper, as well as insights derived from Robert Cummins's

---

(e.g., "a heart") and to traits in a population of organisms (e.g., "the heart").

important paper on functions (Cummins 1975). Cummins has argued that the notion of a function is related to the idea of a functional analysis. In a functional analysis, a structure's function is given by performing an analysis of the effects that the structure possesses relative to a given system (Griffiths 1993, pp. 410-411). A human heart, for example, transports oxygen from the alveoli of the lungs to the capillaries. In turn, the capillaries permit oxygen to be transfused to individual cells. The cells use the oxygen in producing energy, and so forth. A complete picture of the heart's function requires a complete functional analysis of the system of which the heart is a proper part.

Griffiths wants to combine Cummins's insight with Wright's by arguing that a trait's function is given by a Cummins-style functional analysis of the fitness conferred on a population of organisms by the trait's effects. Fitness is (very roughly) the capacity of an organism to survive and reproduce because of its possession of certain traits (Griffiths 1993, p. 412). Griffiths breaks down the overall capacity to survive and reproduce into the capacities of moving around, finding food, avoiding predators, etc. These capacities might in turn be further broken down, e.g. finding food may involve smelling acutely, digging swiftly, etc. Hence, a trait's function is the capacity it confers upon the organism in an overall functional analysis of the organism's fitness (Griffiths 1993, pp. 410-413).

Like Godfrey-Smith's analysis, Griffiths's analysis explicitly excludes all but the most recently selected effects from counting as a trait's function. If some effect of a trait does not presently contribute to an organism's overall fitness, then the effect cannot count as that trait's function. Thus, the blowfish's swimbladder does not have the function of aiding flotation because it does not presently have that function. Unlike Godfrey-Smith's analysis, however, Griffith's hybrid approach is (clearly) able to deal with the variety of vestigial functions exemplified by the human appendix. Because the appendix does not presently contribute to the fitness of human organisms, the appendix does not have a function. Griffiths's analysis provides the correct answer.

However, the etiological approach to functions has still not clearly carried the day. A final class of counterexamples appears to count against all versions of the etiological approach. Some evolutionary theorists believe that feathers were first used by birds for insulation and only later came to be useful for flying. Now, imagine the first occasion in which a bird uses its feathers to fly, instead of merely for insulation. By any etiological account, this bird's feathers do not have the function of making flight possible, because this function has not been around long enough to explain the maintenance of feathers in the bird population (Godfrey-Smith 1993, pp. 357-358; Prior 1985, pp.

314-317).[32] This result is counterintuitive, especially if we imagine the bird taking advantage of its newly discovered flying skills in the avoidance of starvation, predation, etc.

In general, whenever a trait is suddenly used to perform a function for which there is no selectional history, then any account which conceives of functions as the result of a temporally extended process will necessarily fail (Enc and Adams 1992, p. 641). Thus, it appears that the circumstances to which we appeal in assigning functions need to be synchronic, rather than diachronic; that is, the factors relevant to assigning functions should exist *at the same time* as the effects being evaluated. Since this point runs counter to the entire etiological approach to functions, some philosophers have argued that there must be more than one analysis of functions (Godfrey-Smith 1994, pp. 344-345). The etiological view is appropriate in some cases, but another view must supplement the etiological view in other cases.

3.2.3 Toward a Synchronic Account of Function Ascriptions

Rather than accept a bipartite account of functions, I want to build upon an implicit aspect of Griffiths's account. Griffiths apparently thinks of his account as a variation on the etiological approach, because he specifically invokes past selectional history in discussing functions (Griffiths 1992, p. 412). However, I wish to point out that an appeal to history is not required. Griffiths's account relies on a Cummins-style functional analysis of fitness, and such an analysis can be completely synchronic. That is, we can analyze a trait's effects in terms of the contributions that those effects *presently* contribute to the fitness of some organisms relative to others.

[32] In this case, I draw the example from Godfrey-Smith, while the moral of the example is drawn from Prior.

As evidence for this point, I cite the first version of Griffiths's analysis:[33]

> Where i is a trait of systems of type S, a proper function of i in S's is F iff a selective explanation of the current non-zero proportion of S's with i must cite F as a component of the fitness conferred by i.

Now, observe that the word "must" has modal force. The analysis stipulates what *has* to obtain in order for F to be the proper function of i. The contingent, historical facts about F could be regarded as irrelevant, by this analysis.[34] Hence, this first rendition of Griffiths's analysis could be read as stipulating that functional ascriptions are relative to conditions obtaining in the present -- that is, at the precise moment specified in any given application of the analysis.[35] In short, Griffiths's analysis works fine if we read the analysis as a statement of what must *presently* obtain in order for i to have F as its proper function.

A critic might object that the very notion of a "selective explanation" conceptually entails a historical explanation. That is, appealing to conditions that have obtained in the past *just is* what we mean when we talk about explanations by natural selection. But I disagree.[36] The forces of natural selection are surely as amenable to synchronic specification as are other forces in nature. It is surely possible in principle (if not in practice) to specify the *present* rate of mutation, *present*

---

[33] The final change that Griffiths makes to this analysis is redundant, on my reading of his analysis.

[34] Griffiths's exposition of the analysis makes clear that he is thinking in historical terms, but I contend that the analysis does not actually commit him to an historical construal of selective explanation.

[35] We could, for instance, apply Griffiths's analysis to the ancestor of the human appendix several million years B.C. (assuming that we could find out enough about conditions that obtained at that time). At this point in evolutionary history, the appendix's effects might very well have been actively contributing to the preservation of the appendix in ancestors of humans.

[36] See Appendix A for a more detailed discussion of this point.

levels of fitness, etc.[37] With a specification of present conditions and a working knowledge of natural selection, we should be able (again, in principle) to predict the outcome of adding or deleting any given trait from a population.

While Griffiths's account is open to a non-historical conception of selective explanations, his analysis of functions still requires some modification. As stated, the analysis appeals explicitly to biological concepts like "fitness." Since other disciplines (e.g., physics, psychology) do not use these concepts, it is unclear how Griffiths's analysis should be applied in these other disciplines. I shall attempt to provide an analysis which follows Griffith in providing a Cummins-style synchronic account of functions, although being general enough to apply across disciplines.

3.3 A PROPOSED ANALYSIS OF FUNCTIONS

3.3.1 The Formal Analysis

For the purposes of my analysis, we must imagine a physical system with the following elements:

(1) two physical parameters A and B of a physical system S
(2) a relation F
(3) a range of values T (defined over F)
(4) two physical types X and Y

With these elements in mind, I propose to analyze the meaning of "the function of X is F" as follows:

(i) F is a possible relation between A and B
(ii) T is a range defined over F
(iii) the presence of an X token would be sufficient to explain the presence of an F token [38]

[37] If the equations governing the system are time-dependent, then we must know the system's initial conditions.

(iv) the presence of a Y token would be sufficient to explain the presence of an F token that ranges within T

Condition (i) is intended to be a more precise way of expressing Wright's view that a function is something that a structure "does." I have assumed that anything a structure "does" can be expressed as some kind of covariation between (at least) two physical parameters. For example, the function of a muscle might be described as an equation mapping the distance between the two ends of the muscle to the strength of its contraction. Notice that F must be a possible covariation, where "possible" is construed nomologically. For instance, infinite values for the distance between the two ends of the muscle would be inadmissible.

In general, condition (i) relies on the assumption that most things studied by scientists ultimately can be described via some equation or another. Variables in the equation would correspond to the physical parameters in condition (i). If this intuition is mistaken, then (i) can simply be understood as saying that F is what X does, with corresponding translations throughout the rest of the paper.

Condition (ii) requires the specification of a tolerance T within which the covarying parameters A and B are able to range. In the simplest case, this tolerance would consist of a set of numbers specifying the lower and upper bounds on two ranges. Returning to the muscle example, we might adopt the following tolerances:

Lower bounds -- strength of contraction: 0 ft-lbs
-- distance between ends: 12 centimeters
Upper bounds -- strength of contraction: 30 ft-lbs

---

[38] Obviously, the mere "presence" of an X token would not be enough. I am implicitly assuming that the X and Y tokens are causally linked in whatever arrangement is required to produce the relation F between A and B.

-- distance between ends: 6 centimeters[39]

Two points about T are especially noteworthy. First, T is not a tolerance within which A and B *do* range. Thus, the analysis allows for circumstances in which A and B do *not* range within T. Second, T is not a tolerance within which A and B *should* range; that is, T is not a standard by which to assess values of A and B. Instead, T should be understood as specifying some values within which it is at least nomologically *possible* for A and B to range.

Condition (iii) stipulates that tokens of the physical type X must be sufficient to explain tokenings of the relation F. I want to make it clear, however, that it is in virtue of participating in the type X that a particular token of type X explains a particular tokening of the relation F. For example, suppose that X is the human heart and F is a certain relation between the amount of oxygen in the capillaries and the amount of oxygen in the lungs. If F presently obtains in a particular human body, then the presence of a token of type X would be sufficient to explain the fact that this relation obtains.

The token associated with the type X may not be the only available explanation for the token of F. Suppose a doctor places a mechanical heart alongside my natural heart and hooks them up such that one will begin beating if the other one stops. Even if my natural heart is presently doing all the work, the mechanical heart still provides a part of the overall explanation for F. The covariation of oxygen in my capillaries with the oxygen in my lungs is explained by the fact that I have two hearts, each of which is sufficient to maintain that covariation. Or suppose that my mechanical heart is presently doing all the work. My natural heart has the function of maintaining F, despite the fact that it

[39] The tolerance T may also have to include provisions for a sign, in order to indicate the direction of covariation.

is not actively sharing any of the load. My natural heart is a counterfactual-respecting explanation for F. It explains F even if it is not actively expending any energy to maintain F.

Condition (iv) stipulates that tokens of type Y must be sufficient to explain the presence of token F relations which range within some tolerance T.[40] Call tokens of physical type Y "control structures." A token control structure Y monitors the token F relations that are induced by X tokens.[41] Whenever tokens of X are generated which produce a covariation between A and B that falls outside tolerances T, Y reacts. It constrains the generation of further X tokens, such that only X tokens will be produced which covary within the tolerances established by Y.

A token control structure Y is sufficient to explain production of the F relation inasmuch as it monitors and actively constrains the production of the token structures which do the actual work of producing the F relation. The control structure is situated further up in the "causal stream," as it were, but its explanatory efficacy is no less real or potent than that of the structures which do the immediate work of producing the F relation.

Control structures have properties similar to those of tokens of type X. Like X tokens, Y tokens need not be the only structures capable of explaining A and B's ranging within T. If some other structure induces a nomic connection between parameters A and B, causing them to range within T, then the covariation between A and B is simply overdetermined. Both the Y token and the additional structure explain the covariation.

[40] This tolerance T must be imposed systematically on all tokens of type X, if Y is to count as the control structure for X.

[41] The term "monitors" is probably needlessly intentional. The control structure Y could be as mechanical as a thermostat. It "monitors" in the sense that it is causally hooked up so as to produce some effects upon the production of certain other effects.

Indeed, Y tokens *could not* be the only explanation for the relation that holds between A and B, at the risk of incoherence within my proposed analysis. Why? Because X tokens *also* explain A and B's ranging within T. Whenever some F relation ranges within T, *both* X and Y serve as explanations for A and B's ranging within T.[42] Also, I observe that Y tokens (like X tokens) explain in a way that is counterfactual respecting. Thus, Y tokens need not be actively expending any energy in order to explain A and B's ranging within T.[43]

The notion of a control structure introduces the possibility of a hierarchy of functions. In this hierarchy, the generation of tokens of each type of structure at a lower level is controlled by token structure(s) at a higher level. At the *highest* level is a structure which constrains the generation of all lower level token structures but which is itself unconstrained. This structure would not itself have a function. Moreover, any variations in this structure could be interpreted as random fluctuations imposed on the complete hierarchy of functions by conditions external to the system under investigation.

So much for exposition of the formal analysis. I shall now explain the intuitions behind the analysis.

3.3.2 Explaining the Intuitions Behind the Analysis

---

[42] This feature of the analysis makes good sense on independent grounds. Our purposes are often arranged such that the same behavior is explained by many purposes. For instance, I exercise so that can stay in shape, so that I will be healthy, so that I will live longer, etc.

[43] Conceivably, a control structure Y could be monitoring the production of X tokens all of which were operating within tolerances. During these moments the control structure would not be forced to expend any energy to constrain the production of additional X tokens.

On my view, tokens of a physical type X have a function F just in case a control structure Y constrains the generation of tokens of type X to those tokens which operate within tolerances T.[44] If someone alleges that a certain type of structure has a certain function, then *something* must place constraints upon the parameters associated with the function. The parameters cannot range over *any* nomologically possible values; otherwise, the type of structure in question cannot actually have the alleged function. To take an example, hearts typically pump blood at a rate of between 50 and 180 beats per minute. If there *were no* mechanism which presently explained the fact that hearts pump blood within some tolerances or other, then hearts could be "pumping" blood at 3 beats per minute, 1,000 beats per minute, or whatever. But hearts which have *no* practical operating limits are not hearts which have the function of pumping blood.

My point is not (merely) that a heart pumping blood at 3 beats per minute is not "functioning properly." As I said above, the tolerance T is not a normative notion. It does not specify some range within which a type of structure *should* be functioning. The normative dimension of this analysis lies in its application. No type of structure X *counts* as having a function unless the presence of some *other* structure explains the fact that tokens of X are operating within certain limitations. A type of structure which has no constraints on its operations has no function. Furthermore, this analysis solves all three of the problems that I noted in my discussion of the etiological approach to functions.

Consider, first of all, the problem of specifying synchronic circumstances under which to ascribe functions. I mentioned the example of the first bird which used its feathers to make flight possible, rather than to insulate itself. On my account, it is perfectly reasonable to say that the bird's

---

[44] The word "constrains" in this sentence should be given a counterfactual reading.

feathers have the function of making flight possible. We only need to grant that there exists some control structure which presently constrains within certain limits the effects achieved through the use of feathers, in the event that this use should fall outside of those limits. In the present case, the limits would be established by a certain set of environmental conditions. In most environments, every bird born with defective or missing feathers would fail to survive long enough to propagate.[45] Such birds would be unable to compete with the other birds who learned (through the lucky discovery of one bird) to use their feathers to fly.

Thus, from the moment that a bird first flew, feathers possessed the function of helping birds to fly. Why? Because defective feathers are attached to birds that do not pass along their genes; therefore, defective feathers are not reproduced. As a consequence, defective feathers (or feathers

---

[45] This would not be true in all environments. Imagine that we transported a population of birds to an altitude at which the air was too thin to support flight, even with feathers. In that case, feathers would no longer have the function of making flight possible, because feathers would give none of the birds a competitive edge with respect to flying ability. The mechanism which previously constrained the flying effect associated with having feathers would no longer be operating. Granted a long enough period at this altitude, the birds might even lose their feathers and instead grow fur. (Think of penguins.) Of course, feathers in general might still have the function of making flight possible, because the larger population of birds might still enjoy the beneficial flying effects conferred by feathers.

This scenario raises the question of which population should control the type -- do feathers, or do they not, have the function of making flight possible? Following R.M. Hare's approach to certain problems with utilitarianism, I want to argue that the scenario must be filled out more. Our intuitions cannot be expected to operate properly in the absence of more information. For example, does the scenario stipulate that the two populations of birds are *truly* isolated from one another? Suppose, on the contrary, that non-feathered, high altitude birds occasionally enter the population of feathered, low altitude birds, or vice versa. As the frequency of the mixing increases, the two populations approach a single population, and the problem is dissolved.

If, on the other hand, the frequency of mixing approaches zero, then I no longer have clear intuitions that it would be wrong to ascribe to feathers the flying function in one population and not in the other. It would be as if we were imagining two different worlds, each with their own population of birds. In one world, feathers help birds fly; in the other world, they do not.

used improperly) are curtailed to within certain tolerances. Thus, a certain set of environmental conditions, which obtained at the moment of the first flight, provide the explanation for certain limitations that were immediately imposed upon the flying effects achieved through the use of feathers. The presence of these conditions assures us that feathers do indeed have the function of making flight possible.

The problem of proper functions is also resolved by this analysis of functions. Suppose a particular neuron is presently sending its electrical signals so weakly that these signals cannot interact with neighboring neurons. The neuron is functioning improperly; it is operating outside some tolerance within which it usually ranges. If the neuron *truly* has the function of sending electrical signals, then there must be some mechanism which is responsible for keeping neurons of this type within tolerances. Because of the influence of this mechanism, other neurons of this type are still operating within tolerances. Thus, the neuron still has the function of sending signals. But the fact that the neuron is operating outside the tolerances generally forced upon tokens of the type means that this neuron is functioning improperly.

The critic might argue that I have merely moved the question of proper functioning back a step. For, she asks, what sets the values for the constraints that the token control structure imposes on tokens of the type being constrained? I reply that the controlling mechanism may (but need not) itself be constrained by some other mechanism. Thus, the neurons of the original type are constrained within tolerances that are set by the constraints under which the control structure itself operates. If we continue this line of reasoning, we will eventually find some structure which constrains the operations of other structures but is not itself constrained in its own operations (i.e., the highest level in a hierarchy of functions). Consistent with the analysis, I would argue that this structure does not itself have a function.

Finally, I contend that this analysis solves the problem of accidental effects. This problem could be regarded as the acid test for my analysis of functions, because it is the problem which first induced Larry Wright to introduce the etiological (i.e., historical) approach to analyzing functions. By rejecting Wright's approach, I implicitly commit myself to providing a new way of solving the problem.

Recall that Wright's analysis of functions relies on a historical construal of explanation. A function is some effect which a structure is capable of producing and which has in the past contributed to the stabilization of the structure. Thus, effects that should count as functions can be distinguished from those that do not, because effects that do not make a contribution to the structure's stabilization are not functions of the structure.

Under my analysis of functions, a somewhat different method emerges for dealing with accidental effects. Consider any arbitrary set of token structures of a physical type X. Suppose that the X tokens generated at time t1 are each capable of producing effects F1 - Fn. Now imagine a time t2 later than t1. (Assume that the circumstances under which X tokens are generated at t2 are roughly the same as those under which X tokens are generated at t1.) I claim that mere accidental effects could be distinguished from the functional effects if some of the effects F1 - Fn fail to exhibit systematic constraints as t2 becomes much greater than t1. Such effects are the accidental effects.

For example, if the "lub-dub" sound produced by the heart is an accidental effect *and* if we imagine a sufficiently extended period of time in which token hearts are generated, then my analysis predicts that we will witness a proliferation of token hearts that produce noises without any systematic constraints on this effect.[46] Of course, we might say that all effects produced by token hearts are necessarily constrained, inasmuch as hearts cannot produce similar "lub-dub" sounds at a rate of (say) 10,000 beats per minute. However, these sorts of constraints do not count, since *every* effect is constrained by the fact that certain rates are not physically possible.[47] Only constraints that restrict an effect to within some subset of its possible operating range count as function-making constraints.

In order to show that my analysis does not work, it would be sufficient for a critic to produce a relatively realistic case in which a token structure of some physical type exhibits systematic constraints over long periods of time but which we would not want to call a "function" of the structure which produces the effect.[48] Alternatively, the critic could produce a relatively realistic case in which some structure clearly has some effect as its function, despite the fact that the effect does not exhibit systematic constraints.

3.4 APPLYING THE ANALYSIS OF FUNCTIONS TO STICH'S PUZZLE

3.4.1 Several Preliminary Points

Let us apply this analysis of functions to the problem discussed earlier, viz. setting standards of reliability by which to assess indicator relations. Recall the puzzle used by Stich to illustrate the essentially arbitrary nature of the task of setting standards of reliability for indicators. As a bird flies from east to west -- i.e., from a perfect probability of meeting only monarchs to a perfect probability

[46] The term "systematic" could be defined pragmatically using some suitable statistical measure.
[47] The formal analysis takes this point into account, by specifying that the relation F must be a possible relation.
[48] As with the notion of systematic constraints, I believe that the length of this period of time could be established by the explanatory context.

of meeting only viceroys -- when does its indicator change from a monarch indicator to a monarch-or-viceroy indicator? I shall argue that, if the bird's monarch indicator is conceived as part of a hierarchy of functions, then it is possible to provide a principled solution to Stich's puzzle.

Proposing this solution clearly requires making two initial points. First, I must fix a small flaw in the original puzzle. Stich asks when the bird's indicator changes from an indicator of monarchs to an indicator of monarchs-or-viceroys. However, he *also* stipulates that the indicator cannot detect the difference between monarchs and viceroys. (This stipulation must be correct. For, if the difference *could* be detected, then there would be no puzzle. The indicator would simply send out one signal when it detected monarchs and another when it detected viceroys.) But how could the indicator's content possibly be changed if (by stipulation) it cannot detect the difference between these two kinds of entities?

Stich has erred by thinking of the mere *indicator* as the monarch representation, rather than thinking of the internal state which has the *function of indicating* the presence of monarchs as the monarch representation. Stich's puzzle makes sense only if we ask when the bird's representation (rather than the mere indicator) changes from a representation of monarchs to a representation of monarch-or-viceroys. If we do not change the puzzle in this way, then it simply loads the dice unfairly against Dretske's views. The indicator itself cannot be the element in the bird's representational system which switches the contents of the representation from monarch to monarch-or-viceroy; by stipulation, the indicator does not have the resources for accomplishing this task. Hence, either something *additional* is required to accomplish this task, or nothing at all can accomplish it (and Dretske's system is consequently inadequate). I argue that Dretske's system has the resources for addressing Stich's puzzle but that the puzzle must be formulated in such a way as to make it possible to take advantage of these resources.

My second point is a reminder. According to Dretske, a representation is an internal state which has the function of indicating the presence of something or other. I argued in Chapter 2 that we must make a distinction between internal states that have the function of *indicating* and internal states that have the function of *representing* (having previously been given the function of indicating). The first kind of internal state is a mere representation, that is, one which has not become hooked up to systems of motor output. The second kind of internal state is a representation which causes some kind of behavior. This internal state has the function of representing something; it also has the job of initiating behaviors in virtue of what it represents. This distinction is important, because I am going to argue that Stich's puzzle can be solved by applying my analysis of functions at the level of the *second* of these two kinds of internal states. I shall argue that we must consider the constraints imposed upon an internal state which has the function of representing something, that is, representations capable of initiating behavior.

3.4.2 The Solution to Stich's Puzzle

Finally, the puzzle and its solution: when will the bird's monarch representation become a representation of something else? According to Fred Dretske, the bird's representation of a monarch consists of an internal state which has the *function* of indicating the presence of monarchs. I have argued that a structure has a *function* in virtue of being a token of a type which imposes constraints on some parameters, while itself being constrained by some other structure. Stich's puzzle can be solved by using these two concepts in tandem -- that is, by importing the proposed analysis of functions into

Dretske's definition of a representation. A representation is an internal state which indicates the presence of something *and* for which there is some control structure which constrains the indicator relation to within some tolerances. Hence, I propose to let the indicator's control structure, which constrains the operations of the indicator, set the values of reliability for the indicator.

Let me illustrate with the following scenario. The bird is flying west. It has not eaten for a few minutes and is hungry. The bird spots a butterfly that it takes to be edible and gobbles it up. The butterfly turns out to be a monarch, and the bird soon becomes very sick. The bird has mistakenly eaten one of the butterflies that it (and other members of its species) generally avoids. Furthermore, let us imagine that the insects upon which the bird normally preys are somewhat scarce, or that the lighting is particularly poor, or in general that some condition obtains throughout the region such that the bird is forced to make frequent errors. Suppose that the rate at which the bird makes such mistakes is relatively constant. The bird is minimizing as much as possible the number of monarchs it eats accidentally, but it has already reached an optimum level of efficiency in monarch-avoidance.

As the bird journeys west, it will continue to make mistakes and to get sick from eating monarchs. But, eventually, it is likely to encounter a viceroy. And as the bird approaches the western end of the region, the probability of eating a viceroy will continue to increase. Finally, the bird will reach a point at which the probability of meeting a monarch, rather than a viceroy, is so low that the bird can eat viceroys with relative impunity. Presumably, if its learning systems are working correctly, it will add the viceroy to its diet.[49]

[49] Valerie Hardcastle has pointed out to me that real birds do not actually learn to eat something that has once made them sick. Avoidance learning of this sort is apparently "one shot." But the reader is

On my solution to Stich's puzzle, the bird no longer has a monarch representation when the constraints appropriate to a monarch representation are replaced by constraints appropriate to a viceroy representation. Instead of avoiding the butterflies, the bird begins to eat butterflies. The internal state which previously had the function of representing monarchs now has the function of representing viceroys. Put another way, the internal state which used to initiate avoidance behavior now initiates eating behavior. Why did this occur? Because the representation was functioning outside tolerances. It was initiating avoidance behavior when the bird was eating perfectly edible butterflies. The control structure, monitoring the relationship between the representation and the initiated behavior, altered the constraints being imposed in light of the altered circumstances.

Notice that this may not involve any changes at the level of the internal state which has the (mere) function of indicating the presence of the butterflies. The bird may need to pick out viceroys correctly, for eating purposes, with the same level of accuracy that it previously picked out monarchs, for avoidance purposes. However, the purpose for which the butterfly representation exists is constrained in a completely new way. A new behavior must henceforth be initiated upon indication of the presence of a butterfly.

Lastly, this account makes possible genuine precision with respect to fixing the content of the representation. The bird's monarch representation becomes a viceroy representation at precisely the moment when the learning mechanisms that connects indicators with motor systems engages the new behavior. If the bird will tolerate eating 1 monarch for every 9 viceroys, then the bird will acquire a viceroy representation when it has traveled precisely 90% of the way across the region described

---

asked to overlook the element of unrealism in this example, for the sake of continuity with previous (where the ratio of viceroys to monarchs is 9:1). That is the point at which the bird now has a viceroy representation and not a monarch representation. A principled solution to Stich's puzzle can be given by determining precisely when the behavior appropriate to the representation would be hooked to the representation by a learning mechanism.

---

examples.

CONCLUSION

If Fred Dretske is correct, then a physical thing has (non-derived) meaning in virtue of having the function of indicating some condition of the world. And it appears that at least some physical systems (i.e., biological systems of audio localization) have the properties required by Dretske’s theory. However, I have argued that Dretske's theory faces several serious charges. Lynne Rudder Baker charges Dretske's theory with circularity; Stephen Stich charges it with vagueness.

My reply to Baker's charge claims that Dretske's theory is circular unless some indicators have a function without having a structuring causal role. Realizing this possibility requires the notion of a standby function, that is, the notion of a state which has its function solely in virtue of being incorporated into a structure which “weeds out” those states that do not perform their function.

My reply to Stich's charge suggests that a principled solution to his puzzle arises from an appropriate understanding of functions. An effect of a structure is a function of that structure only if there are some tolerances within which token structures of the type are constrained to operate. The tolerances established by the controlling mechanism provide the resources necessary for answering Stich's charge.

In general, these replies suggest that a key element in defending Dretske’s views lies in a careful scrutiny of a basic notion in his theory -- the notion of a “function.” Conceiving of functions as effects of a structure that are imposed on it by virtue of its taking part in a system which lays certain constraints on tokens of the same type as the structure in question allows us to imagine the existence of structures with standby functions. It also permits us to incorporate into Dretske’s conception of meaning the resources needed to make that conception as precise as his critics require.

REFERENCES

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

Baker, L.R. 1995. Explaining Attitudes. New York, NY: Cambridge University Press.

Barinaga, M. 1994. Neurons tap out a code that may help locate sounds. Science, 264(5160), p. 775.

Bechtel, W., and Abrahamsen, A. 1991. Connectionism and the Mind (Cambridge: Basil Blackwell).

Boorse, C. 1976. Wright on Functions. Philosophical Review 85, pp. 70-86.

Cummins, R. 1975. Functional Analysis. Journal of Philosophy 82, pp. 741-765.

Dretske, F. 1981. Knowledge and the Flow of Information. Cambridge, Mass: MIT Press. A Bradford Book.

Dretske, F. 1988. Explaining Behavior. Cambridge, Mass: MIT Press. A Bradford Book.

Dretske, F. 1990a. Precis of "Explaining Behavior." Philosophy and Phenomenological Research, (50)4, 783-793.

Dretske, F. 1990b. Putting Information to Work. In Information, Language, and Cognition, ed. Philip Hanson (Vancouver: University of British Columbia Press).

Dretske, F. 1990c. Precis of "Explaining Behavior: Reasons in a World of Causes." Philosophy and Phenomenological Research 1(4), pp. 783-786.

Dretske, F. 1990d. Reply to Reviewers. Philosophy and Phenomenological Research 1(4), pp. 819-839.

Dretske, F. 1991. How Beliefs Explain: Reply to Baker. Philosophical Studies, 63, 113-117.

Dretske, F. 1992. The Nature of Thought. Philosophical Studies, 70, 185-199.

Dretske, F. 1993. Mental Events as Structuring Causes of Behavior. In Mental Causation, eds. John Heil and Alfred Mele (Oxford: Clarendon Press).

Dretske, F. 1994. Misrepresentation. In Mental Representation: a Reader, edited by Stephen Stich and Ted Warfield, Blackwell Publishers, pp. 157-173.

Dretske, F. 1995. Naturalizing the Mind. Cambridge, Mass: MIT Press.

Enc, B. and Adams, F. 1992. Functions and Goal Directedness. Philosophy of Science 59, pp. 635-654.

Fodor, J. 1994. The Elm and the Expert: Mentalese and Its Semantics. Cambridge, Mass: MIT Press. A Bradford Book.

Godfrey-Smith, P. 1989. Misinformation. Canadian Journal of Philosophy, 19, 533-550.

Godfrey-Smith, P. 1994. A Modern-History Theory of Functions. Nous 28(3), pp.344-362.

Griffiths, P. 1993. Functional Analysis and Proper Functions. British Journal for the Philosophy of Science 44, pp. 409-422.

Handel, S. 1988. Listening (Cambridge: MIT Press). Hanson, P (ed). 1990. Information, Language and Semantics. Vancouver: University of British Columbia Press.

Hardcastle, V. 1994. Indicator Semantics and Dretske's Function. Philosophical Psychology, 7, 367-382.

Irvine, D.R. 1992. Maps of frequency and space in the mammalian auditory system: generation and plasticity. Australian Journal of Psychology, 44, pp. 131-138.

Jenkins, W.M. and Masterton, R.B. 1984. Role of cat primary auditory cortex for sound localization behavior. Journal of Neurophysiology, 52, pp. 819-847.

Jenkins, W.M. and Masterton, R.B. 1982. Sound localization: effects of unilateral lesions in central auditory system. Journal of Neurophysiology, 47, pp. 987-1016.

Konishi, M. 1993. Listening with two ears. Scientific American, 268, pp. 66-73.

Middlebrooks, J.C. and Green, D.M. 1991. Sound localization by human listeners. In Annual Review of Psychology, edited by M. R. Rosenzweig, and L. W. Porter, Annual Reviews, Inc.: Palo Alto, CA, pp. 135-159.

Middlebrooks, J.C. 1992. Narrow-band sound localization related to external ear acoustics. Journal of the Acoustical Society of America, 92(5), pp. 2607-2624.

Middlebrooks, J.C., Clock, A.E., Xu, L., and Green D.M. 1994. A panoramic code for sound location by cortical neurons. Science, 264(5160), pp. 842-844.

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

Neti, C., Young, E.D., and Schneider M.H. 1992. Neural network models of sound localization based on directional filtering by the pinna. Journal of the Acoustical Society of America, 92(6), pp. 3140-3156.

Poon, P.W., and Brugge, J.F. 1993. Sensitivity of auditory nerve fibers to spectral notches. Journal of Neurophysiology, 70(2), pp. 655-666.

Prior, E. 1985. What is Wrong with Etiological Accounts of Biological Function? Pacific Philosophical Quarterly 66, pp. 310-328.

Samson, F.K., Clarey, J.C., Barone, P., and Imig, T.J. 1993. Effects of ear plugging on single-unit azimuth sensitivity in cat primary auditory cortex. I. Evidence for monaural directional cues. Journal of Neurophysiology, 70(2), pp. 492-511.

Segal, G. and Sober, E. 1991. The Causal Efficacy of Content. Philosophical Studies, 63, 1-30.

Stich, S. 1990. Building Belief: Some Queries about Representation, Indication, and Functions. Philosophy and Phenomenological Research 1(4), pp. 801-804.

Stich, S. and Warfield, T (eds). 1994. Mental Representation: a Reader. Cambridge, Mass: Basil Blackwell.

Takahashi, T.T. 1989. The neural coding of auditory space. Journal of Experimental Biology, 146, pp. 307-322.

Warfield, T. and Stich, S. 1994. Introduction. In Mental Representation: a Reader. Cambridge, Mass: Basil Blackwell.

Wightman, F.L., and Kistler, D.J. 1992. The dominant role of low-frequency interaural time differences in sound localization. Journal of the Acoustical Society of America, 172(4), pp. 425-438.

Wright, L. 1973. Functions. Philosophical Review 82, pp. 139-168.

Zakarauskas, P., and Cynader, M.S. 1993. A computational theory of spectral cue localization. Journal of the Acoustical Society of America, 94(3), pp. 1323-1331.

APPENDIX A

CONCERNING EXPLANATIONS BY NATURAL SELECTION

A.1 EXPLANATIONS BY NATURAL SELECTION

In this appendix, I argue that proponents of the etiological approach to analyzing functions have been lead astray by the view that explanations by natural selection are necessarily historical. First, I try to account for the intuition that explanations by natural selection are necessarily historical; that is, I try to show why philosophers might think that nothing could be a selectional explanation without also being a historical explanation. Second, I argue that this intuition is simply false. It is both possible and desirable to recognize a non-historical conception of explanations by natural selection.

A.1.1 What Good are Explanations by Natural Selection?

Classically, explanations by natural selection (henceforth, selective explanations) have been used by researchers to infer the distant origins of species observed in the present. By examining the traits presently possessed by some species and modeling various circumstances under which the traits might have developed, researchers hope to learn something about the nature of ancestral organisms and the environments in which they lived. Of course, there are more purposes to which the theory of natural selection can be applied than the question of origins. However, the question of origins has historically been important.[50] In fact, I submit that the historical importance of this question provides

[50] I do not have the space to defend this point here, but I believe that few will dispute the claim that natural selection was early seen as a means of liberating science from the burden of a supernatural, seven day long conception of origins.

at least one possible explanation for the fact that philosophers think that selective explanations are necessarily historical.

It is sometimes difficult to separate a thing's primary use or purpose from its internal character. For example, guns may be used primarily in the perpetration of brutal crimes in our inner cities. However, it does not follow that guns just *are* crime-inducing devices. It is also possible for gunes to be used recreationally or even for the purpose of halting crimes. But guns are often conceived as crime-inducing devices merely because of their primary use in the perpetration of brutal crimes. The purpose for which a thing is used can often be confused with the nature of the thing itself.

Hence, I suggest that, perhaps because of the importance of origins, philosophers' intuitions have been driven by a failure to separate the use to which selective explanations have generally been put from the internal character or essence of such explanations. Philosophers have come to confuse the question, "what *good* are selective explanations?" with the question, "what *are* such explanations?" While it is true that selective explanations are often useful in explaining matters of contingent historical fact, the explanations themselves can be complete with only a specification of initial conditions and a proper understanding of the relevant laws of nature.

This failure to recognize a conceptual distinction between selective explanations and historical explanations can be further illustrated by examining the term "etiology". This term is perhaps most familiar in the context of medical journals in which doctors are concerned with the "etiology" of a disease. In such contexts, the term can refer (1) to the actual cause of the disease or (2) to the set of events which trace the development of the disease. In the first sense, for example, the etiology of the flu is a certain sort of virus. But in the second sense, the etiology of the flu might include contraction of the virus from a surface, weakening of the body's immune system by external conditions (e.g., severe weather), and consequent development of the related symptoms.[51] Thus, etiology can refer to causes *per se* or to the set of events which (either typically or in a particular case) precede or accompany diagnosis of the disease.

We can express the ambiguity implicit in the meaning of etiology as a question concerning whether a set of causes are synchronic or diachronic with the events being caused. On the one hand, if a doctor uses the "etiology of a disease" to refer to the structures which are presently causing certain rates of tissue swelling, etc., then the causes in question are synchronic with the events being caused. Moreover, this sense of etiology corresponds to the non-historical meaning of explanation, in which only initial conditions and certain laws are required to provide an explanation of a physical system in a certain state.

On the other hand, if a doctor uses the "etiology of a disease" to refer to the events which have lead up to a particular point in the development of the disease, then the causes in question are diachronic with the events being caused. The doctor wants to understand the history of the patient (or the structures) exhibiting the disease. This information may be important in determining the exact trajectory being manifested by the disease or in providing data for a better understanding of the disease.

[51] Since symptoms of most diseases generally appear in a relatively well-understood sequence, it is important for doctors to understand the "etiology" of the disease in this second sense. They want to understand the course of events associated with the disease.

In the next section, I argue, in effect, that the latter sense of the term "etiology" does not exclude the former sense. That is, historical and non-historical selective explanations are complementary, not mutually exclusive.

A.1.2 Recognizing a Non-historical Conception of Selective Explanation

It is possible to distinguish conceptually between historical selective explanations and non-historical selective explanations, while losing nothing of the explanatory power provided by selective explanations in general. That is, I contend that selective explanations can be conceived as nothing more than laws of nature governing particular physical systems beginning with particular sets of initial conditions. First, I clarify the intended sense of selective explanation. Second, I discuss in what sense such a conception of selective explanations would be genuinely "selective."[52] Third, I argue that selective explanations in this sense need not be understood as historical.

Explanations by natural selection essentially involve variation and competition. Granted sufficient variation of basic traits and sufficient scarcity of resources among competing organisms, the theory of natural selection predicts that the dominant traits will be those traits which provide the greatest competitive advantage to an organism relative to the advantages conferred on other organisms by other traits.[53]

Now, there is nothing (in principle) which prevents researchers from discovering the present circumstances which drive both variation and competition. What variations are possible? Which are most likely to emerge? How strong is competition? What would the most likely pool of competing

[52] Valerie Hardcastle first put this question to me.
[53] I am ignoring many important philosophical issues (e.g., what should be conceived as the unit of selection, etc.), as these issues are well beyond the scope of this paper.

traits be like? How scarce are resources? Such questions have answers that can be investigated (in principle) synchronically, that is, at the same time as the organisms in question are being observed. For example, having observed what happens to rabbits that do not reproduce very quickly in an area infested with predators, we explain stabilization of the trait of rapid reproduction as the natural outcome of competition between rapidly and non-rapidly reproducing rabbits.

Of course, we would need to specify initial conditions. For instance, the outcome of a situation in which food sources were scarce for a prolonged period of time might favor the less rapidly reproducing rabbits rather than the more rapidly reproducing rabbits. Thus, in order to predict which trait receives the competitive advantage, we must know something about the surroundings. In order to explain why rabbits have this trait, we must project a time evolution of the entire system involving the rabbits, food and prey, beginning with what we know of the initial conditions of this system. Furthermore, *every* selective explanation can be thought as having for its explanandum a complex system for which the state at any given moment is given by its initial conditions and the laws governing time evolution of the system.

A critic might wonder in what sense such an explanation would amount to a "selective" explanation. After all, on my account, the explanation for a given trait is the inevitable product of the time evolution of a certain system beginning with certain initial conditions. Since the state of the system is inevitable, granted the relevant laws and initial conditions, we may properly ask *what* is being selected. Selection implies singling out one thing from a range of alternatives, but my account does not appear to leave room for a range of alternatives.

I submit that the range of alternatives -- thus, the "selective" character of the explanation -- lies in the specification of the competing traits which are predicted to emerge in the overall system. The explanation is selective because it involves the singling out of one trait as the dominant trait from a pool of competing traits. While the singling out is inevitable, this surely does not undermine the selective nature of the process. If anything, the inevitability of the process should reinforce the intuition that one trait is being selected above the others.

The account of selective explanation that I have proposed relies (putatively) on a purely synchronic specification of the system under investigation and the laws which govern it. However, a critic might point out that the initial conditions would not themselves appear to be synchronic with the system at any point in its time evolution. Indeed, these conditions *necessarily* precede any given moment in the system's evolution, making them necessarily diachronic with respect to any later state of the system. Does this weaken the claim that my conception of selective explanation is synchronic rather than diachronic (or non-historical rather than historical)?

I do not think so. No explanation consisting *merely* of the specification of a physical system governed by certain laws with certain initial conditions constitutes a genuinely *historical* explanation. It is true that the initial conditions are necessarily not synchronic with any state of the system examined at an arbitrary point following initiation of the system’s time evolution. However, it is also true that every state of the system is inevitable once those initial conditions are granted. This inevitability, I submit, makes explanations of the sort I am describing non-historical.

Historical explanations are historical, in part, because the state of a system at an arbitrary moment is *not* an inevitable product of previous states of the system taken alone. The contingent facts of history must be taken into account. Historical explanations just are those explanations to which we resort when we seek understanding of a system whose previous states cannot be predicted granted *only* information about states *internal* to the system. We must take into account conditions imposed upon the system being investigated by forces *external* to the system.

At this point, the critic may feel that a much more devastating criticism of my account of selective explanation has emerged. As we think about selective explanations, it appears that there *are* facts about evolving systems which cannot be predicted by examining the systems alone. For example, suppose a flood occurs in a region populated by rabbits that reproduce rapidly. The flood causes a food shortage, but it also changes the characteristics of the environment. It washes away valuable topsoil. The rabbits' food sources have been diminished by a significant degree for the foreseeable future. This contingent fact necessarily alters the trajectory of the evolving system consisting of the rabbits, food and predators. It causes a different prediction about the outcome of dominance of the trait for rapid reproduction in rabbits versus the trait for non-rapid reproduction. Furthermore, the fact could not have been predicted by characteristics internal to the system. Hence, the proper explanation for the selection of the trait of non-rapid reproduction appears to be a historical explanation. And, as frequently as events like this occur to real systems evolving in nature, it would appear that most realistic explanations would end up being historical.

This criticism brings out the final element in my account of selective explanation. When confronted with a system for which the outcomes are not a function of its initial conditions -- that is, when discontinuities are imposed on the system by external factors -- we can simply enlarge the system under investigation. All processes in nature obey natural law. Hence, even the change in conditions imposed externally upon a time evolving system can be taken into account non-historically, if we will only incorporate into the original system these larger conditions. Furthermore, such a move would be the next logical step for a researcher who could already (by assumption) successfully model the properties of a system based on considerations internal to it.

This possibility seems to undermine the claim that selective explanations are necessarily historical. Any discontinuities imposed on a time evolving system by forces external to the system might (in principle) be accounted for by enlarging the system to incorporate the laws that drive those discontinuities. Thus, we ought always (in principle) to be able to predict the outcome of some structure as a function of its initial conditions.

Of course, the fact that selective explanations are not *necessarily* historical would not mean that they could *never* be historical. Practical limitations laid upon our ability to gather data may *force* us to regard discontinuities imposed on the time evolution of a system as factors external to the system. We cannot "enlarge the system," so to speak, because we do not have the required data to enlarge the system. Hence, we are forced to adopt a historical explanation of the system, inasmuch as a more rigorous explanation is not available even if we desire it. Also, there may be circumstances in which it is simply more desirable to characterize in a general way the discontinuities imposed on a system by external forces than to expand our investigations to include a much larger system.

These points are well-taken. However, they argue at most that historical and non-historical type explanations ought to be conceived as complementary rather than mutually exclusive. Practical limitations are not limitation in principle. If the critic will only admit that a non-historical construal of selective explanations is at least possible, then my point will have been established. Explanations by natural selection can be complete given only data about events synchronic with the events or structures being explained.

VITA

Jason Richardson

Jason Richardson was born in Brindici, Italy, on April 22, 1970 to a U.S. airman stationed at the U.S. Air Force base in Brindici. Jason was reared in Hampton, Virginia, where he graduated in 1988 from Hampton Christian High School as the Valedictorian of his class. He subsequently attended Covenant College in Lookout Mtn., Georgia. For academic excellence and demonstrated leadership abilities, he was awarded the four year Maclellan Scholarship. He graduated Magna Cum Laude in 1992 with a B.A. in Interdisciplinary Studies (concentrations in Math, Physics and Philosophy). He presented a paper at the Spring 1996 meeting of the Southern Society for Philosophy and Psychology, and he has been invited to participate in a symposium at the society's Spring 1997 meeting. This thesis was completed in partial fulfillment of the requirements for the degree of Master of Arts in Philosophy at Virginia Polytechnic Institute and State University. Jason specialized in philosophy of mind at Virginia Tech.