FRED DRETSKE'S INFORMATION-BASED THEORY OF INTENTIONAL STATES

By

BRENT MAXWELL SMART

B.A., The University of Calgary, 1983
M.A., The University of British Columbia, 1988

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Department of Philosophy)

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

October, 1997

© Brent Maxwell Smart, 1997

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Philosophy

The University of British Columbia
Vancouver, Canada

Date Oct. 15/97

DE-6 (2/88)

ABSTRACT

The purpose of this dissertation is to advance our understanding of the intentionality of mental states through a critical examination of Fred Dretske's theory of mind.

Dretske's theorizing is constrained by his desire both to save central features of our "pretheoretic" conception of mind and, at the same time, to be appropriately "materialistic" (and hence "naturalistic"). Dretske's materialism is captured in his claim that: "...the crucial point is that whatever set of facts we select to analyze the mental are facts which, taken individually, are recognizably physical -- the sort that exist, or can exist, in a world devoid of minds" (Dretske 1994a, p. 131). Central features of our pretheoretic conception of mind that Dretske wishes to preserve include:

- that ordinary categories such as thought, belief, and desire capture real categories of mental states,
- that mental states are certain kinds of representational states, and therefore they have semantic content (meaning),
- that a mental state's having a particular representational content is an explanatorily relevant (because causally relevant) fact about it.

Dretske proposes that intentional, representational contents of the cognitive states of a cognitive system are fixed by the history of information-bearing relations (also called "indicational" relations) that those states of the system have had to aspects of the system's environment. Further, the resulting representational contents are themselves essentially and entirely relational in nature. One cognitive system with an internal configuration type-identical to another, but with type-distinct (or in the limiting case no) history of

indicational relations to its environment, will have type-distinct (or in the limiting case no) representational contents.

Within these constraints, then, Dretske must show both how mental representation is possible and also how it is causal-explanatorily relevant to behaviour. As such, it is indicational relations that need to provide the basis both for an account of how a cognitive state comes to have its given content, and also for an account of the causal-explanatory role that its having that very content may have for the behaviour of the cognitive system whose state it is. However, I argue that indicational relations *per se* can do neither. This is so particularly on Dretske's own, strong, conception of indication. Buiiit even on weaker conceptions proposed by others as a remedy, factors which ought to count as intentional by Dretske's own lights (and which are importantly non-relational) become important contributors to intentional content determination in a way that would commit the overall account to a general and pervasive circularity. I argue that some of these factors, having to do with matters of "receptivity" and perceptual "discrimination", are endemic to the very processes of operant conditioning that Dretske must appeal to as contributing to the rise of states with intentional contents.

TABLE OF CONTENTS

Abstract	ii
Acknowledgement	v
Introduction	1
Chapter One Behaviour and Causation in the Information Theory	7
a. Behaviour	7
b. Causation	16
Chapter Two Indication and Information	25
a. Information	25
b. Indication	33
(i) Indication: Type or Token?	36
(ii) Indication and Dependency	40
(iii) H-indication	47
Chapter Three Representation and the Design Problem	63
a. Introduction	63
b. Type I, II, and III Representational Systems	66
c. Solving the Design Problem: When Does Information Work?	70
Chapter Four Motivational States	89
a. Introduction	89
b. Drives and Desires	92
(i) Desires	95
(ii) The Drive/Desire Distinction	101
c. Motivational States and Behaviour: Sniffy Gets His Oats	118
(i) Sniffy	119
(ii) Imposing the Model	120
(iii) Procedural Beliefs	124
Chapter Five The Critics	135
a. Introduction	135
b. Baker's Circle	136
c. What is the Causal/Explanatory Role for Semantic Properties?	145
d. Does Natural Meaning Have a Causal/Explanatory Role?	156
e. Are We Swamp-persons?	170
f. Should the Indication Conditions be Revised?	179
Chapter Six Prospects for Dretske's Informational Approach to Mental States	199
a. Introduction	199
b. The Design Problem Again	200
c. The End	220
Works Cited	228

ACKNOWLEDGEMENT

I wish to thank all of the members of my supervisory committee, Dr. Gary Wedeking, Dr. Andrew Irvine, and Dr. Phil Hanson, for their kind support, help, and encouragement. I am especially grateful to Dr. Hanson for his direction and for patient reading of so much preliminary material. In addition, I would like to acknowledge and thank two of my teachers from the University of Calgary, Gordon M. Grieg and C.B. Martin, for their guidance. Finally, I am above all grateful to N.A. for showing me a new way of life.

INTRODUCTION

Some of the most difficult philosophical questions are those that ask us to make sense of ourselves. In particular, the issue of the nature of mind, how it informs and how it interacts with the natural order is an issue that has lived in philosophical discussion for centuries. It continues to do so.

Fred Dretske is an important contemporary philosopher who addresses the nature of mind through a theory he has developed over nearly two decades. The theory he offers is a twin theory which is intended both as a view on the explanatory role of mental states in behaviour, and as a view on how those states come to be states with meaning-properties. The result is an impressively detailed body of work, consistently and patiently defended.

In the "Prologue" to his recent book, *Naturalizing the Mind*, Dretske gives a name to his approach: "Representational Naturalism". Here is a succinct general statement of what this means:

> The purpose of these lectures is to promote a naturalistic theory of the mind -- something I call the Representational Thesis. The thesis, in two parts, is that, plus or minus a bit, (1) *All mental facts are representational facts*, and (2) *All representational facts are facts about informational functions*. (Dretske 1995, p. xiii)

Some of Dretske's basic commitments are clearly evident in this short passage. He takes one of the traditional conceptions of minds and mental states in his commitment to representationalalism. Mental states generally, in his view, are states which have meaning properties. His account of mental states having such properties will be in terms of "informational functions". So the approach Dretske takes to mental representation is one in which representational properties are founded upon a notion of information or information-carrying. His preferred technical term for the carrying of information is "indication".

Dretske develops his indication-based theory of mental representation within relatively austere constraints. He intends his theory to be naturalistic: what mental states are and how they come about must be explicable in terms of the "natural order". Dretske takes this as implying that there are ways that states can mean something without their having been assigned a conventional interpretation in the way that, say, a word in a language may be assigned an interpretation. One being's mental states, for example, cannot simply depend for their representational properties on an interpretation given them by another mind.

In a short article which nicely lays out Dretske's working commitments, he begins as follows:

> I am a materialist. Mental states are simply physical states of the organism. The physical states (or facts) with which mental states (or facts) are to be identified may be very complex, of course; they may be, or include, facts about the *relations* that exist (or existed) between an organism and its environment. The physical relationships underlying the mental life of an organism may be causal, informational, functional, or historical. I happen to believe that the important relations are informational and historical: facts having to do with the way an organism developed (during learning) to more efficiently service its own needs. But, from the point of view of materialism, the crucial point is that whatever set of facts we select to analyze the mental facts are facts which, taken individually, are recognizably physical -- the sort that could exist in a world devoid of minds. (Dretske 1994a, p.131)[1].

The "set of facts" Dretske selects to analyze mental phenomena are relational ones. The representational content of an organism's mental states will be a function of historical, informational relations. In effect, mental representational content is an interpretation of an organism's physical (neurological) states. The interpretation's being naturalistic depends on the possibility of such interpretations being able to come about (through historical and informational relations) in a way that does not depend on an interpreting agent. I take this to

mean, not only that the "interpretations" of my mental states cannot ultimately depend on other cognitive beings and their mental states, but that some base level mental states, at least, cannot depend on other mental states of mine, either. The analytical problem of naturalistically accounting for mental representation will not have been solved if, at the level of basic states, determination of representational content depends on still other mental states.

Dretske sets for himself the difficult problem of making mental states matter within the constraints of his materialist version of naturalism. He is clearly committed to such states as beliefs and desires being explanatory, and being so because they are causal:

> I think we really do have thoughts and desires. Our thoughts are *about* things and our desires are *for* things. Furthermore, these thoughts and desires are causally efficacious, and they influence behavior in virtue of what they are thoughts about and desires for. We really do see, know, and remember, and that is why it is predictively useful to describe ourselves as seeing, knowing, and remembering. It is not, as some have argued, acceptable (or true?) to talk this way merely because it is predictively useful or convenient. (Dretske 1994a, p.131)

Nor is he satisfied simply with states that happen to be meaning-bearing mental states being causal. He commits himself both to the first task of giving an account of what mental states (eg. beliefs) are, but also of what, *qua* mental states, they do. As Dretske says "[t]here is no use having a theory of representation, and showing that the brain is a representational device, if the fact that the brain represents conditions in the animal's environment doesn't help to explain something about the way the animal operates in that environment" (Dretske 1994a, p. 135).

That a physical structure could have a meaning property and be efficacious without the meaning property (or the fact that the physical structure has it) itself being efficacious is what has come to be called "The Soprano Problem". The problem owes its name to an example

Dretske uses to illustrate it:

> Something possessing content, or having meaning, can be a cause without its possessing that content or having that meaning being at all relevant to its causal powers. A soprano's upper-register supplications may shatter glass, but their meaning is irrelevant to their having this effect. Their effect on the glass would be the same if they meant nothing at all or something entirely different. (Dretske 1988a, p. 78)

Solving the Soprano Problem for mental states is a matter of showing how they are causally relevant in virtue of being states with representational properties. This, in addition to giving a plausible account of mental states' being representational states, is what Dretske takes to be his main task:

> That, I say, is the second major task that a materialist (who, like myself, is a realist about the mind) faces. This is the task of saying what explanatory difference mind-facts (in this case, facts about what the brain represents) make in the life of an organism. This is the task I have undertaken in *Explaining Behavior*; there I argue that it is only in the learning situation, only in the *ontogeny* of cognitive systems, that the representational aspect of brain events, the fact that the brain represents things as thus and so, acquires a genuine relevance in the causal explanation of the animal's behavior. (Dretske 1994a, p. 135)

To sum up, Dretske's development of a theory of mental representation occurs within a framework of materialistic naturalism. Mental states are physical states, the interpretation of which is a function (largely) of historical, informational relations. As a result of a state having a history of information-carrying (a history of indicating something) it can acquire a function of indicating. For a mental state, such as a belief, to be a representation is for it to have a function to indicate. When this function comes about in the right way, it has an explanatory role to play in behaviour. These are theoretical details that will require careful development.

Many have been pessimistic about the prospects of succeeding in the task Dretske sets

for himself within the constraints he sets for himself. I am among the pessimists. I do not believe that the informational relations that Dretske identifies are the crucial determinants in the determination of mental representational content. Nor do I believe that they are sufficient for such determination. I will argue that Dretske's indication relation does not do the work he requires of it, and that it must do in shaping behaviour, in order for it to determine the contents of a cognitive system's representational states. I will claim both that indication is not causal in the way that it needs to b e, and that the determination of mental content depends on other factors as well.

My focus in developing this work has been Dretske's account of intentional mental states. This account is mainly found in *Explaining Behavior* (Dretske 1988a). Because it is intentional states upon which I focus, and because Dretske thinks of mental states generally as representational, I use both "representational content" and "intentional content" to refer to the same phenomena.

In the following chapters, I explicate and examine Dretske's view of the representational content of intentional states, including his account of motivational states. Crucial to an appreciation of Dretske's theory is his view of behaviour as a process, structuring versus triggering causes, his notions of information and indication, and the distinction between a state being an indicator and a state having the *function* to indicate. It is the latter that is required for the sort of meaning that one would call representation (on Dretske's view). The development of Dretske's theory occurs in Chapters 1 to 4, although in Chapter 4 I begin to raise critical questions concerning Dretske's account of motivational states, their classification, and their role in the formation of learned behaviour. Chapter 5 is given over to the development of

important criticisms of Dretske's approach: that his theory is either circular or else fails to make mental representation explanatorily relevant, that the Dretske conception of indication does not do the causal work it needs to do, and that the kind of criterion indication is (based on reliable correlation) is the wrong kind of criterion as a general determiner of content. Chapter 6 brings together my views as to why the basis of Dretske's account is insufficient to account for the meaning properties of intentional states. In addition to concerns about the causal relevance of indication in developing behaviour, which Dretske requires, I argue that there are traces of mental representation already required in the situations (learning situations) which are supposed to produce it. As Dretske is providing a "ground-up" account of how active representational properties of intentional states are possible, this will not do.

NOTES

1. Of course it is notoriously difficult to specify exactly what would be included as "materialistic" or "recognizably physical". Certainly neurophysiological events and states of organisms would be materialistic. So, too, would recognizably physical events such as the presence of water, the presence of electricity in a circuit, the illumination of a lamp, the barking of a dog . Psychological events or states, couched in mentalistic language referring to thoughts, beliefs, desires, expectations, suspicions, and the like may be physical, but they are not as such "recognizably physical". It would remain to be shown that an event or state described under terms of mentalistic language was indeed "recognizably physical". It would thus remain to be shown that such a phenomenon was in Dretske's sense naturalistic.

CHAPTER ONE

Behaviour and Causation in the Information Theory

Two unique pieces of Dretske's approach are his accounts of causation and of behaviour. What he says about these two topics is central to an understanding of his information theory. Causation can be second-order (causings of causings) as well as first-order (causings of events). Behaviour is itself a causal process. This conception of behaviour is vital to the overall theory, for Dretske will claim that the causal role played by information is the role of structuring (types of) behaviour processes. I will discuss the Dretskean view of behaviour first.

a. Behaviour

Dretske argues that it is a mistake to think of behaviour as a simple event, to identify it as bodily movements for example. Suppose I am carrying a cup of coffee and Jones bumps into my arm so that coffee is spilled. A movement of my arm is involved in the spilling of the coffee, but that movement is not behaviour. In Dretskean terms, it is something that happened to me and not something that I did. By contrast, a case of intentional action, my deliberately moving my arm so as to spill coffee on Jones, is behaviour. The moving of the arm is something that I do. However, the distinction is not an intentional/unintentional one. If my arm moves so that coffee spills when I try to read my watch, I have neither intentionally spilled coffee nor have I intentionally moved my arm. The arm does move in such a case and its moving, on Dretske's classification scheme is something that I do.

So the distinction between a behaviour, which includes a bodily movement, and non-behavioural bodily movement[1] lies in where the subject (S) figures in the movement. A mere bodily movement (M) that is not part of a behaviour is something that happens to S. Roughly

speaking M has an immediate cause which is external to S. In the example described previously, the immediate cause in question is Jones (or Jones' contacting my arm). Behaviour requires a bodily movement the cause of which is internal to S. So Dretske identifies behaviour as a process, the process of some internal (physiological) state's (C's) causing M.

Extended processes will count as behaviour. On Dretske's conception behaviours can be "nested" within other behaviours. An internal state C of John Wilkes Booth (to use a Dretske example) causes a certain set of hand/arm movements M. This is a piece of behaviour (B_1). B_1 is "nested" within B_2: C's causing of M causes the firing of a revolver (R). R causes Lincoln to be struck by a bullet (S). By transitivity of causality, C causes S (via M), B_2 is itself nested within a more spatio-temporally extended piece of behaviour B_3. Lincoln's death (D) is caused by S and therefore by C. So B_3 is itself nested within a still more extended behaviour B_4. Each of the described behaviours, and any process with respect to some subject must be a process which has an internal state's causing something (a movement).

Behaviours may be components of more (spatio-temporally) extended behaviours. There may thus be no very definite answer to the question of when a given piece of behaviour took place. When did Booth assassinate Lincoln? The behaviour in question occurred over some period of time so there is no instant that one could identify as being the time at which Booth assassinated him. This view is one that Dretske thinks is a consequence of viewing behaviour as a process of one thing's causing another and it is a view which he thinks helpful in clearing up potential puzzles about action and time. B_1 is C's causing M. B_4 is C's causing D (via C's causing M which causes S which causes D). M's causing S is a process, and part of Booth's behaviour. It is not by itself a piece of Booth's behaviour, however, because M's causing S is not a process of an internal state's

(of Booth) causing a movement.

Although he is anxious to distinguish his approach from Donald Davidson's, Dretske's notion of behaviour still does assign a kind of special status to bodily movement. At the core of each behaviour will be an internal state's causing of a bodily movement. This is an agent's actual contribution, and it is the end of their contribution to any "extended" behaviours. One might call this core contribution "narrow behaviour".

Dretske construes behaviour processes so as to include more than overt bodily movements. Refraining from moving may also count. He writes:

> I have, for convenience of exposition, always spoken of behavior as involving movement of some kind. Although movement often occurs, it is clearly not necessary. Hatching eggs is a perfectly respectable form of bird behavior that doesn't require movement. The fact that a bird moves, and, given its physiology cannot avoid moving during this period is irrelevant. The point is that such movements are not logically *required* for the hatching of eggs. (Dretske 1988a, p.28)

What seems to be intended here is that, while a bird may shift position or engage in other movements while it is hatching eggs, these movements are not part of what is meant by the behaviour "hatching an egg". In the same way, if I scratch my ear while watching a turtle, ear-scratching does not thereby become part of the turtle-watching behaviour.

The refinement made here may well broaden what is to count as behaviour beyond what one would ordinarily think of as behaviour. Blushing apparently counts for Dretske. Referring to his use of 'M', he writes,

> It might be a change in color, temperature, or pressure, changes which may require movements at the molecular level but not what we ordinarily think of as movements of the animal. M can even stand for the *absence* of movement. *Non*-movement also has its causes, and when the cause is internal, the person is standing still, holding her breath, waiting, resting, pointing, aiming, watching, hiding, or whatever. If, on the other hand, the primary cause of non-movement is external, then no behavior is

> occurring. If Blackie is buried in concrete, his arms and legs are prevented from moving by external, not internal, causes. (Dretske 1988a, p.29)

Dretske acknowledges that there is a measure of vagueness in his distinction between behaviour and non-behaviour. That vagueness is exemplified in the foregoing passage by the expression "primary cause". If the distinction between a "doing and a "happening" is made in terms of whether the primary cause of some event is internal and external, then the distinction is only clear to the extent that the notion of a primary cause is clear. Dretske does not think that this notion is a clear one. He endeavours to sidestep the trouble that this might cause by stating that for his purposes what does or does not count as behaviour is not so important as what one is saying about the structure of some phenomenon in identifying it as an instance of behaviour:

> I have no philosophical interest in playing umpire in these disputes, no interest in trying to decide specific questions about what is and isn't behavior. My interest centers on what it is that one is identifying something as when, and if, one identifies it as behavior. It may be arbitrary whether something should be classified as behavior or not, but not at all arbitrary that, once so classified, it is a causal process of the sort described in the preceding section. The project is to understand how behavior is to be explained, and, specifically,how it is, or how it *can* be, explained by reasons. (Dretske 1988a, p.25)

The "preceding section" referred to in this quote describes behaviour as a process of an internal state's causing a movement. This particular quote is also notable in its reminder to the audience of what Dretske's chief purpose is: it is precisely an attempt to show how behaviour can be explained by reasons.

The move that Dretske makes in the passage is a little fast. Certainly his project does not require that he give an account which settles all disputes over whether particular examples are to count as behaviour or not. By adopting criteria, even for purposes of being clear about what is being said of a phenomenon that one proposes as an example of behaviour, one is still ruling on

what cannot count. For any arbitrary event related to some organism, that event cannot be construed either as behaviour or as a component of behaviour if it cannot plausibly be made to fit the proposed criteria. The cracking of my fingernail is not a candidate unless it is *plausible* to identify this occurrence as a process with an internal cause. Perhaps the point that Dretske means to make is simply that, since he is endeavouring to show how reasons can cause behaviour, he need only concern himself with what is clearly behaviour and with the properties that makes those phenomena behaviour.

Galen Strawson has provided an interesting discussion of the concept of behaviour and some of the difficulties in its classification[2]. Strawson lists a set of criteria "that narrows the set of possible instances of behavior in a familiar way" (Strawson 1994, p. 292). It should be noted that he does not agree with the all of the criteria that he lists. Strawson writes,

> **Definition** To be an instance of behavior on the part of a being B is to be a motion or change, or lack of motion or change, M that is a motion or change in or of B and that fulfils the following conditions:
>
> [O] M is *ordinarily other-observable*.
>
> [D] It is correct or reasonable to describe M as something B *does*, rather than something that just *happens* to B. Or, it is correct or reasonable to describe M as a self-motion or self-change on the part of B.
>
> [I] The *proximal cause* of M is internal to B.
>
> [C] The *proximal cause* of M is some activity of the *central control system* of B. (Strawson 1994, p.292)

The italicized terms in this definition are terms that Strawson wants identified as "importantly vague". One of Strawson's main points is that plausible candidate conditions for behaviour by no means settle questions about what is or is not behaviour. Distinctions between doings and

happenings and identification of central control systems are two sources of difficulty.

Dretske's view of behaviour clearly commits him to both [D] and [I]. The other two conditions are much less clear. Dretske provides some evidence for thinking that he would reject [C] as necessary for behaviour. He expresses sympathy with classifications of "behavioral biologists, embryologists, endocrinologists, and pharmacologists":

> These behavioral scientists have no trouble classifying as behavior -- and by behavior I mean human and animal behavior, not merely the behavior of glands and organs -- such things as respiratory and cardiovascular activity (Engle 1986), penile erections (often said to be part of an animal's "display behavior"), the secretions of endocrine and exocrine glands, muscle spasms, convulsions, seizures, involuntary eye movements, the regulatory activities of the autonomic nervous system, and all sorts of reactions, *including reflexes*, whose internal production remains well below the level of conscious or voluntary control. (Dretske 1988a, p.7)

Depending on how narrowly or broadly one construes the notion of a "central control system", some of these activities may well fall outside the governance of such a system. Certainly it would be too narrow to think of the central control system as the controller of voluntary behaviour, but does the central control system extend beyond the scope of an organism's major perceptual systems? Cardiovascular activity is not under proximal control of any perceptual system (although what I perceive may have effects on cardiovascular activity among its effects). Dretske is quite vague on what kinds of internal causes are necessary for behaviour or if any internal cause as primary cause will do. As such, it is not clear whether he would accept Strawson's [C] as a general criterion or not. On the other hand, in discussion of systems such as humans, Dretske is most naturally understood as thinking of brain states[3] as the internal causes of movements. Brain states would certainly seem to qualify as states of the central control system.

The most interesting aspect of Strawson's discussion centers on condition [O]. Strawson

himself argues against this condition, claiming that it is an epistemological criterion which has no place in the metaphysics of behaviour. Its exclusion allows Strawson to include as behaviour mental calculation and other forms of deliberate thought. These are activities which satisfy [D], [I], and [C] if any activity does. Whether or not Dretske would be willing to embrace mental activity as forms of behaviour is also unclear. None of the examples he uses indicate that he would, although watching could certainly be construed as having an important mental component. In a recent article, Dretske writes as follows:

> If behaviour is understood, as I think it must be, as a process that typically culminates in overt movement of some kind (e.g. raising your arm is a causal process that has arm movement as its product), then causal explanations of why a person moved her arm can be quite different even when the process (=behaviour) is the same. (Dretske 1996a, p.84)

This is a very slim piece of textual evidence. It is certainly possible that Dretske is simply being a little careless in his formulation, as the construal of behaviour is far from the main purpose of this particular article. This nevertheless suggests a commitment to criterion [O] -- a criterion which Strawson himself rejects -- and this would (ordinarily) exclude mental activity as behaviour. This is unfortunate, for Dretske's speculations on how his theory extends to more complex mental representation from the simple cases he discusses might actually be helped by Strawson's proposal[4].

Finally, Dretske's notion of behaviour as the process of a C's causing M comes out of the general observation that behaviour should be that which a subject does. This observation technically does not support the conclusion that behaviour must be identified with such processes, though. One could insist that behaviour is just bodily movement, for Dretske's claim actually consists of two parts: 1) behaviour is distinguished from that which merely happens to a subject by having an internal cause C, and 2) behaviour is itself the process of C's causing M. The

examples used previously, of my arm's being moved as opposed to me moving my arm, show 1), but they do not compel us to accept 2). One could instead identify behaviour with bodily movements whose immediate cause is an internal physiological state. That is, one could use the previous examples to conclude, not that behaviour is C-causing-M processes, but that it is M's which have certain kinds of immediate causes.

Making the extra step and identifying behaviour as a C-causing-M process (hereafter C-M process) is vital to Dretske's theory, however. As other critics have pointed out[5], Dretske defines his task in such a way that an explanation by appeal to reasons (i.e. intentional states) must not simply collapse into some lower-level physical explanation. He needs to give a unique and genuine causal (hence explanatory) role to the semantic component of intentional states in explanations of behaviour. Conceiving behaviour as C-M processes makes room for such a role for intentional states. Thus, Dretske writes,

> This way of conflating behavior and output does no great harm when one is thinking about amplifiers and their behavior; an engineer is typically interested only in an amplifier's output and in those conditions in the amplifier that are responsible for this output. But if one uses this model to think about human and animal behavior, great harm can be done. One can easily be misled into thinking that the cause of behavior is necessarily the cause of output. And once this confusion is in place, one will have no option but to identify causal explanations of why we *do* the things we do with causal explanations of why our body *moves* the way it does. One will, in other words, have succeeded in confusing psychological explanations of behavior with neurobiological explanations of motor activity. Reasons -- our *thinking* this and *wanting* that -- will have been robbed of an explanatory job to do. And with no explanatory job to do, reasons -- and by this I mean the beliefs, desires, intentions, and purposes that common sense recognizes as reasons -- will have been robbed of any scientifically reputable basis for existing. (Dretske 1988a, p. 36)

The position Dretske adopts on behaviour is thus clearly an integral part of his theory. Unless behaviour is a C-M process, there will be no causal role and hence no ineliminable

explanatory role for the semantic components of intentional states. Only if there is some ineliminable explanatory role for such states (specifically for the meanings or contents of such states) are these "scientifically respectable", and only if they are scientifically respectable is there reason to think intentional states exist.

To summarize, I have suggested that there may be some grounds for challenging Dretske's conception of behaviour. The terms in which he tries to delineate behaviour and non-behaviour are rather vague. It is also not entirely clear that Dretske's reasoning forces one to accept the idea of behaviour as a process. In spite of this, I shall in what follows grant Dretske's general conception.

Dretske's conception of behaviour is that behaviour for some subject S is a process of an internal state of S's causing a bodily activity of S. The bodily activity is usually but not always a bodily movement. Dretske does generally write as though behaviour always does involve bodily movement, but it is clear when he is being careful that this is *only* for expository reasons. In actual practice, the internal state C which is the cause of bodily movement M will be one among a number of causal contributors. For C's causing M to be a behaviour, C must be the primary cause. It is safe to say that Dretske is committed to Strawson's conditions [D] (behaviour as a *doing*) and [I] (behaviour as having an internal state as *proximal* cause). It is less certain that Dretske is committed to [O] (other-observability) or [C] (the internal proximal cause is a state of the central control system). There is some evidence to suggest that Dretske does accept [O]. It is still less certain that he accepts [C].

There are two absolutely crucial points about behaviour for Dretske:

1. Behaviour is a process (a C's causing M) and must not be confused with the product (M)

of the process.

2. What is explained by an appeal to reasons (i.e. beliefs and other intentional states) is the behaviour and not the movement.

b. Causation

Another important distinction which Dretske draws on in setting up his theory is between triggering and structuring causes. The main significance of this distinction is that Dretske identifies reasons as a species of structuring cause. This is an important and controversial point. It assigns reasons a different role in the explanation of behaviour than common sense would assign. As I shall later show, it also assigns the explanatory power to the wrong properties[6].

A triggering cause is an efficient cause of an event. It is a more or less contemporaneous event which sets a particular instance of an event or process in motion. The flipping of a switch triggers the illumination of a light. My seeing a looming dark shadow triggers my ducking. One kind of request for an explanation is answered by an appeal to a triggering cause. Why did the lights go on just now? The motion detector was activated. Why did Brent duck just now? He saw something moving quickly toward his head.

A structuring cause is a kind of second-order cause. It is not the same as the notion of background conditions. Oxygen's being present is a background condition required for the striking of a match to cause its ignition, but this is not what Dretske means by structuring. For an event (or event type) to be a structuring cause, it must cause a (causal) process. An electrician's connection of a motion detector to a circuit connected to an outdoor floodlight is an example of a structuring cause. The electrician has structured a system so that the circuit is completed and the floodlight goes on when the motion detector is activated. The result of the electrician's work is a process: the

motion detector's activation causing the illumination of a floodlight. Similarly, Dretske would say, my seeing a dark shadow looming causing me to duck is a process that has been structured. This may have been structured by hard experience or it may be a process that I have as a result of my genetic programming[7]. In the former case some learning process structures the process of my seeing a dark shadow causing a certain movement. In the latter case, the processes of natural selection on my progenitors structures this process of a visual state causing a movement.

In Dretske's own words,

> In looking for the cause of a process, we are sometimes looking for the triggering event: what caused the C *which* caused the M. At other times we are looking for the event or events that *shaped* or *structured* the process: what caused C to cause M rather than something else. The first type of cause, the triggering cause, causes the process to occur *now*. The second type of cause, the structuring cause, is responsible for its being *this process*, one having M as its product, that occurs now. This difference, a difference I have elsewhere (1972) described in contrastive terms, is familiar enough in explanatory contexts. There is a clear difference between explaining why, on the one hand, Clyde stood up *then* and explaining, on the other hand, why what he did then was stand up (why he *stood up* then). He stood up *then* because that was when the queen entered, or when he saw the queen enter the room. He *stood up* then as a gesture of respect. The difference between citing the triggering cause of a process (the cause of the C which causes M) and what I have been called its structuring cause (the cause of C's causing M) reflects this difference. (Dretske 1988a, pp. 42-43)

It is interesting but not surprising that the distinction drawn in the foregoing passage has as much to do with explanation as it does with causation. It is Dretske's view that to explain some event is to cite a cause (or causes) of that event. The distinction between "triggering" and "structuring" in fact is probably more easily made out in terms of explanation rather than causation.

Appeals to triggering causes are both immediate and particular in a way that an appeal to a structuring cause is not (or need not be). Explaining why Clyde stood up *now*, on Dretske's scheme something amenable to a triggering cause appeal, is a matter of pointing to some particular event

or events, usually ones which are relatively contemporaneous with the instance of Clyde's standing. Such events are particular occurrences (the queen's entering the room, Clyde's seeing her) which lead to a particular occurrence.

A structuring cause is a cause of a process: a queen sighting causing a Clyde-standing. One can know that Clyde stood up or even that Clyde stood up because the queen entered the room without exhausting explanatory questions. The outstanding question is the question of why the queen's entering prompts Clyde to stand up (rather than, say, clapping or whistling or doing nothing). The answer to this question is a background story, presumably along the lines of Clyde's having learned how to show respect to dignitaries, wanting to show respect when appropriate, and learning that a queen is a dignitary. Similarly, one may ask for the structuring cause when the flipping of a switch causes the light to go on. This is to ask why it is that a switch-flipping caused a light-illumination (as opposed to causing something else). An answer to a request for a structuring cause in this case will be a story about the electrician's connecting of wiring from a power source, through the switch, to a light. Possibly statements about an agent's desires, intentions, or beliefs will figure in an explanation of a switch-flipping's causing a light's illuminating.

What is revealed by appeal to structuring cause is the reason why a causal mechanism has come about. The sorts of mechanisms in which Dretske is primarily interested are those which connect internal physiological states of organisms to their bodily behaviour. An internal state C_1 causing bodily activity of M_1 is a (type[8] of) process. The existence of the mechanism such that instantiations of C_1's type have M_1 (rather than some other movement M_2 or M_3) as their effect is what structuring causes explain. As Dretske uses them in his theory, structuring causes explain the

establishment of a type of process.

A structuring cause may thus be spatially and temporally distant from an instance of a process explained by referring to it. Why it is that standing is what Clyde did today when the queen entered may be "structurally explained" by etiquette instruction Clyde received in 1968. Why it is that a light's being illuminated is what happened when I turned on the hallway switch may be similarly explained by the work of an electrician long since retired. Once the mechanism is in place so that appearances of the queen will cause Clyde to stand, or so that switchings will cause lightings, the work of the structuring cause is complete[9].

Dretske argues that 1) one sort of perfectly legitimate type of explanation could be in terms of a structuring cause of some process, and 2) explanations of behaviour in terms of reasons are these kinds of explanations: "When we seek an explanation of behaviour in terms of the agent's reasons, we are, I submit, always looking for a structuring cause." (Dretske 1988a, p. 50)

Regarding 1), as previously noted, one kind of explanation is an answer to the question, why did x occur? (Eg. why did a light come on at a particular time?) This may be answered reasonably by an appeal to a triggering cause. When one is requesting an explanation for the occurrence of an event a at particular time *t*, what Dretske says is wanted is identification of that which triggered the event so that it occurred at *t*. Why did Brent run just now? He saw Edgar the neighbourhood thug. Why did the light come on (at midnight)? The automatic timer through which the light connected was set for midnight. These are the kinds of examples which Dretske identifies as appeals to triggering causes, events whose occurrence causes some event (Brent's running, the light illuminating).

In connection with behaviour, a triggering cause might be an instance of a

perceptually-caused neurophysiological state (**n-state**). The occurrence of the n-state triggers certain bodily movements. Why is Brent in the kitchen making a sandwich? He smelled food. In this context, Dretske wants to link explanation by triggering cause with explanation of some output or performance. Such an appeal does not explain *why* the smelling of food causes sandwich-making or why an Edgar-sighting causes Brent's running. It only explains that the current bout of sandwich-making was set off by an instance of food-smelling, but not why that type of cause has the effect it has. A "structuring cause" explanation would explain what is unexplained by an appeal to a triggering: why it is that the perceptual state (smelling food) causes that particular output (sandwich-making).

An answer to the question of why a floodlight, rather than an alarm is activated by a motion detector appeals to the electrician's connecting the motion detector to a circuit which has a floodlight (and not an alarm) at the end of it. One may also appeal to the intentions of the various agents involved in the building of this structure. The electrician wants a light to be activated by a motion detector because Ms. Ramsbottom (the home-owner) wants this and she wants this because she hopes that sudden illumination will both scare away intruders and provide light for unexpected guests. An alarm might also do the first job (scaring intruders), but it would not do the second.

This example hints at another feature of structuring causes which Dretske takes to be important: two numerically different examples of the same mechanism may have different structuring causes. Let us alter the example and suppose that Ms. Ramsbottom has a light connected to a motion detector only as a convenience for unexpected guests. Her neighbour, Mr. Larch, who does not want to be disturbed by an alarm, has installed the same mechanism to scare

intruders. The conditions under which the light will be triggered are identical but the (structuring) explanation of why it is illumination that is caused by Larch's motion detector is different from the (structuring) explanation for Ramsbottom's mechanism.

Regarding 2), Dretske contends that appeals to reasons in explaining behaviour are appeals to structuring causes. A reason, once again, is an intentional state for Dretske. a belief, desire, hope, expectation, suspicion, and so on. Discussion generally focuses on beliefs as standard examples of doxastic states and desires as standard examples of motivational ones. So Dretske apparently contends that when one gives an explanation of a cognitive being's behaviour by appeal to beliefs and desires (where what the belief is a belief about or what the desire is a desire for is relevant), one is appealing to a structuring cause.

This is an important point because it undergoes some modification even within *Explaining Behavior*. Dretske's view assigns to internal, movement-causing states two importantly different semantic properties: representation and indication (these will be detailed in the next section). According to Dretske's theory, for C to be explanatory as a reason, it should be that the fact that C is a representation of something that is explanatorily relevant. It is a belief-state's status as a representation that makes it a reason and allows it to enter into psychological explanations. However, the view that Dretske appears to take in more careful moments and, as we shall see in my critical discussion, the view he seems forced to take is that the fact that C-tokens are indicators is what (if anything) structurally explains a behaviour.

A consequence of making explanation by reasons explanation by structural cause is that intentional explanation is an explanation of the existence of a *type* of process. If an appeal to intentional states is an appeal to a structuring cause of some type of (behaviour) process, then what

the appeal explains is the existence of the type of process. So intentional explanation is an appeal to the past, to what it is about past occurrences of some n-state in virtue of which states of that type now tend to trigger a physical movement of a certain type. This move is startling and certainly not uncontroversial. One should think that if Jones goes to the refrigerator to get a can of orange juice that the role played by wants and beliefs would be triggering. After all, it is Jones' now wanting juice together with his now thinking that there is some in the refrigerator which causes his going to the refrigerator and removing a can of juice. Yet, this is a picture which Dretske's theory seems to deny, for he has distinguished between two sorts of explanations: 1) the explanation in terms of relatively proximal causes of what brought about a particular event now, and 2) the explanation in terms of relatively distal causes of why there should be a type of process such that a type of cause C should have an effect M. According to Dretske, only the latter are properly psychological.

In summary, behaviour is a process, C's causing M. Instances of a process have two types of causes (at least): triggering and structuring. For a given instance of C's (now) causing a token of M, to explain that M occurred (now) either because a C-token occurred, or because condition F (which caused the C-token) occurred, would be an explanation in terms of a triggering cause. To explain that it is because of condition F's relation to tokens of type C that it is M that the C-token caused (rather than nothing or something other than M) is to explain in terms of a structuring cause.

This is an important distinction for Dretske, for he uses it to attempt to carve out a unique explanatory role for reasons. Physiological causes of movements are triggering causes. Reasons, which explain behaviour, are structuring causes. Structuring causes can be spatio-temporally distant from their effects. Structuring causes can also be different for two mechanisms even when

the operation of the mechanism is the same.

NOTES

1. To forestall an obvious complaint, Dretske does acknowledge that restrainings and failures to move can be behaviours. These can be accomodated with a liberal interpretation of "bodily movement". It seems to me that nothing very great is gained by objections about possible behaviours involving non-movement. Exercises such as remaining motionless or in a fixed position or restraint often require just as much muscular activity as does an overt movement.

2. Strawson 1994, pp. 291-316.

3. In fact Dretske refers to the internal states as brain states in his "Absent Qualia" (Dretske 1996a, p. 84).

4. As Dretske is mainly concerned to show how representation is possible within a naturalistic framework, he has much to say about perceptual beliefs and relatively straightforward connections to behaviour. In later chapters of *Explaining Behavior*, he speculates on higher level intentional states but it is generally unclear in his descriptions whether Dretske is entitled to the mechanisms to which he helps himself. Nothing at all that is said in his basic theory indicates how various "reasons" (intentional states) can be interactive *as intentional states*. His chapter six clearly assumes that they must be. The idea that some of the behaviour that is "structured" by "design solutions" (these are notions that will be explained later) is mental behaviour or brain behaviour might actually help to give a better idea of what intentional state interactions are going to look like on a Dretskean view.

5. Jaegwon Kim, for example, usefully distinguishes some problems related to causation/explanation and over-determination which Dretske's theory must overcome (Kim 1991, pp.53-57).

6. See Chapter 5, section c.

7. The ways in which a cognitive being's behaviour can be structured will be revisited in detail later.

8. The subscripts I use here are meant to indicate that the items denoted are tokens. My use of the lettering "C" and "M" is intended to follow Dretske's. "C" denotes a state tokens of which represent

(and/or indicate) distal conditions. It is in Dretske's use an internal state of a representational system (usually an organism). "M" stands for "movement", usually referring to a bodily movement of an organism. I will use these letters to denote state types.

9. This is a little misleading. Some processes or causal mechanisms for processes need maintenance. Processes established by operant conditioning figure centrally in Dretske's theory and it is not true that all conditioned behaviour, once conditioned, will persist indefinitely without further reinforcement.

CHAPTER TWO

Indication and Information

Dretske's notions of indication, information, and (natural and non-natural) meaning are crucial to an understanding of his theory of intentional states as representational states. The thesis Dretske is defending is that mental states (hence intentional mental states) are representational states. Representation is explicated in the theory in terms of indication, or information-carrying. Representational theories of mind hold that there is a meaning-component of intentional states, something in virtue of which they are "about" other states of affairs, and this component is a representational content[1]. In Dretske's theory, a mental representation is a function to indicate. The notion of indication is defined in terms of information: a token r indicates that s is F if and only if r carries the information that s is F. Indicating and functioning to indicate are different notions and it is important to keep the two distinct. Dretske sometimes uses the expression "natural meaning" to refer to indication and the expression "non-natural meaning" to refer to functioning to indicate. I will begin the explication with "information".

a. Information[2]

In *Knowledge and the Flow of Information*, Dretske's concept of information is set out in detail. Much of this detail is missing in later discussions employing the notion of indication. Both information and indication, as Dretske uses them, are objective, mind-independent relations (although he allows that they *can be* relative, as we shall see):

> I have elsewhere (1981, 1983), under the rubric *information* tried to say something more systematic about the idea of an objective, mind-independent, indicator relation... Sufficient unto present purposes is the assumption -- an altogether plausible assumption, I hope -- that there is something *in* nature (not merely in the minds that struggle to comprehend nature), some objective, observer-independent fact or set of

> facts, that forms the basis of one thing's meaning or indicating something about another. (Dretske 1988a, p. 58)

The early characterization of information (and thus the indication relation) is adapted from mathematical information theory. The focus of this branch of mathematics is on information quantity, which is roughly speaking measured in terms of reductions of possibilities, perhaps in some independently delineated field of possibilities. Dretske identifies three desired criteria for a signal r to bear information that something is F. Here are paraphrases of the first two (Dretske 1981, pp. 63-4):

(A) The quantity of information borne by r must equal the quantity of information resulting from s's being F.

(B) It must be the case that s is F.

These first two conditions are drawn out of Dretske's initial discussion of mathematical information. One of the basic distinctions made is between information at a source and information at a receiver. I will borrow the type of simple illustrative example to which Dretske appeals (Dretske 1981, pp. 4-5): Imagine that one employee of some company must be selected to perform a particular job. The boss picks eight employees and leaves them in a room to decide among themselves which one of the eight will do the job. When the employees decide who the individual is, information is generated. The decision procedure resulting in Jones' being selected generates information at the source. However, until the employees send word out, however they do this, so that the boss (for example) receives information about the employees' decision, no information has been transmitted. Once a slip of paper containing the name "Jones" (for example) has been sent from the room and is now in the hands of the boss, information has been transmitted. There is

information at a receiver. Condition (A) is a reference to this transmission of information. If Jones' being selected generates some specific quantity q of information, then the quantity of information on the slip of paper, in order to transmit the information that Jones was selected, must be equal to q^3. If it contains less information, then it is not possible for the exact outcome of the selection procedure to be conveyed by the note. Imagine a situation in which the name written on the note bears less than q information: The employees simply write the letter "J" on the note, neglecting to tell the boss whether they intend to indicate their choice by first or last name. As it turns out, two of the eight employees (Betty Jones and Jerry Smith) have a name beginning with a "J". Receiving a note with the letter "J" on it in this case does not carry enough information to indicate that Jones has been selected. It only carries enough information to indicate that either Smith or Jones has been selected.

Condition (B) is self-explanatory but it is very important in the development of Dretske's notion of indication. It ties information to truth. A signal r cannot carry information about what is not the case.

Dretske notes that criteria (A) and (B) cannot be sufficient. There can be a range of equally probable outcomes of some situation, each one of which would thus generate the same quantity of information without it being true that some signal indifferently carries the information that each of them occurred. A runner may pass on a slip of paper to a supervisor containing a number which corresponds to the number which comes up from a toss of one die. If (A) and (B) were sufficient, any outcome coincident with the toss of the die would be carried by the runner's note. Suppose there are six bettors, each of which must bet on a different number. The number five comes up and Brown (who bet on one of the six numbers) bet on number five. If the criteria stated so far were

sufficient, then the note containing the number five must also carry the information that Brown's bet was a winner. The supervisor clearly can be informed as to the winning number without finding out that Brown or any other particular individual won the game. Thus Dretske adds a third criterion (Dretske 1981, p.64). Again in paraphrase,

(C) r's bearing the information it does must be dependent specifically on s's being F (even if s is G and s's being G carries the same amount of information as s's being F).

To relate this to the previous example, a written numeral "5" on the runner's paper is presumably dependent upon the face of the die with five dots having landed up. It is not so dependent upon Brown's bet being the winning bet. Supposing that the numeral scrawled on the paper is meant to correspond to the number of dots on the face of the die which is up after a toss, the numeral five would have appeared even if Brown had bet on number two. If the "two" face had landed up but Brown had bet on five, conversely, the numeral "5" would not have appeared on the runner's paper.

Now the runner *could* be conveying information about whose bet won, rather than information about which number turned up. Without a change in the amount of information or even in the physical means of conveyance (use of numerals on paper), the runner could number the bettors from one to six so that the appearance of "5" on the paper depended specifically on Brown's winning, regardless of the number of the die upon which he bet. Of course the runner could also convey information about both, although this would involve a change in the amount of information carried by the note.

The result is the following formulation of r's carrying the information that s is F:

Informational content: A signal r carries the information that s is F = The conditional

> probability of s's being F, given r (and k), is 1 (but, given k alone, less than 1). (Dretske 1981, p.65)

The letter k is a reference to background knowledge, and Dretske's inclusion of it, he says, is to acknowledge that the information two receivers obtain from the same signal is partially determined by prior knowledge[4]. Its inclusion at this point in Dretske's account of information is a little odd. Communication theory itself is not specifically connected to transmission of information to receivers that have cognitive states. Perhaps the prominence of background knowledge in *Knowledge and the Flow of Information* is due to Dretske's epistemological purposes. In subsequent discussions, especially, where he is giving an account of mental representation, there is little mention of the role of background knowledge. This is no surprise. Obviously a *ground-level* account of the intentional content of mental states could not rest importantly on states of knowledge without being circular. I do not think Dretske means knowledge states to be essential to information (hence to indication, hence, ultimately, to the formation of mental states with intentional content). In *Explaining Behavior*, Dretske again notes that what is indicated to a given subject is relative to what he or she knows about other possibilities. He then goes on to say (in a footnote),

> This, however, doesn't mean that natural meaning is *subjective*. A person's weight isn't subjective just because it is relative, just because people weigh less on the moon than they do on the earth. If nobody knew anything, things would still indicate other things. They just wouldn't indicate the specific sort of thing... we now describe them as indicating. (Dretske 1988a, p. 58)

Dretske's comments concerning indication translate more or less directly into claims about information. So r's carrying information, according to what is said in the previous quote, does not require a receiver with background information. While in some cases background knowledge is

required for certain information to be carried (indication), background knowledge is not required for every case of information's being carried.

It is clear from Dretske's initial account of information and his subsequent account of indication that what is of fundamental importance is that s's being F cannot be irrelevant to the occurrence of a state r of a receiver. This is the point of criterion (C) (there must be a dependence between r and s's being F). What is meant to be excluded is information being carried by vacuous signals. The conditional probability of s's being F given that r occurs will be 1 in cases where the simple probability of s's being F is 1. In such a case, one would not want to say, when r occurs, that r carries the information that s is F. For example, the conditional probability of the sun's being hot given that my watch shows five o'clock is 1, but one does not want an account of information on which my watch carries the information that the sun is hot. Why not? The conditional probability of the sun's being hot given that my watch does not show five o'clock is also 1. The requisite dependency between the state of my watch and the sun's being hot is absent.

There are ways of defining informational content along Dretske's lines so as to satisfy this desideratum without specific reference to background knowledge. For example, one might try either of the following:

IC1: A signal r carries the information that s is F = the conditional probability of s's being F, given r, is 1; and the conditional probability of s's being F, given that r does not occur, is less than 1.

IC2: A signal r carries the information that s is F = the conditional probability of s's being F, given r and background conditions c, is 1 but, given c alone, is less than 1.

IC2 is merely a slight generalization on the original definition that does not specifically

mention knowledge. In communication systems in which the receiver is a cognitive being it is still perfectly possible to note that the relevant background conditions would be or would include the receiver's background knowledge. Where the receiver may not be a system with knowledge states (a computer for example), the definition still applies. The relevant background conditions might be programming or available data.

On any of these three definitions, it has to be the case that there is a difference in the conditional probabilities of s's being F given r and s's being F given that r does not occur. According to Dretske's criterion (C), this is what is wanted: a dependence between a received signal and some source-condition. Were the aforementioned conditional probabilities the same, of course, s's being F and r would be independent events. So IC1 is a completely general way of defining informational content which captures dependence without making any claims about the nature of the dependence. Its advantage over IC2 is that it avoids potential worries about what constitutes a background condition and what kind of distinction is being drawn between a background condition and a signal. The advantage both IC1 and IC2 have over the original is in not mentioning any kind of cognitive state. The account of information eventually is used to supply an account of intentional states and their contents. These include (primarily) beliefs. The normal construal of knowledge is as a species of belief. Thus in employing Dretske's original definition, one runs the risk of appealing to intentional states in order to explain intentional states in information terms.

Finally, in an article critical of Dretske's reliance on indication, Carol Slater interprets Dretske's definition of information along the lines of IC1:

> Although Dretskean indication... requires that internal state C never occur in the absence of the relevant environmental state of a affairs -- p(F/C) must equal 1 -- the probability of F *in the absence of* C -- p(F/~C) -- can be anything at all so long as it

falls short of 1. (Slater 1994, p. 177)

An attached footnote then refers the reader to the passage, in which Dretske defines information, that I have already cited (Dretske 1981, p. 65) as authorizing her view[5]. In his response to Slater's article, Dretske does not take any exception to Slater's characterization of his notion of indication. I will therefore take IC1 to be a reasonable general construal of Dretske's notion of information.

Two further important features of information need to be mentioned here. First, the relation as it has been discussed so far is a relation between particulars. A given state r of a receiver carries the information about a given condition s's being F. As the related notion of indication is developed into a theory of representation, it is sometimes unclear, both in Dretske and in some of his commentators whether "indicates" is to be treated as a relation between particulars or between types. I will treat this issue in detail in the following section.

Second, the dependency which exists between a signal r and a state of affairs, s's being F, about which r carries information is a "one-way" dependency. The probability criterion spelled out by Dretske does not suggest anything about the conditional probability of a signal r given that s is F. When this relation is fit into a theory about dependencies between internal states of cognitive beings and external conditions, the required dependencies will be from internal state to external condition. As the relation has been so far elaborated, there can be no occurrences of a signal r in the absence of s's being F. There is no corresponding converse claim, however. The theory makes no claim at all about whether (or how often) the condition of s's being F may be present without an occurrence of the signal r. There is no requirement in the theory, for example, that the conditional probability of some internal state of mine occurring given that some external condition

is present be 1 (or any other value). Of course it cannot be 0, but there is no theoretical reason, on Dretske's view, why it could not be any other possible probability value.

One advantage of an information-based theory is its generality. Any theory which seeks to account for the semantic component of intentional states by means of some particular kind of dependency between a state of affairs and the intentional state is compatible with an informational framework. This is because, while information relies on some dependency, the form of dependency (causal, nomological, or logical) is not specified.

To summarize, on Dretske's concept of information:

a. a given signal r cannot carry the information that s is F unless s is F,

b. r's occurrence must be dependent on s's being F such that the conditional probability of s's being F given r is 1, and the conditional probability of s's being F given that r does not occur is less than 1; the dependence could be expressed still more generally by saying that the conditional probability of s's being F must be marginally higher than the conditional probability of s's being F given that r does not occur[6],

c. there need be no corresponding converse dependency of s's being F on r.

With this understanding of Dretske's use of "information" in hand, I will now go on to detail Dretskean indication.

b. Indication

In discussions of information in relation to philosophy of mind, such as is found in *Explaining Behavior*, Dretske has adopted "indication" as his primary technical term. Thus there are in Dretske's writings three different terms used to express the same basic concept: indication, natural meaning, and carrying of information. Dretske writes:

> In what follows I shall occasionally, partly as terminological convenience but also partly to exhibit the deep connections between representational systems and information-processing models of human cognitive, advert to the idea of information. Talking about information is yet a third way of talking about the fundamentally important relation of indication or natural meaning. So, for example, if S (sign, signal), by being a, indicates or means that O is A, then S (or, more precisely, S's being a) carries the information that O is A. What an event or condition (whether we think of it as a signal or not is irrelevant) indicates or means about another situation is the information it carries about that other situation. (Dretske 1988a, pp. 58-59)

Dretske may speak of a condition r indicating P, r carrying the information that P, or r meaning that P. In the latter case, it is important to keep clear about what Dretske specifically means when he uses expressions like "r means that P" in discussion of *natural* meaning. Indication is not the same as representation but there is a sense of "means" (i.e. "*non*-naturally means") which is the same as Dretske's use of "represents".

There are two important issues in bridging the gap between Dretske's early discussion of information and his use of indication as the basis of (mental) representation. First, the initial discussion treats information-carrying primarily[7] as a relation between a signal (token) and some distal (token) state of affairs. In later works it is not always clear whether indication is a relation between types (r-type signals as indicators of condition type s's being F). The role that indicators play in Dretske's theory in grounding representation make it convenient to talk about indication as a "type" relation. I will develop both a "type" and a "token" version of "indication". Second, there is the issue of the extent to which Dretske's early views on information survive intact in more recent writings. It is my view that Dretske's view changes very little.

In *Explaining Behavior* details of Dretske's account of information are omitted. Once the mathematical details are left aside, criterion (A) comes to play no role in the discussion. The importance shifts to criteria (B) and (C) (see above pp. 31-34). These are the two criteria in any

case which Dretske originally identifies as constituting "the semantic conditions on information" (Dretske 1981, p. 64). Because the detailed account of information is omitted from later writings, it is open to some speculation whether or not Dretske's notion of indication maintains the very strong information-carrying conditions elaborated in *Knowledge and the Flow of Information*. It is my view that Dretske holds substantially to his original conception.

It will be useful first to look at what Dretske says about "indication". On Dretske's conception, indication/natural meaning is objective and not subject to error. Indication is objective in the sense that r's indicating P (= s's being F) is not subject to agent's interpretation or understanding of r as indicating what it does. Second, just as r cannot carry the information that P unless P obtains, r cannot indicate a condition or state of affairs P unless P obtains. If r cannot indicate P if it is not the case that P, then r cannot *mistakenly* indicate P. Misindication is a conceptual impossibility. Dretske writes:

> A person can *say* and *mean*, that a quail was here without a quail's having been here. But the tracks in the snow cannot mean (in the natural sense of "meaning") that a quail *was* here unless, in fact, a quail was here. If the tracks were left by a pheasant, then the tracks might, depending on how distinctive they are, mean that a pheasant was here. But they certainly do not mean that a quail was here, and the fact that a Boy Scout *takes* them to mean that cannot *make* them mean that. (Dretske 1988a, p.56)

(Other things being equal) a set of bird tracks in the snow caused by a quail indicate a that a quail was here. They would do so even if quail were so elusive that no-one had ever encountered one (and so no-one knew what in fact had caused the tracks). Indication or natural meaning is a function of facts of the matter, not interpretations[8]. One might also notice here that the foregoing quote seems to refer to a *particular* set of tracks.

Dretske continues:

> Furthermore, even if P does obtain, the indicator or sign does not mean (indicate) that P is the case unless the requisite *dependency* exists between the sign and P. Even if the tracks in the snow *were* left by a quail, the tracks may not mean or indicate that this is so. If pheasants, also in the woods, leave the very same kind of tracks, then the tracks, though made by a quail, do not indicate that it was a quail that made them. A picture of a person, taken from the back at a great distance does not indicate *who* the picture is a picture of if other people look the same from that angle and distance. (Dretske 1988a, p.56)

Clearly, a signal r and a distal event P cannot be independent events. The signal must bear some dependency relation to P. In the quail example, the dependency happens to be causal. The set of tracks was caused by the passing of a quail. However, nowhere does Dretske suggest that his view is strictly a causal view. This is a respect in which his theory of representation ultimately differs from earlier informational approaches such as Dennis Stampe's[9], for example.

There are two important components, then, to the notion of information-carrying and its successor "indication". One component is about tokens. The state r cannot indicate that P unless P obtains. A given set of tracks cannot be quail-tracks unless it was a quail that caused them. The other component is about types. The signal r must be of a type such that it would not have occurred independently of the occurrence of a condition of P's type. A set of tracks (whether quail-caused or not) is not of a type that indicates quail unless tracks of that type would not have occurred if no quail had passed by.

(i) Indication: Type or Token?

Lynne Rudder Baker actually gives two separate definitions of indication, one concerning types and one concerning tokens. Baker uses the lettering convention which Dretske employs in *Explaining Behavior*. "C" stands for a type of internal (indicating) state of a representational

system, "F" for a type of environmental condition external to that system[10].

Baker first defines type-indication as follows:

> A state of *type C indicates* an external condition *F* iff there is a dependency of tokens of *C* on occurrences of *F* such that, in normal conditions, all and only occurrences of *F* cause tokens of *C*. (Baker 1991, p.104)

Next is the token relation:

> A token *t indicates* an external condition *F* iff (i) *t* is of a type that indicates *F* and (ii) the conditions in which *t* occurs are normal.(Baker 1991, p.104)

The definitions given by Baker will not do as they stand. First, Dretske clearly does not define the dependency required for indication in terms of causation. A causal relation is one form that the dependency may take, but it is not the only form[11]. Of course in the majority of examples, perhaps in all standard examples, the dependency relation will be based on causation. Baker's oversight is not an especially serious one.

Second, it can be confusing to think of Dretske having two distinct notions of indication, one token and one type. His own discussion perhaps encourages talk of (type) C indicating a type of condition F. The original Dretskean concept of indication is founded on the idea of particular signals carrying information. In his article "Misrepresentation", Dretske comments that "[i]n speaking of signs and their natural meanings I should always be understood as referring to *particular* events: *this* track, *those* clouds, and *that* smoke" (Dretske 1986, p.20). I am therefore inclined to conceive of indication as primarily a relation between particulars C and F. On the other hand, it is true that the relation requires a dependency between a C-token and an F-token, and this involves a relation (causation, for example) which concerns type C and type F. We will also see a little later that the step from indication to C-tokens being of a type that represents something as

F requires that tokens of C's type generally are indicators[12]. Neither do these points force one to identify specifically a relation of type-indication nor do they suggest that defining "token-indication" in terms of type-indication is a good idea. It will be useful in subsequent discussion to have a notion of an indication relation that relates types, but also useful to have this identified in a way that makes it easily distinguishable from the basic "token" indication relation. In the following discussion I will illustrate how a problem over "type" versus "token" indication can arise and I will address the problem.

Third, still another difficulty with Baker's "type" notion of indication is that there is an addition to the dependency requirement which is not present in Dretske's theory. This point will require a short digression. As I noted in my previous section on information, Dretske's account makes no claim about the conditional probability of an occurrence of a C-state given an occurrence of an external condition F. Baker's definition of the C-F dependence (replacing the notion of "cause" with the more general "dependent") is that:

(a) all tokens of (type) C are dependent on tokens of (type) F, and

(b) only tokens of (type) C are dependent on tokens of (type) F.

Dretske's view is (a) only. One way of putting this is to say that Dretske is committed to an internal state tokens of (type) C being reliably, and non-accidentally, correlated with F. In fact, the initial Dretskean view has it that C must be perfectly correlated with F (over some period of a system's history). However, there is in general no corresponding commitment to tokens of external condition (type) F being perfectly (or even reliably correlated with) C.

Peter Godfrey-Smith also interprets Dretske's theory as I have, referring to it as an "asymmetrical theory" because the dependence is such that internal states must bear a reliability

relation with a type external distal condition only. One can put this asymmetry by saying that the direction of reliability of the correlation between C and F is C's being reliably correlated with F. Godfrey-Smith writes:

> It is suprising that there is still widespread uncertainty in the literature on the question of the direction(s) of dependency involved in indication. Dretske's original account was unambiguous. If the representing system would not be in the state it is in without the world being a certain way, then that state of the system indicates that condition in the world. He did not require that the dependency hold in the converse direction; it is not necessary that the state of the world indicated always produce an indicator in the representing system. (Godfrey-Smith 1992, p. 286)

There are a couple of additional useful ways of thinking about directions of dependency. The first recognizes the fact that the sort of internal states will in the most interesting cases be neurophysiological states of beings such as ourselves. Hartry Field (as Godfrey-Smith notes) refers to Dretske's clearest reliability commitment as "head-world reliability". The converse direction of dependency is "world-head reliability" (Field 1990, p.106). Second, one can think about this in terms of types of errors. A "false positive", for some token of state C (that is a prospective indicator of F) would be an occurrence of a token of C in the absence of a token of F. On Dretske's early account of information at least, there can be no "false positives". A C-token does not carrying the information that something is F if that token is not dependent on an occurrence of a token of F. A "false negative" is an occurrence of a condition F in the absence of a token of C. If one thinks of indication as a type relation, Dretskean indication allows the latter but forbids the former[13].

Godfrey-Smith notes that Dretske's view on the direction of dependency is less clear in his more recent work than it was in *Knowledge and the Flow of Information*. He writes:

> It must be admitted that Dretske's 1988 treatment is not as univocal on this point as his earlier work. His main discussion (1988, pp. 54-59) gives no formal definition. Some of the cases he mentions are clearly one-way. Doorbells indicate someone at the door,

> but you can stand at my door without ringing. Fingerprints tell us who held the gun, but a gloved killer leaves no sign. He also refers to his (one-way) 1981 account as a generally accurate formalization of his 1988 concept. However, around, pp. 96-97 there are puzzling expressions that suggest a two-way understanding of indication. Generally, I will take these to be aberrations, and assume that Dretske's theory involves only a one-way, head-world, dependence. (Godfrey-Smith 1992, p. 287)

The 1988 references are to *Explaining Behavior*; the 1981 reference is to *Knowledge and the Flow of Information*. Godfrey-Smith's view on this point is the safe one. Nowhere in Dretske's specific development of his concept of indication does he suggest a requirement of "world-head" reliability in addition to "head-world". Where two-way dependency does creep into the discussion is afterward, when Dretske's interests have shifted to putting indication to work in accounting for representation and explanations of behaviour. In these later contexts, Dretske is simply not as careful he ought to be about his own conditions on indication, especially when elaborating his theory with examples.

So, to put together the discussion to this point, I give the following definition of Dretskean indication:

A token of C indicates s's being F iff

1. s is F
2. C is a type that is dependent on s's being F (i.e. in the absence of something's being F, a token of C would not have occurred).

(ii) Indication and Dependency

Baker's version of Dretskean indication also picks up on a reference Dretske makes to normalcy conditions in *Explaining Behavior*. Here is an example from Dretske:

> If squirrels changed their habits (because, say, doorbells were made of nuts), then a ringing doorbell would no longer mean what it now does. But as things *now* stand, we can say that the bell would not be ringing unless someone was at the door. It therefore indicates or means that someone is at the door. But this subjunctively expressed dependency between the ringing bell and someone's presence at the door, though not a coincidence, is not grounded in natural law either. There are surely no laws of nature that prevent small animals from pressing, or randomly falling meteorites from hitting, doorbells. There is certainly nothing in the laws of physics that prevent an occasional short-circuit in the electrical wiring, something that might cause the bell to ring when no one was at the door. *Normally*, though, these things don't happen. (the last emphasis is mine) (Dretske 1988a, p. 57)

Such comments suggest a softening of the strength of the dependency requirement for indication.

As we have seen in Dretske's view of information in *Knowledge and the Flow of Information*, Dretske takes a very clear and strong view of the requisite dependency between an information-carrying signal and a distal state of affairs. It must be the case that the C-token (or signal r) could not have come about in the absence of condition F. The strength of the dependency needed to ground the subjunctive condition is that the conditional probability of an F-token given a C-token is 1. What kind of dependency is required for the conditional probability of a quail's having passed by given this set of tracks to be 1? The suggestion in Dretske's earlier work is that the correlation between C and F must be grounded logically or nomologically:

> Correlations, even pervasive correlations are *not* to be confused with informational relations. Even if the properties F and G are perfectly correlated (whatever is F is G and vice versa), this does not mean that there is information in s's being F about s's being G (or vice versa). It does not mean that a signal carrying the information that s is F also carries the information that s is G. For the correlation between F and G may be the sheerest coincidence, a correlation whose persistence is not assured by any law of nature or principle of logic. All Fs can be G without the probability of s's being G, given that it is F, being 1. (Dretske 1981, pp. 73-74)

Thus there is an implied connection between the conditional probability of F given C being

I and the dependency of C on F being grounded in logic or natural law. However, this requirement that such dependencies be grounded either in natural or logical laws is qualified in *Explaining Behavior* to include grounding in some "locally valid" law:

> Chance correlations between two variables, no matter how prolonged, are not enough. In order for one thing to indicate something about another, the dependencies must be genuine. There must be some condition, lawful or otherwise, that *explains* the persistence of the correlation. This is the difference between a lucky run of heads with a fair coin and the not-at-all lucky run of rings when someone is at my door, a difference that enables my bell (but not coin flips) to indicate something about the correlated condition. This, of course, is a fact about *my* house, *my* neighborhood, and *my* doorbell wiring. If your house or neighborhood is different, maybe the ringing of *your* doorbell means something different. (Dretske 1988a, p. 57)

Dretske then continues:

> In many cases of biological interest, a sign -- some internal indicator -- on which an animal relies to locate and identify, say, food -- will only have this kind of local validity. It will, that is, be a reliable indicator only in the animal's natural habitat or in conditions that approximate that habitat. (Dretske 1988a, p. 57)

The requisite relation between conditions of C's type and those of F's type, necessary for C to indicate that F, may not meet conditions for constituting a (natural) lawful or logical relation, but the relation is still not mere coincidence.

Perhaps the criteria of dependency is meant to apply to environments or contexts[14]. It must be the case that in this environment/context nothing but quail cause tracks like *these*[15]. It must be the case that quail cause such tracks and non-coincidental that "locally" nothing but quail cause them.

Such "local laws" will turn out to include regularities assured by designers and/or managers of artificial, experimental environments[16]. No natural law guarantees that food will be available to a rat in an experimental habitat every time a given signal (a light, for example) is activated. That

food is available when a light is illuminated in this instance is a dependency accounted for by the actions of the experimenter. The correlation between availability of food and the illumination has an explanation and in this sense is not merely accidental, but neither is it grounded in a natural law.

Let us imagine three different cases.

Case (1): Nothing but quail cause x-tracks in a certain region of broken woodland (call it Mirkwood). Pheasants also cause x-tracks, but no pheasants have ever passed through Mirkwood. Pheasants are found in other areas like Mirkwood, and there are some pheasants in fields neighbouring Mirkwood. However, as a matter of fact, no pheasants have ever found their way into Mirkwood.

Case (2): Nothing but quail cause x-tracks in Mirkwood. Pheasants also cause x-tracks, but no pheasants have ever passed through Mirkwood. Pheasants are not found in areas of Mirkwood's type because the lack of tall grasses in wooded areas of Mirkwood's elevation make it unsuitable for pheasants to nest and to hide from predators.

Case (3): As far as any relevant expert can tell, pheasants could thrive in Mirkwood. Members of this species in fact do exist in ecological systems like Mirkwood in other parts of the world. However, as a result of quirks of evolution pheasants do not occur in Mirkwood[17].

In case (2), but not in case (1), a set of x-tracks in Mirkwood (supposing that nothing other than pheasants and quail produce x-tracks) indicates a quail. In neither case is there a natural law (where these are construed as global principles) which grounds a dependency of x-tracks on quail. However, there is some local "ecological" law which grounds a local dependency of x-tracks on quail in case two. Of the creatures that could cause x-tracks, only quail are found in Mirkwood-type environments because that type of habitat is unsuitable for birds of pheasant's

constitution.

The interpretation suggested may still be too strict. Dretske especially wants to allow "locally valid" dependencies of a kind that are found in biology, and that are exploited by organisms to support indicator relations. In many cases of such exploited indicators, there may be no reason why some state has managed to be an indicator in one environment but not in others. A C-type state in a predator in environment # 1 may be caused both by potential prey and by some poisonous non-prey. In environment # 2, let us suppose, members of the poisonous non-prey species happen to be absent even though it is an environment in which they *could* thrive. State type C would appear to be a prey-indicator in environment # 2 but not in environment # 1. Yet, there is no ecological or other law-based explanation for the absence of poisonous non-prey in environment # 2.

Perhaps one could liberalize the dependency condition to say that for C to indicate F, C-types must be dependent on F-types and it must be the case that, in a given context or environment, no signal of C's type has occurred that was not dependent on a condition of F's type. This would mean that indication is satisfied in the previous case (1) (where pheasants are around Mirkwood and could have found their way into Mirkwood but as a matter of fact none ever have). I think Dretske would regard such a case as too weak to support an indication relation.

What about case (3)? As with the previous example of the absence of poisonous non-prey, we are to suppose that there is no reason (based on the biological constitution of a pheasant) why pheasants could not exist in Mirkwood. So in a sense it is merely accidental that x-tracks in Mirkwood are never produced by pheasants. It is therefore accidental that only quail have produced x-tracks in Mirkwood and thus one might conclude that case (3) fails to provide the

necessary dependency for a set of x-tracks to indicate a quail. However, the sort of locally valid "ecological laws" which Dretske wants to include as providing the necessary dependency for indication are precisely insensitive to just this kind of situation. A creature with an interest in quail but not pheasants and which develops an x-track detector does so presumably without regard for the reasons for the persistent absence of pheasants. It is enough that distinguishing and following x-tracks leads to quail and not pheasants. It surely does not matter whether this is an accident or a matter of physical necessity[18].

Of course, case (3) is in another respect one in which the absence of pheasants is non-accidental anyway. There may be no biological reason why pheasants could not have evolved and thrived in Mirkwood, but, given that the species did not evolve in that part of the world, the absence of pheasants in Mirkwood is not a coincidence. The explanation would appeal to geography and the flight and migratory habits of pheasants. The populations of pheasants that do exist have evolved in regions distant from Mirkwood (for whatever reason). Pheasants have not subsequently found their way to Mirkwood because they are not particularly mobile birds. It is not clear to me that this kind of case would ground a dependency of x-tracks on quail (in Mirkwood) in any kind of law, even a locally valid one. It does make it for practical purposes impossible that a set of x-tracks should appear in Mirkwood in the absence of a quail (again given the assumption that nothing besides pheasants or quail produce x-tracks in this world).

The strength of dependence of C-types on F-types lies somewhere between case (1) and case (3). It is too weak to say merely that, as long as tokens of type C have always occurred in the presence of an F-token, a current C-token (occurring in the presence of a token of condition F) is an indicator of s's being F. It is too strong to say that the inevitability is a matter of logical or even

nomological necessity. If one says of indication that a C-type could not occur in the absence of F, the implicit notion of possibility seems to be set at practical possibility. We do not worry that practical jokers could conceivably re-create x-tracks, or that the local woods, Mirkwood, might be invaded by space-aliens with feet like quail.

This apparent broadening of the grounding for the dependency between C and F in fact raises two issues in Dretske's theory. The first is about the notion of indication itself. Can one still make sense of the idea that a C-token does not indicate that something is F unless the conditional probability of an F given a C is 1? There is some flexibility in the standards one can require for some event (or event type) to be assigned a probability of 1: the standard could be that the event's occurring is a logical necessity, or that it is a nomological necessity. Are the standards so flexible that the probability of something's F occurring given the occurrence of a C-token is 1 so long as it is non-accidental that C should occur when F occurs?

This first issue is not an especially pressing one. Given the kinds of exceptions that Dretske mentions, it is still clear that, as he means to use "indication", C cannot indicate F unless the occurrence of a "false positive", a C-type in the absence of an F-type, is an extraordinary occurrence. Within the normal constraints or conditions of a local environment, tokens of type C cannot occur in the absence of the condition of something's being F. Of course the extraordinary might occasionally occur: rare ringings of my doorbell might actually be the product of a short-circuit or the work of a clever squirrel. As these events fall outside the normal conditions of my environment, they do not count against a token of a ringing of my doorbell being a person-at-the-door indicator.

(iii) H-indication

The second issue is whether Dretske's requirements for a type of state tokens of which are indicators to become a type of representational state are weakened by the broadening of the dependency requirements for indication. Tokens of C being *representations* of s's being F, rather than their being indicators of s's being F, require that a representational system have the right history. For systems whose representational states come about naturally rather than conventionally[19], C's (now) representing something as F requires C-type states of the representational system in question to have had a history of indicating F. There has to be a history of a non-accidental correlation between past tokens of C's type (C-tokens, for short) and the external condition F. Sometimes both Dretske and critics alike refer to a type of state's having a history of indicating F as "C's indicating that F".

How strong must this correlation be in order for it to be possible for tokens of C to be F-representations (representations of thing's being F)? On Dretske's earlier view, the answer is again quite clear. Since the probability of an F given a C must be 1 (for indication), a current token of C cannot be an F-representation unless all of the past tokens of C (in some system) were indicators of F[20]. We could thus define a notion "H-indication" (for "has a history of indication") for a state of type C. Following the earlier Dretskean development of information-carrying, we define H-indication as follows:

H-indication $=_{df}$ A state of type C H-indicates condition F for some system S for some period of time t iff every token of C in S prior to t has indicated F.

H-indication is a necessary condition on the earlier Dretskean view for representation. One might suspect, however, that the Dretske of *Explaining Behavior* weakens the conditions for

H-indication. This would amount to weakening the indication-history requirement for a state of a given type to be a representation. The requirement embodied in the foregoing definition is a 100% correlation of C-tokens with the condition F (although F can occur without a given representational system tokening C). It is also built in, because it is built into the notion of "indication", that the correlation is non-accidental. Does the Dretske's requirement of perfect correlation for H-indication weaken as his view of the requirements for indication weakens? Does Dretske allow that a current C-token in some representational system can be a representation of F even if some past C-token in that system has not been an indicator of F?

The question of what the dependency requirement is for a token of C to be an F-indicator is a different question than the question of what the required correlation between a state C and condition F is in order for C-tokens to be F-representations. The first question is, what is the force of the modality in the counter-factual "If F had not been present, then C would not have occurred"? The second question is, how reliable must the correlation of tokens of C with condition F be for C to represent F? Dretske changes his answer to the first question, but there is no reason to think that he thereby changes his answer to the second. The later Dretske still holds that representation requires H-indication.

Godfrey-Smith does not separate the two matters I have identified. Nor does he keep clear about indication being a "token" relation which has a "type" component. He notes the broadened allowances for dependency between a sign and what is signified:

> The 1981 notion of information seems so demanding that one wonders whether messages with interesting contents are ever really transmitted. By Dretske (1986), the requirements had been loosened considerably. There 'information' is replaced by 'natural signing'. (Godfrey-Smith 1992, p. 288)

He then goes on to cite a couple of supporting examples from Dretske:

> Natural signs are indicators, more or less reliable indicators, and what they mean is what they indicate to be so. (Dretske 1986, p. 18)

> In most cases this relation is causal or lawful, one capable of supporting a counterfactual assertion to the effect that if the one condition had not obtained (if Tommy did not have measles), neither would the other (he would not have those red spots all over his face). Sometimes there are mere regularities, non-lawful but none the less pervasive, that help secure the connection between sign and significance. (Dretske 1986, p. 19)

One might notice that the discussion in these quotes concerns what it takes for something to be an indicator. On Dretske's view, this is different from what is required for that same "sign" to be a representation. Dretske's theory, unlike some other similar ones[21], bases representation only indirectly on indication (or some corresponding physical dependency relation). The first quote is also an illustration of Dretske's own relative indifference to a type/token distinction, which is a cause of some of the difficulties of interpretation.

Godfrey-Smith then continues,

> Much of the 1988 presentation echoes the 1986. Perfect correlation is apparently not required, and non-lawful regularities can contribute to indication. The requirements for being an acceptable, non-lawful element in indication are spelled out in more detail, however, and while many of the same examples are used in the 1986 and 1988 versions, the latter may be a little stronger. In the 1986 version it appears that the sheer 'pervasiveness' of a regularity is sufficient for it to contribute to indication. (Godfrey-Smith 1992, p. 288)

Godfrey-Smith then illustrates with a couple of quotes from *Explaining Behavior*. The quotes used are parts of passages I have cited previously in developing the idea that a C-token's indicating s's being F has to be grounded in a correlation between (type) C and (type) F that is at least non-coincidental.

The points of Dretske's to which Godfrey-Smith draws attention, first of all, must be

understood as pertaining to the requirements for a token of C to *indicate* something's being F. Nothing in the discussion so far pertains to the requirements for C-tokens to *represent* something as F.

There are, moreover, (at least) two ways of reading the foregoing Dretske claim that "[n]atural signs are indicators, more or less reliable indicators, and what they mean is what they indicate is so". One reading (which I take to be mistaken) is to suppose that Dretske has "loosened" his indication requirements so that all that is required is that a C-token be of a type that normally (or reliably) occurs in the presence of F. This reading would be to loosen the concept of indication so that it is possible for a C-token to indicate an F-token even if F is not present. A second reading is that C-tokens must be reliably (but not necessarily perfectly) correlated with condition F in order for *any* given C-token to be a candidate for F-indication[22]. However, such a token C is still not an indicator (even if it is of a type instances of which can be F-indicators) *unless it does occur in the presence of F*. Dretske never does waiver on this point.

Godfrey-Smith seems to read Dretske's comment the first way and so gives a modified (weakened) account of indication (understood by Godfrey-Smith as a type relation) on which all instances of state of C's type are F-indicators if and only if type C is reliably and non-accidentally correlated with type F[23]. The suggested standard of reliably is that F be more probable than not, given a token of C. He notes, however, that his revised notion might not do:

> Some may find this far too weak an account of indication. It is certainly too weak to meet many of Dretske's original (1981) requirements of a theory of information. And though he does not come out and say it, there are signs in the 1988 discussion that Dretske still thinks of indication as requiring perfect correlation: Why else would he insist that there is no such thing as misindication (1988, pp. 55-56)? (Godfrey-Smith 1992, p.289)

This is an interpretive matter that Godfrey-Smith does not pursue further because his own purposes do not require him to do so. It is clear from the comments here that Godfrey-Smith is conflating two issues. He is reading comments of Dretske's which suggest that Dretske loosens the dependency requirement for a token of C to indicate s's being F. Dretske does loosen this requirement. From the loosening of this requirement, it does not follow that Dretske is allowing that a token of type C could indicate s's is F in the absence of s's being F. Nor does it follow that Dretske is lowering the requirements set for state type C of some system to be "promoted" to a representation of F. These are separable issues.

There is a fairly straight-forward way of understanding what is going on in Dretske's view. The only change that occurs from the earlier to the later Dretske is a relatively minor one of the strength of the modal dimension of indication: what is required for it to be true that if F had not been present, C would not have occurred.

Dretske still does require that there is no such thing as misindication[24], but this has nothing to do with the strength of (type) C's correlation with F. Once again a token of C, indicates s's being F if and only if the following two conditions are satisfied:

(a) s is F

(b) If s had not been F, the token of C would not have occurred.

Talk of reliability of correlations between (type) C and (type) F, and the grounding of those correlations all pertains to condition (b). It still remains, whatever strength of correlation is required and whatever grounding of the correlation is required, that the token C does not indicate F if s is not F. This is why, in spite of qualifications of the initial (1981) account, there is still no such thing as misindication.

If the impossibility of misindication has nothing to do with correlations between type C states and type F conditions, what *does* correlation have to do with in Dretske's theory? One area which is touched by correlation is the second condition on a C-token's being an F-indicator. The grounding of the correlation between C and F must be such that (within an environment) if there have been any C-tokens which have occurred in the absence of condition F, they must have been *aberrations*. On Dretske's later (1988) view, a given C (occurring as it must in the presence of F) apparently can indicate F even where some C-tokens have been false positives. The stipulation is that such occurrences must be abnormal occurrences of type C.

The other area in which correlation fits into Dretske's theory is in the "promotion" of a type of indicator to a representation. The details of such promotions will be described later on. For present purposes, the rough idea (for natural systems of representation) is that a system S comes to have a state type C which represents things as F if S's behaviour has been structured (or modified) and would not have been so if S's past C-tokens had not indicated condition F. Type C's (now) being a representation depends in part on (type) C's having a certain correlational history with F: past C-tokens (for some S, for some period of time) have been F-indicators. It is convenient and therefore very tempting to refer to this historical relation simply as "C's indicating that F", or "C's having indicated that F", where "C" in these formulations refers to the type C and not a token C. I have tried to distinguish this notion by introducing the expression "H-indicates" so as to avoid potential confusions between a C-token's indicating F, and C's being a type tokens of which (for some S) have indicated F or do indicate F.

So Godfrey-Smith does not properly distinguish between two separate points in Dretske's theory. One point is the strength of dependency of C-tokens on condition F. This strength of

dependency (which appears to have been weakened in *Explaining Behavior*) was initially characterized in terms of conditional probability: the conditional probability of an F given a C must be 1. The second point is the strength of correlation required for H-indication. Obviously, it will have to be 100% on the earlier, strong view about the strength of dependency for indication. It is obvious that the strength of correlation for H-indication is not *required* to be 100% if, for indication dependency, the conditional probability of an F given a C is less than one. On the other hand, it is not at all obvious that Dretske changes his view about the strength of correlation required for H-indication. In fact, there is good reason to think that he does not[25].

In relation to notions of error or failed co-occurrences, we are now in a position to distinguish a number of different questions on the *Explaining Behavior* view:

(1) Can a C-token be an indicator of F even if some tokens of type C have been false positives of F[26]?

(2) Can *any* token of type C be an indicator of F if that token is a false positive of F?

(3) Can a state of type C, for some representational system S, be a representation of F if some of S's C-tokens have been false positives of F?

The answer to (1) is yes, provided that within the normal constraints of the given environment there have been no such false positives (i.e. any false positive is an aberration). The answer to (2) is no, absolutely not. There is no such thing as misindication (being an indicator which fails to indicate). The answer to (3) is yes. After type C has already been "promoted" to an F-representation, there can be C-tokens which are false positives of F. In fact, they can all be false positives. However, prior to type C's promotion, there cannot be any C-tokens which are "false positives". For the period during which type C is in the process of being promoted, C-tokens must

be specifically correlated with occurrences of F. That is, there cannot be any mutually exclusive condition G in the presence of which some C-token also occurs.

A response to Slater serves to show Dretske's continued strong view on the requirements for promotion to representation. A couple of introductory comments are necessary. First, Dretske uses "indication" in the passages I will quote in the way in which I have used "H-indication". Certainly, he refers to types as indicators rather than a token as an indicator. Second, instead of talking about representation, Dretske refers to concepts. Third, "discrimination learning" is (as will be explained later) one of the ways in which a type of indicator can be "promoted" to a representation. Fourth, "contingency" refers to less-than-perfect (and perhaps accidental) correlation between an internal state and some external condition. In this first selection, Dretske reveals the distinction between contingency and indication (H-indication):

> I appealed to discrimination learning because I thought primitive concept formation (the only kind at issue here) is a species of discrimination learning, not because I thought all discrimination learning involved concept formation (and, thus, indication). I agree, in other words, that it is contingency, not indication, that is important in discrimination learning. There doesn't have to be a worm under *every* leaf (thus making leaves indicate worms) for birds to learn to look under leaves for worms. Birds will learn to do this if there is the appropriate contingency, if the probability of a worm given a leaf is significantly higher than the probability of a worm given no leaf. (Dretske 1994b, p. 204)

There is much that needs to be explained and questioned about this passage. For present purposes, it is enough that one finds here some guidance on what sort of history is required for representation (concept formation) and what merely supports discrimination learning. Extrapolating from the bird example, it seems that in order for a state caused by the perception of a leaf to represent the presence of a worm (for some bird), leaves need to be indicators of worms. For leaves to indicate worms, there needs to be a worm under every encountered leaf. In the

formation of this "concept", the bird cannot encounter false positives: leaves with no worm underneath. Even though the modal component of a token C's indicating F has apparently been loosened between 1981 and 1988, there is evidence here that representation of F can only come about if (in fact) every occurrence of a token C is an F-indicator. (This strong requirement on indication will be the subject of later criticism.) The notion of "indication" Dretske employs here is stronger than mere "contingency", which seems to be simply a positive correlation between (for example) leaves and worms. The most important strength difference between the two notions here will turn out to be *frequency*. This is already fairly clear in the bird example.

Dretske asks that we consider an example of the learning of a colour concept in order to understand the need for "indication" (i.e. H-indication):

> To understand why contingency is not enough (though it may be necessary), why indication is needed, suppose that we have three properties -- purple, dark red, and pink -- and we are trying to teach S the color purple -- what it is to be purple -- in lighting conditions (or maybe it is just S) in which purple and dark red both look the same to S although both look different from pink. Both purple and dark red look 'dark' while pink looks 'light'. We nonetheless teach S to say 'purple' to everything that looks dark and not to say this to things that look light. If there are an equal number of purple, dark red, and pink stimuli, there will be a contingency between something's looking dark to S and its being purple. It will be purple 50% of the time it looks dark, while it will be purple 0% of the time it looks light. So we have contingency between its looking dark to S (this is the internal C) and its being purple. But we have no indication. C (looking dark) does not indicate that the color is purple. The most it indicates is that the color is *either* purple or dark red. (Dretske 1994b, p. 205)

The conclusion he draws from this example is that in the described circumstances S might learn the concept DARK, but could not learn the concept PURPLE. Notice that as the example is described it is more than a mere possibility that some dark stimuli are dark red. S is exposed to equal numbers of the three stimuli, so some of the dark stimuli that cause C-tokens will not be purple.

Here is Dretske's explanation of why S cannot acquire PURPLE in the described conditions:

> What is missing in this case is something inside S that indicates the stimulus is purple. Though there is a subjunctive condition in S that raises the probability of a stimulus being purple (from 0% to 50%) there is nothing in the in the child (C) that, during learning, raised the probability *enough*. Nothing indicated that the color was purple. If you are going to learn what purple is, you have to be exposed to this color in conditions in which this color looks different from the range of colors that this concept is supposed to contrast with (I am not suggesting that one cannot learn colors in a context in which dark red is not contrasted with purple). To learn colors in this way you must be exposed to colors in conditions in which the way colors look (C) is the way colors are (F). (Dretske 1994b, p. 205)

If one reads "indicates" as meaning something like what I have labelled "H-indicates", it is possible to make sense of the whole passage. S does not acquire the concept PURPLE because no type of internal state C, during the prospective "promotion period" (learning) has the appropriate indicator history. The appropriate indicator-history would be for every token of some type C to be an indicator of purple. Every C-token, in other words, must (in this case) be caused by a purple sample and it must be no accident that there were no false "purple" positives. I take this to be implied by the last two statements: the second to last says, in effect, that internal states caused by presentation of purple samples must be different from the internal states that would be caused by contrasting colours. Internal state type C cannot (even possibly) sometimes be caused by samples of dark red if it is to be an H-indicator of purple. The last sentence says that the conditions in the "promotion period" must be such that "purple looks" (instances of type C) must be caused be actual purple samples. Nothing suggests that Dretske is offering a qualified view on which some or most "purple looks" must be purple-caused. The general claim seems to be straightforwardly a universal one: at least for the period during which S is in the process of acquiring a representation of F, for type C to be that representation, every C-token has to be an F-indicator. The importance in

Dretske's view of no false positives in representation acquisition is confirmed by the first of these three quotes when he says that a worm under every leaf would make leaves a worm-indicator.

Again, I believe the strong position reflected in Dretske's reply to Slater (written in 1996) is compatible with the sort of qualifications that Godfrey-Smith picks up on if one sees Dretske's qualifications in *Explaining Behavior* as nothing but a loosening of the modal condition on C's indicating F. A C-token can indicate F as long as type C is dependent on F such that within the normal constraints of some environment tokens of type C would not occur in the absence of F. A token of C still cannot indicate F if F is not present. Moreover, a representation of F still cannot be acquired if, during the "promotion period", there are any aberrations (C-tokens which are false positives). H-indication still requires, as suggested by the worms and the colours that every C-token be a "normal" token (occurring when F is present). During the promotion of type C to a representation of F, the correlation of C with F has to be perfect.

Such a requirement in fact allows a tidy "solution" to a disjunction problem: how could an internal state which is dependent on either of two mutually exclusive condition types, F and G, manage to represent something as F? The Dretskean answer is that it cannot. If, during promotion, some C-tokens are F-dependent and some are G-dependent, C-tokens at best indicate "F or G" (or some condition that captures this disjunction). So if type C comes to represent anything at all, it represents "F or G" (or some condition that captures this disjunction). Even if, during promotion, a C-token might (within the normal constraints of the environment) have been G-dependent instead of F-dependent, type C will not come to represent F.

To summarize, Dretske's basic concept of indication undergoes some loosening from earlier to later writing, but the loosening is not especially significant. It pertains to the modal requirement

(the second condition) for a token C to indicate that F. Once again, (for some environment E) a token of C indicates that s is F iff

(a) s is F

(b) If F had not been present, then a C-token would not have occurred.

The force of the counter-factual is that, within the normal conditions of E, type C is dependent on F. If one likes, one can say that within E's normal conditions or constraints, the conditional probability of an F-token given a C-token is 1 and the conditional probability of an F-token given that there is no C-token is less than 1.

This is a "token" notion of indication. However, it is defined partly in terms of types, for no C-token could be an F-indicator if (type) C was not normally dependent on F. It is also useful to have a historical "type" notion of "H-indication":

For some representational system S, state type C H-indicates F over some time t iff every C-token of S over t indicates F.

It is important to remember also that in Dretske's theory the concept of representation is not founded directly on indication. Representation is thought to have important properties that Dretskean indication lacks. One that has been discussed already is the potential for error. Misrepresentation is possible; misindication is not.

It is important to emphasize one additional point about indication. Indication is quite non-specific. For example, if a C-token, C_t, indicates that something is F and, logically, things are F if and only if they are G, then C_t indicates that something is G. The dependency of C_t on G cannot be accidental since C_t indicates F and the relation between F and G is one of non-accidental dependency. The "chain of indication" can continue as long as there are further conditions

non-accidentally related to F or G that meet the conditional probability requirements. So if conditions of G's type are always caused by those of H's type, then C_t, by indicating F and G, also indicates H. Also, if nothing causes conditions of H's type without causing a condition of I's type, then C_t indicates I as well. I am of course assuming that the conditional probabilities of an I given that C does not occur and of an H given that C does not occur are each less than 1. The point will generalize for H-indication. If, for some system S, state C has H-indicated condition F over some period of time, then it will also (in the circumstances described) have H-indicated G, H, and I.

The non-specificity of indication is quite important, for the target notion of representation is supposed to have the kind of specificity that indication lacks. A token of C, in the cases of representation in which we (and Dretske) are interested, can represent something as F without representing it as G even if, non-accidentally, something is F if and only if it is G. Whatever it is that "promotes" a state C such that its tokens are representations of F (from C's having H-indicated F) must somehow account for this specificity of representation.

NOTES

1. Dretske actually seems to deny this in relation to desires, though he certainly holds it in relation to beliefs. However, in denying, or seeming to deny that desires are representational, he also denies that these kinds of states are *about* anything, preferring instead to say that their intentionality consists in their "*for*-ness". I will have more to say on this subject in my discussion of motivational states.

2. In this section, I follow the abbreviation conventions found in *Knowledge and the Flow of Information*. From the discussion of indication onward, I use those found in *Explaining Behavior*.

3. In *Knowledge and the Flow of Information*, Dretske explains the mathematics behind the calculation of information quantities, measured in bits in some detail. In the kind of case at hand, suppose that the probability of any of the eight employees being selected is equal, the information resulting from the selection of Jones is 3 bits. This is calculated by taking the $\log_2$ of 1 over the probability of Jones being selected (.125), that is, $\log_2 8$.

4. He also concedes that there is more than one way of looking at this issue (Dretske 1981, pp.78-80). The example he employs to illustrate is of two subjects who observe the third of four shells being lifted to reveal that there is no peanut underneath. Supposing that there is a peanut under one of the four shells, and that subject one but not subject two has seen that there is no peanut under the first and second shell, subject one obtains the information that the peanut is under the fourth shell. Subject two does not obtain this information from the observation that the peanut is under the third shell. Dretske notes that one could reasonably argue that subject one does not obtain the information that the peanut is under the fourth shell *from the observation that the peanut is not under the third shell* either. One could say that the signal received by seeing under the third shell is itself the same for both subjects, but that subject one has received information from other signals that subject has not received. Dretske rejects this view, opting to "relativize" information to receivers: "I propose to follow Dennett in relativizing the information contained in a signal because I think this accurately reflects our ordinary ways of thinking about such matters" (Dretske 1981, p.79). His reasoning for this relativization is obviously not especially strong. One has to think that this issue of whether or not (for example) different receivers receive the same or different information from the same signal (depending on background knowledge) is of little theoretical concern to Dretske. As far as I can see it is an issue of no concern to the development of his theory of mental representation.

5. There are a couple of respects in which Slater's passage is ahead of my explication. It is stated in terms of indication, not information. However, on Dretske's use of indication, x indicates y if and only if x carries the information that y. Also, Slater uses the letter C, which Dretske uses to stand for internal states of systems that are caused by some external condition of a local environment and are at least potential components of behaviour. In the case of cognitive beings, these would be neural states.

6. Weaker information-theoretic accounts of mental representation have been derived from this more general criteria. Two examples are Dan Lloyd's theory, developed in his book *Simple Minds* (Lloyd, 1989) and a theory developed by Don Ross and Tad Zawidzki Jr. in their article "Information and Teleosemantics" (Ross and Zawidzki, 1994).

7. Even this "token" relation still depends in some way on a relation between signal r's type and the condition s's being F's type, but the relation "carries the information" itself relates tokens.

8. Actually, this deserves some qualification. What s can indicate will turn out to be partly dependent upon what observers are capable of distinguishing.

9. Stampe's view is given in his article "Toward a Causal Theory of Linguistic Representation" (Stampe 1977).

10. "C-token" replaces "r" and (an instance of) "condition F", "s's being F", or (for short) just "F" replaces "P".

11. For one thing, Dretske allows the possibility of indicators arising in contrived environments such as those constructed by psychologists for purposes of studying animal behaviour. One example Dretske employs, an example I take up in detail in Chapter 4, is of a rat trained to press a bar when a light is illuminated (because pressing the bar releases food from a hopper). The experimenter has contrived the environment in such a way that there is a (temporary at least) dependency between the presence of food in the hopper and the light's being illuminated. There is also similarly a dependency between the presence of food and the perceptual state of the rat which is caused by the light's illumination. However, neither the perceptual state nor the illumination of the light is caused by the presence of food. The coincidence of these conditions is caused by the experimenter and his/her purposes.

12. Actually, the step requires that C-tokens of a system, over some developmental period, be indicators of s's being F.

13. The terms "false positive" and "false negative" also appear in Godfrey-Smith's article "Indication and Adaptation", especially section 5 (Godfrey-Smith 1992, pp. 297-309). I introduce this terminology now not only in hopes that what Dretske is and is not committed to will be perfectly clear but also because it will be useful later on.

14. I add "context" here because it seems odd to talk of one's neighbourhood or one's house as an environment.

15. It is the notion of an environment or a context which would defy a very precise formulation. This is what makes the dependency condition that Dretske means somewhat vague.

16. See, for example Dretske 1988a, pp.115-117.

17. A good example of the kind of case I have in mind is that of North America and the European starling before the nineteenth century. There were no such birds in North America prior to their artificial introduction. We now certainly know that starlings are very well suited to North American environments -- there is no biological reason why they could not thrive here. However, they did not evolve here, and, but for an accident, there is no reason think they ever would have found their way to this continent.

18. Thispoint willactually prove to be something of a problem for Dretske.

19. The distinction between natural and conventional representations and representational systems will be spelled out in my section on representational systems.

20. Actually, it should be "all of those up to the type C being 'promoted' to an F-representation". The meaning of this qualification will be explained under the discussion of Dretske's "Design problem", which deals with the issue of "promotion" of a type of state from one whose tokens indicate to one whose tokens represent some condition.

21. Such as Lloyd's in *Simple Minds* (Lloyd 1989), Robert Stalnaker's in his book *Inquiry* (Stalnaker 1986), or Dennis Stampe's in his article "Toward a Causal Theory of Linguistic Representation (Stampe 1977).

22. In the dead of a winter's night, the quail-footed aliens land in Mirkwood. A scout is sent out to look around. It reports back that there is nothing of interest on this little planet and the aliens leave to seek out strange new worlds. The set of x-tracks left behind by the alien scout fail to indicate a quail. Next week, though, when I encounter yet another set of x-tracks (one that was caused by a quail in a Mirkwood in which only quail cause such tracks, given the normal constraints and conditions of the local environment), the tracks will indicate a quail.

23. On page 289 of his 1992, Godfrey-Smith initially uses "only if" and not, as I have, "if and only if" when formulating his revised version of indication in the first full paragraph on the page: "That is, s indicates that p only if p is more likely than not, given s". In the next paragraph, however, he suggests that this notion of indication could be taken as a sufficient condition as well.

24. By this I mean (as does Dretske) that no token of a type C can be an indicator of s's being F that fails to indicate s's being F.

25. I should note here, however, that Godfrey-Smith's failure to distinguish these features of Dretske's account does not affect the critical points he makes against Dretske. See chapters 5 and 6.

26. That is, tokens occurring in the absence of F.

CHAPTER THREE

Representation and the Design Problem

a. Introduction

In this first section on representation, I will outline briefly what Dretske's notion of representation is. I want to highlight certain features of it: the employment of the notion of function, that the account has a teleological as well as a statistical-informational strain, and that the formation of representational content is intimately connected with the formation of behaviour. I will then proceed in the next sections by discussing Dretske's typing of representational systems and his "Design Problem".

Indicator relations are (for Dretske) the building blocks of representation:

C represents something as F if and only if it *functions* to indicate F.

A function to indicate comes about (partially) as a result of C-tokens having a history of indicating that something is F. For C-tokens to have the appropriate history of indicating that something is F is for type C to H-indicate condition F. The period of time over which C must H-indicate condition F is the "promotion period" during which (type) C is coming to acquire its function. To acquire a function of indicating condition F, some behaviour of the representational system S must come under the control of internal state C. State (type) C must come to cause a movement[1] M that is in some way relevant to F[2]. That is, it must be that C would not have come to cause M if C-tokens had not had a history of indicating F. How a given type of movement can come to be under the control of a given type of internal state of a system is what Dretske calls "the Design Problem". A proper understanding of the Dretskean theory of representation will require elaboration of this problem and its solutions.

The concept of function is not at all carefully defined by Dretske. It is absolutely clear, however, that he has in mind some teleological notion, rather than a causal notion. A function is not merely what some structure (causally) has the power to do. A function is the "job" some structure has acquired or has been assigned, what it is supposed to do. "Function" is a purposive notion as Dretske employs it.

By defining representation in terms of what some state is supposed to indicate Dretske hopes to account for two difficult properties of representation: error and opacity (or specificity). Misindication is impossible. C cannot indicate F in the absence of F. C can, however, *function* to indicate F while occurring in the absence of F:

> Put a frog in a laboratory where carefully produced shadows simulate edible bugs. In these unnatural circumstances the frog's neural detectors -- those that have, for good reason, been called "bug detectors" -- will no longer indicate this (even when they are, by chance, caused to fire by real edible bugs) because their activity no longer *depends* in the requisite way on the presence of edible bugs. Taking a frog into a laboratory is like taking a compass down a mineshaft: things no longer work the way they are supposed to work. Indicators stop indicating. If we suppose, then, that it is the function of the frog's neural detectors to indicate the presence of edible bugs, then, in the laboratory, shadows are misrepresented *as* edible bugs. The frog has the analogue of a false belief. (Dretske 1988a, pp. 68-69)

Error comes to a C-token's failing to perform what it is (type) C's function to perform.

Intentional states, the type of representational state of primary interest, are more specific than physical dependency relations. Using the frog as an example, the neural state (C) of a frog which is a bug-indicator will equally bear a dependency relation to bug-facing-surfaces and bug-shaped retinal images. Moreover when the condition (F) of a bug's being present causes a C-token, so too will the condition (G) of there being a bug-shaped retinal image being present. If tokens of C indicate F, then, it seems, they indicate G as well. However, an account of representation that will

capture desired features of intentional states ought to allow that (for example) frogs have can states that represent bugs (F-representations) without those states having to be G-representations as well.

The idea of representations as indicator *functions* is supposed to allow for this. The frog's C-tokens which are F-indicator must also be G-indicators, but the function of those states is to indicate F. It is incidental that they also indicate G. This is possible because C tokens, while indicating both F and G, can acquire the *function* of indicating F without thereby acquiring the function to indicate G. Dretske's idea is that a type of internal state C can have a role to play in the economy of a representational system capable of some behaviour in virtue of C-tokens having been F-indicators without (necessarily) also having a role to play in virtue of having been G-indicators.

Determining what internal indicator-states are supposed to do is a matter of the behavioural role those internal indicator states play in the representational systems in which they occur: what type of movements do they cause and why did they come to cause those movements? In a biological example, such as the frog, Dretske's explanations of why some internal indicator plays the role it does will be given in terms of advantages or benefits to organisms. Edible bugs are of some benefit to a frog whereas bug-shaped retinal images are not. Teleology again enters into Dretske's account. Generally speaking, the determination of representational content for Dretskean representations, at least in biological systems, will be linked to utility: what would confer an advantage (adaptivity) or what would be of some benefit (desire-satisfying).

In the last section, I very briefly introduced the idea of a "promotion period". This is a period during which C, tokens of which are indicators of some condition F, comes to be a representation of F. This element of Dretske's theory is straightforwardly informational: C's coming to be a representation of F is a function of a statistical history of C-tokens indicating F. This particular

strain, the basing of representation, roughly speaking, on correlation Godfrey-Smith calls "reliabilism". Thus it will turn out that Dretske's view is distinguished from a straightforward informational theory, a kind of theory in which representation was based solely on some version of a statistical relation, by having both a reliabilist and a teleological component.

Representation and behaviour are clearly intimately linked in this theory. Dretske is taking on two broad tasks at once in developing his theory of representation, for he is trying to provide a theory which accounts for representation in intentional states in such a way as to show how these states are parts of genuine explanations of behaviour. Thus he is at once giving an account of how behaviour is structured as well as giving an account of representation such that certain kinds of states with representational properties figure in the explanation of behaviour. To develop the theory of representation in more detail requires discussion of two related topics from Dretske: types of representational systems and his so-called Design Problem. The latter will make clear how in Dretske's view the formation of behaviour determines the representational content of internal "indicators".

b. Type I, II, and III Representational Systems

Dretske's twin analyses of behaviour and representation distinguishes three types of representational system (called simply types I, II, and III). Among these only types II, and III manage to be solutions to the "Design Problem" (the problem of getting an internal indicator-state to trigger selectively a system output). Only type III representational systems are natural representational systems (organisms). These systems are the only ones that Dretske allows can come to a non-conventional or natural solution to his "Design Problem".

Type I systems are those systems in which what is represented, and the means by which

representing structures manage to represent are both controlled by some external agent. Dretske's example is of bits of popcorn and coins being used to represent basketball players in order to reenact a game play. The bits which stand for basketball players are dubbed tokens: they stand for basketball players because (intentional) participants in a certain kind of symbol game agree to thus use them. As dubbed tokens, the bits of popcorn and coins are conventional representations of basketball players. This is a functional role that has been conferred upon the tokens by some agent.

The popcorn and coins are conventional in another sense as well. Without the guiding hand of a symbol-using agent, the dubbed tokens cannot represent the court activities of the basketball players. In order for the pieces to be part of a system that represents basketball plays, there must be some isomorphism between the movements of the pieces and movements of basketball players. The patterns of pieces' movements and their relations to one another need to bear a systematic relationship to patterns of ball-player movements. The pieces have no intrinsic power to do this, but must depend on the manipulations of an agent who is in a position to translate basketball player movements into coin and popcorn movements. Type I representation is conventional in what is represented and conventional in its power to represent.

A type II system is still conventional in what it represents, but not in its power to represent. Dretske's examples of type II systems are often gauges or other measuring devices. A thermometer provides a good, simple example. The placement of mercury in a graduated tube, suitable marked represents ambient temperature. The device's representing temperature is something that is arranged by a designer who knows how to exploit the reaction of mercury to temperature change and who knows how to suitably calibrate the tube. The actual tracking of temperature is not something that an agent needs to manage, however. In a thin glass tube, mercury covaries with

temperature without any further extrinsic help. Unlike the coin, the mercury's states co-vary with temperature in virtue of intrinsic, natural powers of the mercury. The designer need only know how to make use of these powers.

Often conventional devices like gauges and metres are simply supposed to carry information to an intentional agent who can then use that information for various purpose. Some systems, however, are systems of automatic behaviour. A device such as a thermometer, a thermostat, or a motion detector can be connected into a system so that states of the representing device control states of some other part of the system. A motion detector can be wired into a system of lights and alarms so that when it is activated (a state that represents unauthorized entry) it turns on the lights and sounds the alarms. It manages to control the operation of lights and alarms (activating them when there is a medium-sized object moving within the field of the detector) because of causal powers of the components of the detector. The components have the particular effects that they have because a designer, who knows how to exploit these causal powers, has built the detector into a system such that its activation closes a switch which sends electrical current to a set of lights and alarms. The power of such devices to represent what they represent is natural but their representing what they do is conventional.

Type III and type II systems have the common feature that they represent in virtue of some intrinsic power to respond selectively to external conditions. How the two types differ is in how the power comes to be configured in such a way that it does represent (the presence or absence of) an external condition. Type II systems are contrived: constructed to do so by a designing agent. Type III systems represent in virtue of natural capacities or attributes to do so (namely, by learning or by genetic programming).

Put simply, there are two distinguishing components among representational systems. One is the source of a system's power to somehow "track" external conditions, to change state selectively. This, one might say, is a system's or state's power to be an indicator of some external condition. The other dimension is the means by which a system comes to represent some external condition. A system may have the capacity to indicate naturally (in virtue of some intrinsic causal power) or conventionally (in virtue of the continuous manipulation of some outside agency). A system may represent (function to indicate) by some natural process or may be constructed so as to be representational (conventional representation). So a rough way of differentiating the systems may be given as follows:

	POWER TO INDICATE	MEANS OF REPRESENTATION
TYPE I	conventional	conventional
TYPE II	natural	conventional
TYPE III	natural	natural

The cases of greatest interest (type III) have both some intrinsic power in virtue of which they indicate as well as their own power to represent. Let us call type III representational systems natural systems of representation. Correspondingly, a type III representation will be a natural representation.

Although the representational systems that are of primary interest are natural systems, Dretske draws out many of his observations about these from his discussion of type II systems. The advantage and danger of doing this is the same: type II systems provide an abundant stock of fairly simple examples. The advantage is an understandable analogical model of representation for

natural systems. The danger is that the model may be overly simple, for type II systems lack the crucial and puzzling ability that make at least some natural systems interesting: What is at issue is when representational systems manage to inform *themselves* by their representations. That is, when are the semantic components of a system's representations causes of the system's behaviour? In Dretskean terms, one wants an account of type III representation which makes sense of a representational state of a system S being a representation "to or for" S. (Dretske 1988a, p. 94). On Dretske's view, not all natural representations are representations "to or for" the system in which they are representations. Whether the representational properties of S are "to or for" S is a function of how the representing structures come about. This is the topic of the next section.

c. Solving the Design Problem: When Does Information Work?

States of things, on the Dretskean view, may co-vary with all kinds of external conditions. As previously noted, meeting the right correlational criteria is not representation for Dretske. A type of internal state of some system represents when it meets the right correlational criteria and has a behaviour-producing function to perform in relation to some external condition. However, still not all internal states which meet these conditions for being representations meet the conditions for being behaviour-producing states in which their being representational is explanatorily relevant. An internal state C's being a representation is explanatorily relevant only when it acquires its status as a representation in the right way. As Dretske puts this, it must be the result of the right kind of the solution to his Design Problem.

The Design Problem is explained as follows:

Consider the following common problem, whose general form I shall call the Design

> Problem: we want to build a system that will do M when, but only when, conditions F exist. How do we build it? Or, if we are already talking about an existing system, how do we get it to behave in this way? (Dretske 1988a, p. 96)

Dretske's Design Problem[3] has three solutions: the engineer's solution, the selectional solution, and the developmental solution. Here are concise descriptions of the three solutions:

(1) **The Engineer's Solution**: A designer gets a certain type of output out of a type II representational system by making use of natural electrical, chemical, mechanical, etc. relations in order to configure it to produce the desired kind of output. A designer, for instance, exploits the effect of ambient temperature on a bimetallic strip by fixing it into a thermostat. The result is a system that automatically turns a furnace off and on at selected temperatures.

(2) **The Selectional Solution**: Certain internal states of a type III representational system are hard-wired to produce certain kinds of output as a result of evolutionary (selectional) pressures on the system's genetic predecessors.

(3) **The Developmental Solution**: Environmental pressures (positive/negative reinforcement) on a type III system bring about modifications in the output produced by a given type of internal state. This is what I take it that Dretske has in mind by learning. Various terms are employed to describe this process, such as learning, operant conditioning, classical conditioning, instrumental learning, or discrimination learning.

Any solution to the Design Problem involves a system getting behaviourally structured. Some type of internal state C, which has been reliably and non-accidentally related to a condition F, comes to produce a movement of type M. Through an application of a Design Solution, state type C is "recruited" as a cause of M. C comes to cause M because past C-tokens have been

indicators of condition F. In (1), a designing agent exploits C-tokens' indication of F. In (2), an organism genetically inherits a system in which C has caused M in the presence of F. In case (3), an organism (or other cognitive being[4]) becomes conditioned so that C-tokens cause M. In each of these cases the resulting type of representational state C has a content we might refer to roughly as *that F* or *that F is present*.

Dretske explains the connection between behaviour-structuring and representation as follows:

> Once C is recruited as a cause of M ---and recruited as a cause of M *because of what it indicates about F* -- C acquires, thereby, the function of indicating F. Hence, C comes to *represent* F. C acquires its semantics, a genuine meaning, at the very moment when a component of its natural meaning (the fact that it indicates F) acquires an explanatory relevance. (Dretske 1988a, p. 84)

Ultimately, this account of representation is supposed to show the relevance of intentional states in virtue of their being representational states. Dretske thus continues by relating his discussion to beliefs:

> This, indeed, is why beliefs are maps *by means of which we steer*. An indicator element (such as C) becomes a representation by having part of *what* it indicates (the fact that it indicates F) promoted to an explanatorily relevant fact about itself. A belief is merely an indicator whose natural meaning has been converted into a form of non-natural meaning by being given a job to do in the explanation of behavior. What you believe is relevant to what you do because beliefs are precisely those internal structures that have acquired control over output, and hence become relevant to the explanation of system behavior, in virtue of what they, when performing satisfactorily, indicate about external conditions. (Dretske 1988a, p. 84)

In the first of these two quotes, it is clear from the contrast that "genuine meaning" refers to non-natural meaning (representation). So the view is that a system S, with an internal state type C, comes to have a representation of F when and only when C comes to cause M in virtue of C H-indicating F. C H-indicating F, once again, means that over some period a system S's C-tokens have all indicated that something is F. C's having a "genuine meaning" is directly bound to its

having a behavioural role (that it acquired because of past indication-relations). As we shall see in Chapter 5, there is substantial question about how this manages to give an explanatory role to belief *qua belief*, which is precisely what Dretske does need to show. Even in this passage, "representation" emerges as a rather puzzling and oddly transparent sort of property. It is sometimes difficult to know whether Dretske sees it as some additional property of a type of state that somehow "emerges" from a Design Solution, or in a more deflationary way, as just coming to "the property of causing M in virtue of H-indicating F".

Now it should be fairly clear that being an indicator of F has itself to be a feature of C's coming to cause M in order for state type C to become an F-representation. This is confirmed in a footnote:

> C will normally indicate a great many things other than F. Its indication of F is, therefore, only "one component" of its natural meaning. Nonetheless, it is this single component that is promoted to *representational* status, to a form of non-natural meaning, because it is C's indication of F, not its indication of (say) G or H, that explains its causing M. Hence, it becomes C's function to indicate F, not G or H. (Dretske 1988a, p.84)

In Dretske's terms, it is C's *indication of* F that explains its coming to cause M. As I would put it, it is C's H-indicating F, or C-tokens having been indicators of F, that explains the behaviour-structuring. In this passage, one can also see how Dretske would propose to deal with issues of the selectivity of intentional states and the so-called "depth problem". Suppose that there are appropriate dependencies between state C and conditions F and G. Perhaps G causes C and F is such that it is present when and only when G is present. Suppose that it is C's H-indicating F (and not, for example, its H-indicating G) which explains C's coming to cause M. So even though C-tokens have indicated (and may continue to indicate) both F and G, F is selectively represented.

The explanatory role of "F-indicating" accounts for the representational depth of (type) C being fixed at F.

Of course, this ability of a Design Solution to distinguish one of several H-indicated conditions requires an explanatory mechanism in virtue of which a particular indicated condition (F) is a cause of the solution (the behaviour structuring) in a way that other such conditions are not. For the engineer's solution, the distinguishing mechanism is the designer's intentions. The engineer connects a detector with type C states to an output mechanism which produces M, choosing the detector precisely because the resulting structure will do M in condition F. So it is because the designer wants an F-detector that the system's C-tokens can be said to have the function of indicating F. For the selectional solution, the distinguishing mechanism is adaptivity. It was (past) C's indicating F (but not their indicating G) which conferred a competitive advantage upon members of a species. For the developmental solution, based as it is on the notion of operant conditioning, the distinguishing mechanism is reinforcement. For some system capable of learning, doing M in the presence of F is reinforcing whereas doing M in the presence of G is not.

The foregoing outline of the Design Problem and its solution suggest the following summary of Dretske's view of representation (as the function to indicate). C represents F if and only if

(1) C H-indicated F,

(2) C comes to cause M,

(3) C would not have come to cause M if C had not H-indicated F.

While the three types of Design Solutions are solutions to the same problem, there are acknowledged differences. The selectional and developmental solutions ((2) and (3)) are distinguished from the engineer's solution in that the engineer's solution is conventional rather than

natural. It is through the knowledge and skill of an intentional agent that a bimetallic strip comes to be part of a switching system that links the activation of a furnace to room temperature. The selectional solution depends on the effects of environmental pressures on populations which have or lack some genetic mutation, while the developmental solution depends on (positive or negative) reinforcement of behaviour. Conditioning of a given subject's behaviour is something that can be managed by an intentional agent, but the intervention of an appropriately knowledgeable engineer is not generally essential to the developmental solution.

On the other hand, the engineer's and developmental solutions are Design Solutions in individual systems or organisms, whereas the selectional solution is a solution over a kind or a population. The connection of my telephone to a particular answering machine, or the connection of a particular set of motion detectors to floodlights are applications of the engineer's solution. My learning, in early childhood, to avoid electrical wall-sockets through the rough experience of having poked my finger in a couple of them is an application of the developmental solution. The result of these processes is a "behaviour structure": activations of a motion detector causing floodlight illuminations or wall-socket sightings causing avoidance movements. The structured processes are processes within particular systems. The structuring does not involve or have any implications for other systems of the same kind.

The selectional solution does involve kinds. If a particular organism S now has an inherent behaviour, one that is genetically encoded, that it does so is accounted for by environmental pressures on the species of which S is a member. At some time in the past history of the species-population, members of that population appeared with the encoding for a "C's causing M in the presence of F" structure. This structure, due to the existing conditions of the environment

(and of population members), conferred some "advantage" on members which had it. This is to say that those organisms which had the C-M structure were able to reproduce more often than the other members of the population. Over generations, the size of population which has the C-M structure increases as the pressures of the environment, by process of elimination, "de-select" organisms in which C-tokens fail to cause M.

To make use of a previously cited example of Dretske's, let us suppose that a current frog's (call this frog Perry) predatory reaction to its sightings of moving black, bug-shaped spots is an inherent piece of behaviour. This would then be an example of the selectional solution. Internal state-type C *of Perry* is not necessarily connected to tongue-movements (i.e. bug-catching movements) in virtue of *its* indicator history (i.e. *Perry's* C-tokens having been indicators). The story which makes it reasonable to identify Perry's C-states as representations of bugs or of food (for example) is a story about *Perry's genetic ancestors*. In past populations of frogs, a mutation in which a tongue-shooting response to a type of visual stimuli was a piece of behaviour that was relatively successful in delivering food to frogs who were genetically wired for the behaviour. They thereby obtained some reproductive advantage over frogs whose genetic package did not include that particular type of predatory behaviour[5]. The relevant features of the environment are such that frogs genetically tuned for the behaviour tended to reproduce more frequently because, being more successful in obtaining food, they tended to live longer. The genetic package becomes more prevalent in the frog population over generations because of the advantage (the difference in reproductive rate). Something roughly like this story is what explains why Perry exhibits the pattern of behaviour in which bug-sightings are met with tongue-dartings. That the behaviour would not have been propagated had the tongue movements not been advantageous in delivering

food (which for frogs happens to be bugs) suggests that had other frogs' C's not been dependent upon the presence of food, Perry's C's would not have caused M. Perry (if he existed) would have had a different genetic history[6]. That the existence of the behaviour structure has this explanation, Dretske contends, means that the type of internal state in question has the function of indicating the presence of food (or perhaps bugs).

Thus, when solved selectionally, the Design Problem is not solved by a particular organism coming to be structured such that C-tokens cause M. Rather the Problem is solved for a population of which S, the current organism, happens to be a member. That S (now) does M in presence of F-like conditions (eg. conditions that cause C) is explained by the development of the species-population of which S is a member. The behaviour pattern of C's causing M may be of no particular advantage to S at all. It may even be a disadvantage[7]. Moreover, what S receives is a (genetic coding for a) type of behaviour. S receives this as part of the genetic package which is transmitted via biological reproduction. The "solution" is literally reproduced: behaviour types are propagated as parts of replicating entities.

In a sense, then, the selectional solution does not explain particularly why Perry does M, or has come to do M, in the presence of F[8]. What it does explain is that most any frog of the current generation is likely to exhibit this behaviour because, by elimination, the genetic package of which the behaviour is a part has been selected for replication. It has been selected for replication because of its reproductive advantage for ancestors of the current generation. Perry, therefore, is of a type such that in most or all moving bug-shaped sightings cause a tongue movement.

Dretske explains this point referring to a species of moth with an inherent internal state C that causes evasive maneuvers. The type of state in question is (has been) typically triggered by

high-pitched bat squeaks. He writes,

> Even if through recent freak of nature (recent enough so that selectional pressures had no time to operate) the occurrence of C in contemporary moths were to signal not the approach of a hungry bat but the arrival of a receptive mate, C would still produce M -- would still produce the same evasive flight maneuvers. What C indicates in *today's* moths has nothing to do with the explanation of what movements it helps to produce. And the fact that tokens of C indicated in remote ancestors the approach of hungry bats does not explain -- at least not causally (developmentally) -- why *this* (or indeed, why *any*) C produces M. Rather, it explains (selectionally) why there are, today, predominantly moths in which C causes M. (Dretske 1988a, p. 93)

The selectional and engineer's solution thus each differ in some important respect from the developmental solution. The engineer's solution, requiring the intervention of an intentional agent is conventional. What the structures in an engineered system represent depends on the intentions of the engineer. If Dretske is to give an account of how cognitive beings, by having intentional states, constitute representational systems, he needs a natural (non-conventional) solution to the Design Problem. The developmental solution is natural at least in the sense that it does not generally require the external intervention of a designing agent. The engineer's and developmental solution are similar, however, in being particularistic. The Design Problem is solved by these solutions for individual representational systems.

The selectional and developmental solutions are similar in that they are both non-conventional solutions. They differ (at least) in that the selectional solution is a solution over kinds. It is not particularistic.

All three of the Design Solutions provide states with representational content, with what Dretske would regard as "the right sort of meaning". The right sort of meaning is a kind of non-natural meaning: functioning to indicate. Nevertheless, Dretske claims that the semantic properties of states that acquire their functions through engineering are causally impotent. So too

are selectional promotions. In order for meaning to be "relevantly engaged" (to use an expression of Dretske's), the state must acquire its function developmentally.

The reason for this claim, in the case of an engineered system seems to be that the development of such a system and the selection of an indicator-function for indicating states of the system is managed entirely extrinsically by an intentional agent. What a system state C comes to represent is a function of the designer's intentions, beliefs, and desires. Thus if C-tokens either indicate or function to indicate to or for anything, one might contend, they do so to or for the designer. Why mechanical, electrical, chemical, or other parts are assembled as they are in a conventional system has only to do with the assembler's interests.

Why is selectionally acquired representation not relevantly engaged? Dretske does spend some time arguing for this[9], but the reasoning is not entirely clear. The idea seems to be that a selectionally acquired indicator-function is not part of the individual history of an organism. A current organism's process of C causing M is not caused by what *that* moth's C-states have indicated; and thus what C-states in that moth have indicated does not explain why its C-states cause M. This seems a bit strong. Surely it amounts to claiming that the selectional solution is not a solution that "promotes" an indicator to an indicator-function. If C's indicating F does not explain in some sense the fact that in any particular moth C causes M, then how could it be true that C would not have come to cause M if C-tokens had not indicated F? If the counterfactual is not true, then C fails to meet the conditions required for it to be an F-representation. Dretske must mean something along the following lines: In selection what a state of type C indicates over the history of a population is part of the causal conditions for the increasing proportion of "C-causing-M" organisms without C-tokens indicating F being directly explanatory or directly structuring in any

particular organism's C-tokens causing M. The main point, then, in denying that selectionally acquired indicator-functions are effective is that they are, in an individual organism, inextricably fixed into a (tropistic) behaviour pattern. They cannot be modified in a particular system regardless of what that individual's C-tokens happen to indicate. In the previously cited moth example, selectionally fixed C-tokens in a current moth would continue to cause evasive movement even if, throughout the life of that moth, C-tokens had all been caused by the approach of a mate and not by a hungry bat. In other words, when a C's causing M process is inherent, the informational relations borne by an individual's C-tokens are powerless to change the probability of future occurrences of C's causing M in that individual. Detection of high-frequency sounds in a current particular moth will continue to cause erratic flight movements come what may. So, Dretske reasons, what that moth's internal states indicate is irrelevant to its behaviour.

Dretske's idea that history of indication in an individual system is required to "engage" meaning is reinforced in his main discussion of this point in *Explaining Behavior*. He says that inherent behaviour control states, by having a biological function, are representational. They are representational "maps" in the sense that they do "say" something about external conditions. The representational content of such states, is not, however, properly engaged and such states are therefore not even candidates for being beliefs. He writes,

> For to qualify as a belief it is not enough to *be* an internal representation (a map) that is among the causes of output, something that helps us steer. *The fact that it is a map*, the fact that it *says* something about external conditions, must be relevantly engaged in the way it steers us through these conditions. What is required, in addition,... is that the structure's indicator properties figure in the explanation of its causal properties, that what it *says* (about external affairs) helps to explain what it *does* (in the production of output). That is what is missing in the case of reflexes, tropisms, and other instinctive behaviors. Meaning, though it is there, is not relevantly *engaged* in the production of output. The system doesn't do what it does, C doesn't cause M, because of what C (or

> anything else) means or indicates about external conditions. Though C has meaning of the relevant kind, this is not meaning it has to or for the animal in which it occurs. That, basically, is why genetically determined behaviors are not explicable in terms of an actor's reasons. That is why they are not *actions*. What (if anything) one wants, believes, and intends is irrelevant to what one does. (Dretske 1988a, pp. 94-95)

My central purpose here is simply to identify Dretske's reasons for claiming that only the developmental solution to the Design Problem yields meaning that is "relevantly engaged". However, it is also quite important as one examines Dretske's comments on this issue that one keep in mind that the Design Solutions pertain to *structuring causes*. The Design Problem is the problem of structuring a system so that it will give an output M in particular kinds of conditions. So when Dretske is referring to the relevance of tokens of a type of state C being indicators of F, the relevance of the semantic property "indicating F" must be (or should be) with respect to state C's being recruited as a cause of M. What is not at issue, for example, is the relevance of a current C-token of a system S being an F-indicator in *triggering* a token of the movement M.

The indicator/representation distinction is very important in Dretske's theory because indicators need "promotions" to become representations. Also, it is one of the central points of the distinction that a state that represents something as F need not be an F-indicator. Otherwise, misrepresentation would be impossible. It is certainly worth noting, therefore, that there are opportunities for confusion of "representation" and "indication". For example, in the foregoing quote "belief" is identified as a representation, and representation as a "map" (by means of which we steer). In the second sentence of the quoted passage a state's being a map is spoken of as its *saying* something about external conditions. It would be reasonable in this context to think of representations as states that say something about external conditions. When the fact that they say something is "relevantly engaged", they are beliefs (or belief-like states[10]). What is it for such

states to be relevantly engaged? Dretske describes this in the foregoing passage by saying that "...the structure's indicator properties figure in the explanation of its causal properties, that what it *says* (about external affairs) explains what it *does* (in the production of output)". Given the idea of solutions to the Design Problem, the natural reading of this is as a reference (for a particular system) to an internal state's tokens having been indicators figuring in the explanation of the output movement M that state comes to cause. In that case, the reference in the previous sentence to what an internal state says about external affairs is a reference to what its tokens have indicated (not what they represent). So within two sentences, there are uses of "what C says about external conditions" which refer both to what C represents and to what C indicates.

To return now to the central issue, Dretske's view about what is lacking in the selectional solution is that tropistically connected internal states cannot in an individual system affect future behaviour. From this view about what is lacking in the selectional solution, it should be possible to sketch a view of what makes the semantic (indicator) properties of some internal state of an organism relevant in the developmental solution. C-token's having indicated a condition F is not "relevantly engaged" on selection, in Dretske's view, because with respect to a particular system (organism) whether or not its state C has ever been dependent on F is irrelevant to the effects that its C-tokens has on output (M). The developmental solution is the solution on which indication is "relevantly engaged". This must therefore mean that, with respect to a particular system in which C's causing M is a developmental solution, its C-tokens having indicated F must be part of the explanation of C's causing M.

The developmental solution is a solution by operant learning. Dretske's view is that developmentally acquired behaviour is conditioned behaviour. Thus C-tokens having indicated

F, or (as I have put it) C's H-indicating F, has to be part of the explanation of the conditioning process by which type C comes to cause M. The idea that the property of being an indicator must be among the causes of C's coming to cause M is an important point to establish. In Dretske's theory, this and only this role as a structuring cause of behaviour is what gives the semantic properties of a representational system an explanatory role.

Let us revisit the example of Perry the frog and imagine (contrary to what is probable for frogs) that Perry *learns* to catch bugs with a flick of his tongue. Through some exploratory process, Perry manages to catch a bug by darting his tongue at a dark, moving speck that also causes a token of an internal state C. Catching a bug[11] is reinforcing. C-tokens indicate bugs. Repeated trials of the predatory behaviour "C's causing of M", where M is the tongue movement are thus reinforced. So, the Dretskean story goes, Perry continues to exhibit the behaviour because of C's H-indicating bugs reinforces tongue-darting. That Perry's internal C-states have indicated bugs, and that C's causing M (tongue-dartings) would not have been reinforcing without C's relation to bugs is what accounts for the continuance of C's causing M.

The mechanism for the promotion of developmentally acquired representations on this view is reinforcement. Dretske writes,

> The kind of learning we are talking about is a special form of *operant* or *instrumental* learning, a kind of learning that is sometimes called *discrimination* learning. One learns to identify F, or at least to distinguish (discriminate) F from other conditions, by having particular responses to F (or particular responses *in* condition F) rewarded in some special way. (Dretske 1988a, p. 99)

Dretske goes on to note that such discrimination learning requires two things[12]. In his order of discussion, the second is that an organism needs some specific sensitivity to condition F: the perceptual machinery in virtue of which it can distinguish F from other conditions. The first

Dretske describes as follows:

> First, there is Thorndike's Law of Effect, which tells us that successful behavior tends to be repeated (Rachlin 1976, pp. 228-235). More technically, a reward (alternatively, a positive reinforcement) increase the probability that the response that generates it (or with which it co-occurs) will occur again in the same circumstances. (Dretske 1988a, p. 99)

While the selectional and developmental solutions differ in some important respects, it is clear that Dretske means there to be parallels. In particular, the selectional solution is a model which shows that functions can be acquired by non-conventional means. In his most recent book, Dretske comments,

> I assume that there are naturally acquired functions and, thus, natural representations. I do not argue for this; I assume it. This view, I know, is not universally accepted (Dennett 1987, especially pp. 287-321 and Searle 1992, p. 52, for example, deny it). I return to this point in chapter 5, but for the present I follow the lead of Wright (1983, 1987), Kitcher (1993), Godfrey-Smith (1994), Millikan (1984), Neander (1991a, 1991b), Papineau (1993), Bennett (1976) and many others in supposing that bodily organs and mechanisms can, in the relevant sense, be designed to do a certain job -- and, thus, have the function of doing it -- without being designed by anyone to do it. Philip Kitcher (1993, p. 380) puts it thus: "one of Darwin's important discoveries is that we can think of design without a designer." (Dretske 1995, p. 7)

The context makes it clear that what Dretske has in mind here is biological functions which are the product of selectional pressures. However, in an associated footnote, his view that the developmental solution yields natural functions is made explicit:

> I think that natural selection is not the only source of natural functions. Learning is also a source. This was the basis of my (Dretske 1988a) account of belief. I return to this point shortly. (Dretske 1995, p. 169. The 1988a reference is a reference to *Explaining Behavior*.)

There is an analogy being drawn between selection and development (operant learning). The selectional solution to the Design Problem allows for (mutated) behaviour patterns to arise and to be reproduced and propagated (when successful). The developmental solution allows for

behaviours to arise and (according to Thorndike's Law) to tend to be repeated (when rewarded).

It is important to be clear that the analogy is only a rough one with some crucial points of difference. Rewarded behaviour is reproduced in *some* sense, but not in the sense that adaptive (inherent) behaviour is reproduced. The selectional solution propagates a *type* of behaviour over generations of a species population. It does so via biological reproduction. The beings in which the behaviour-structure is instantiated literally replicate that structure by passing on the genetic package responsible for the type of behaviour in question. Systems reproduce more systems. The spread of the behaviour in a population is the result of the "eliminative" effects of the environment. Some genetic packages are passed on less frequently than others because of the effects of the environment on those organisms' ability to reproduce. The selectional solution solves the Design Problem for a species population because of the differential effects of an environment on the reproductive rates of species-members with different sets of inherited traits.

In operant conditioning, what is "reproduced" is not a type of behaviour, or a physical structure which is responsible for a type of behaviour. What are multiply produced are instances of some type of behaviour. Unlike inherent behaviour traits which are passed on and replicated as parts of a genetic structure, behaviour instances are not part of such a replicating structure. Repetition because of reinforcement is obviously not biological reproduction. Because selection is a solution for a kind, there is no sense in wondering with respect to some particular organism what the mechanism was by which it came to do M in the presence of F. It does make sense to wonder about this in connection with the developmental solution. As Dretske is no doubt aware, an appeal to Thorndike's Law of Effect does not explain how or why rewarded behaviour tends to be repeated (or how it is rewarded). That is, it does not explain the mechanism by which

reinforcement solves the Design Problem. It is important to bear this in mind, for one might well suspect that the operation of reinforcement, part of the machinery which is supposed to show how mental states with active semantic properties are possible, requires mental states that are representational in order to work. This is a point I will return to in some detail in chapter 6.

One common thread in all of Dretske's solutions to the Design Problem is the idea that representation is a functional notion and that representation is an extrinsic relation. Dretske is in a strong sense an "externalist" about meaning. A representation's content depends in an essential way on a relation to an external environment, and what the content of an intentional representational state is predominantly determined by the external term of the relation. For example, in a response to Rudder Baker, he writes "there is the fact that meaning is an extrinsic, in particular an *historical*, property". He continues as follows:

> The fact that meaning is an historical property reveals why meanings retain their causal efficacy even when they are false, why false beliefs are as effective as true beliefs in explaining behaviour. Meanings are effective even when false, even when Cn fails to indicate what it is its function to indicate, because Cn's meaning F is, in part, constituted by past tokens of this type *doing* (i.e., indicating F) what it is (as a result) Cn's function to do. (Dretske 1991a, p.117)

To summarize, on Dretske's view representation is intimately linked to behaviour through the Design Problem. A representational state is a functional state, the content of which is determined by the external condition in virtue of which the state comes to be recruited as a component of behaviour. On the engineer's solution, what the state C of a system acquires is determined conventionally by designer intentions. The two solutions (selectional and developmental) for type III representational systems are natural (non-conventional) solutions. The content of a recruited C-state on the selectional solution is determined by the indicated condition

of the environment which makes C's causing a movement M adaptive. The content of a recruited C-state on the developmental solution is determined by the indicated condition which makes C's causing M reinforcing (or in which C's causing M is reinforcing). Indication is, for Dretske, a crucial ingredient for any of these solutions, for it must be the case (for C to represent something as F) that C would not have been "recruited" as a cause of M had C not H-indicated condition F.

NOTES

1. The notion of "movement" here is to be loosely construed as per the earlier discussion of Dretskean behaviour.

2. Movement M, for example, could be a movement by which S tends to obtain F. A movement toward a certain stimulus for some organism (one type of representational system) may tend to put that organism in the proximity of a potential mate, for example.

3. Dretske here characterizes the design problem as the problem of getting a system to do (movement) M when and only when F *exists*. Elsewhere, he characterizes the concern of getting a system to do M when and only when it *detects* F (See, for example Dretske 1988a, p.98). The two characterizations are obviously different but which one Dretske is committed to will not affect the points I plan to make.

4. I am merely leaving open the possibility of non-biological systems being capable of learning.

5. No doubt the details of this story are greatly over-simplified. It is unlikely that there was a population of frogs divided between those which shot their tongues at bug-shapes and those which did not, for example. Such an over-simplified story is good enough for my purposes, however, which is merely to illustrate the features of Dretske's selectional solution. That a complex behaviour such as predation is the likely result of a long series of subtle mutations, that the competition between Perry's ancestors and other frogs was likely between frogs with only slightly different responses to moving bug-shapes are details of biology that do not affect the general idea of the selectional solution.

6. I want to avoid a distracting debate about Perry-identity. It does not matter to my purposes whether one describes the counter-factual situation as one in which the current frog is a different frog (with a different genetic history) or is Perry (with a different genetic history).

7. I do not mean to suggest this a criticism of Dretske or as an oversight on his part. He is well aware of this feature of the selectional solution. Using the example of a species of moth in which sounds associated with the approach of predatory bats have come to cause evasive movements, Dretske remarks: "Even if, through a recent freak of nature... the occurrence of C in contemporary

moths were to signal mot the approach of a hungry bat but the arrival of a receptive mate, C would still produce M -- would still produce the same evasive maneuvers." (Dretske 1988a, p.93)

8. It is probably more appropriate to say "F-like conditions", for Perry's internal C-tokens, as has been noted, need never have been dependent on the occurrence of an actual F (an actual bug).

9. In recent conversation (Fall 1996), Dretske suggested that he may be rethinking this claim about selectionally acquired states.

10. Dretske qualifies his views at various points so as to be neutral on the question of whether even simple organisms are capable of actual belief-states.

11. "Food" might be a more reasonable choice here than bugs. The general illustration works just as well with either "bug" or "food" as the indicated condition F, however.

12. He later acknowledges additionally that desires or other motivational states are required to make some particular outcome one that is rewarding. Dretske's discussion of motivational states is the topic of the following two sections.

CHAPTER FOUR

Motivational States

a. Introduction

An analysis of behaviour and how behaviour is structured by the developmental solution is incomplete without inclusion of motivational states. An explanation of a system's (organism's) execution of a piece of conditioned behaviour in terms of reasons will surely appeal to desires (or some variant) as well as beliefs (or some variant[1]). So any complete, general account of how the developmental solution structures behaviour has to make some reference to motivational states. An organism with informational states will not acquire any new behaviour unless it is motivated, unless it has receptivities[2]. These are points which Dretske notes in incorporating desires into his account. The bulk of his theory concerning intentional states is given over to the development of a view of doxastic ones (beliefs or belief-like states). He does, however, include a view on motivational intentional states.

The discussion of motivational states in *Explaining Behavior* is relatively compressed, difficult to follow and, by Dretske's own later admission, a section of his work with which he had some difficulty. In a symposium response to Michael Bratman's critique of his account of motivational states, Dretske writes,

> Michael Bratman focuses on an issue that caused me hours of grief when I was writing the book: am I giving a theory about how desires that explain behavior get their (intentional) object or am I, instead, giving a theory about how desires with intentional objects explain behavior? Or both? (Dretske 1990, p. 834)

He goes on to acknowledge that there is "some ambiguity" in the text (*Explaining Behavior*) and that the view he opts for there is not the view that he should have endorsed. The two

alternatives which Bratman highlights, and Dretske's views on these, will be dealt with in detail shortly.

In this section, I want to place desires in his framework of explanations of behaviour, highlight the Dretskean view of the intentionality of desires, and examine how motivational states fit within the developmental/selectional distinction. Dretske appears to commit to the view that motivational states are intentional without being representational, although his commitment to this view is rather weak and his reasons are not clear. Also, he has some trouble over the account of basic sorts of motivations and whether these are acquired by something akin to development or to selection.

The addition of motivational states initially involves two complications in the Dretskean picture of behaviour as it has so far been elaborated. So far "C's" have been conceived simply as indicator-states that can acquire representational functions. They thereby become beliefs or some "doxastic variant", as Dretske puts it. An internal state C is more properly thought of on the elaborated account as a complex of information and motivation. Dretske uses "B" (suggestive of "belief") for the doxastic component of C, and "D" (suggestive of "desire") for the motivational component. The two components, B and D, are "necessary parts of a sufficient condition" for the production of M in the presence of F (Dretske 1988a, p.113).

The second complication is the addition of "R", which stands for "reinforcing outcome". This is some outcome in relation to F which changes the probability of further instances of C-complexes causing a movement M. Generally, Dretske conceives of R's as conditions which are in some way beneficial to an organism, though he reminds us that negative reinforcement can also affect the probability of (future) C-tokens causing M-tokens[3].

The general approach taken with respect to desires is closely analogous with the account of doxastic intentional states. In fact, in the previously quoted reply to critics, Dretske makes it clear that he sees very strong parallels between beliefs and desires. both in terms of how they come to have intentional properties and in the function that these states which come to be beliefs and desires have in a structuring explanation of behaviour[4]. He writes,

> By understanding that the behavior beliefs and desires are invoked to explain is a causal *process*, we make such behavior accessible to both triggering and structuring (causal) explanations. By understanding that both belief and desire derive their content -- what is believed and desired -- from the learning process in which such behavior is structured, one makes the content of our internal states relevant to the explanation of this behavior. This is true of both belief *and* desire. (Dretske 1990, pp. 834-835)

A little later the (postulated) role of beliefs and desires is described as a "parallelism".

One ought to be a little suspicious at efforts to draw too many lessons from belief in accounting for desire, however. Motivational states do not arise from indication relations. Moreover, desires (or desire-like states) *qua* desires seem to be intimately connected with a condition's being reinforcing in a way that beliefs are not[5]. Commonsensically, nothing is a reward unless a motivation is satisfied by obtaining it, and the developmental solution (applications using positive reinforcement at least) requires some outcome to be a rewarding one. Dretske's proposed parallel with belief, if drawn very closely, would have a state's becoming a desire for some outcome as a product of a developmental solution. On the contrary, one would think that it is that some state is a desire for some outcome which makes that outcome reinforcing and thus makes conditioned behaviour possible. This appears to suggest that *some* desire is required for a developmental solution. Let us try to sort out this issue by laying out the proposed account of desires.

Dretske's outline of his more complete theoretical structure is as follows:

> Think of an organism learning to do something in a specific set of conditions: it learns to produce M in conditions R by having rewards R for producing M contingent on M's production *in* F. As we saw in the preceding chapter, such a process will result in the recruitment of an F-indicator as an internal cause of M. We have relabeled this internal indicator B. So, if learning is successful, B is enlisted as a partial cause of M. Since D is the internal state on which R's effectiveness as a reinforcer depends, successful learning also requires the animal to occupy state D when movements M are produced. R will not be effective in promoting the production of M unless the organism is in *both* state B and state D. Since M doesn't lead to R except in F, and since R isn't reinforcing unless D, learning requires that *both* F and D exist for the production of M. Since this is so, R will recruit, as a cause of M, both B and D. Or, if you please, the occurrence of R will recruit B as a cause of M only if B is accompanied by D. The only way to arrange things so that M is produced when both B and D exist, but (in the interest of economy of effort) not otherwise (i.e., when either B or D exists alone), is to make B and D necessary parts of a sufficient condition for M. Hence, this kind of learning results in the recruitment of B and D as *partial* or *contributory* causes of M. (Dretske 1988a, pp.112-113)

There are a number of important points about the elaborated model: what the individual symbols, B, F, M, R, and D denote in specific cases, what the model shows about Dretske's account of the structure of reasons-explanation, and how Dretske means to promote the basic account as showing how higher order intentional content could come about. These are issues to which I will return. The first order of business is to outline the account of motivational states itself.

b. Drives and Desires

On Dretske's view, motivational states are divided into three broad categories: drives, pure desires and derived desires. There is some difficulty, however, in this account on how exactly the line between drives and desires is to be drawn. For as I have already noted, Dretske does indicate a shift in view on motivational states between what he endorses in *Explaining Behavior* and what he accepts in responding to Bratman. The change of view may well represent a change in how

drives are conceived and in how they are distinguished from desires. For a start, I will explicate the view developed in *Explaining Behavior* on how drives and desires seem to be differentiated[6].

Here is a preliminary effort at delineation: Drives are not states with intentional properties (if they have any) that could contribute to an explanation of a piece of behaviour in terms of reasons. In fact, Dretske takes the view that drives do not have intentional properties. Examples that Dretske employs suggest that these are to be thought of as partial causes of inherited behaviour patterns. The patterns are invariably executed under certain conditions of organism stimulation regardless of the outcome of the behaviour. The reasoning is that, as the whole behaviour pattern is the product of some selectional process, a drive whose activation causes an instance of the behaviour lacks an intentional dimension of a desire. A drive is not *for* some outcome. It is likely the case that a behaviour pattern is selected because it happens to be beneficial (members of a species wired to execute the behaviour obtain some evolutionary advantage). However, an individual organism so wired cannot avoid the behaviour regardless of what it brings about.

Dretske's main example to illustrate drives is a description of the feeding behaviour of the blowfly (Dretske 1988a, pp. 123-124). Blowflies have acquired chemical detectors in their knees which are sensitive to sugar water. When the appropriate receptivity-state (a state which is governed by a nervous connection between their stomach and brain) is activated and the blowfly makes perceptual contact with sugar water, it will automatically extend a probiscis to take up the liquid. When and for how long the probiscis is extended is governed partially by a receptivity controlled by signals sent from the stomach to the brain. The drive in this example is merely a deficit-state. When active, it causes probiscis extension given a certain set of stimulus conditions.

There is an (adaptive) selectional explanation of the development of the blowfly's feeding

system available. In blowfly environments, the sugars typically found in water are sustaining, so blowflies equipped with the feeding system in question can take advantage of a source of food that allows them to survive. The feeding system is proliferated through reproduction.

However, a blowfly equipped with this system will extend its probiscis under the relevant perceptual and motivational circumstances come what may. If the drive is continually activated, the fly will continue to eat, even to death. If a blowfly is placed in an environment in which sugar-water is not sustaining, that blowfly will nevertheless continue to execute the behaviour despite its continual lack of success (success being the obtaining of sustenance) for the organism.

The blowfly feeding system is thus blind to its success in obtaining what it was selected to obtain. So, Dretske says, such a drive-governed system is not governed by a motivational state that is *for* something. The invariant behaviour pattern was selected (by natural selection) because of the effect the behaviour pattern has tended to have on blowfly populations over generations. Because it is invariant the drive which governs the probiscis movements is a drive *toward* a certain result, but it is not *for* that result:

> If a drive produces movements M that typically yield a beneficial result R, and the behavior (d's causing M) was selected for because it tended to yield R, then we say that d is a drive *toward* R. But just as one can drive *toward* Chicago while having no intention or desire to go to Chicago, so a drive toward R is merely one that typically causes movements that result in R and which, because R is beneficial, may have been selected for this job because of this fact. But the fact that d was selected for this job (producing M) by evolutionary processes does not mean that the behaviour is goal-directed. Behavior in which drives rather than desires figure may exhibit some of the external marks of goal-direction, but it is, at best, behavior directed *toward* a goal, not *by* a goal. (Dretske 1988a, p. 124)

From the example and accompanying discussion that Dretske provides, there would appear to be two important identifying features of drives. First, they are receptivities which are

selectionally wired into behaviour patterns. Second, they may be identified as directing behaviour toward a goal, but it is not in virtue of being *for* some goal that they have a movement-causing role[7]. They are simple deficit-states that switch on (and switch off) an organism's propensity to execute some inherent behaviour, but these deficit states' bringing about a beneficial outcome (or not) has no effect on the probability of the behaviour's being repeated by the organism.

By contrast, then, desires are receptivities available[8] for the structuring of behaviour. They do have the intentional property of "*for*-ness" (being *for* some outcome). Desires have a role to play in behaviour in virtue of their being for a goal. That some state is a receptivity which, if satisfied, would or could alter the probability of a given organism's repeating a behaviour seems to be the feature which marks a receptivity as intentional, being *for*, and not merely tending toward, an outcome R.

The explication given so far is quite preliminary. More certainly needs to be said about desires. It is difficult to get exactly clear on the desire/drive distinction in order to understand what Dretske takes the intentional component of desires to be, and in order to understand how he conceives of them in relation to behaviour. The question Dretske eventually must confront is whether tokens of a type of receptivity state must have actually been satisfied by some outcome R in order for the receptivity to be *for* R. My explication will lead toward that issue.

(i) Desires

Among desires themselves Dretske distinguishes between what he calls "pure desires" and "derived desires". The former are the more basic of the two. The distinction is most easily outlined by analogy with the distinction in ethics between intrinsic and instrumental goods. The former is that which is a good for its own sake, the latter is that which is a good for some further benefit it

brings about. A pure desire has an intrinsic object: a desire *for* R is a pure desire if it is a desire *for* R as such, or *simpliciter*. "Derived desire" is short for "cognitively derived desire". This is a desire which is derived from other cognitive states: desires for other outcomes and beliefs about how to achieve them. If an organism desires R because it believes that R will yield R' and the organism wants R', the desire for R is a derived desire. If I want a glass of water and believe that by turning a tap I can get this, then a desire on my part to turn the tap on is a desire derived from prior cognitive states. Dretske explains:

> The use of letter D, suggesting desire, to designate these receptive states is deliberate. Such states function as *motivational* states. They are what I shall call *pure* desires, and they are desires for whatever condition or outcome they make the organism receptive to. There are as many different (pure) desires as there are distinguishable states of receptivity (i.e., states of receptivity for different things). Other desires -- what I shall call (cognitively) *derived* desires -- are generated by beliefs about what will secure the objects of pure (and other derived) desires. Without pure desires, though, there would be no desire at all, and hence no motivation, no purpose, no behavior explicable in terms of an agent's *reasons*. (Dretske 1988a, pp.110-111)

He goes on to note that discussion of derived desires is to be left aside. It is worth bearing in mind how Dretske has here distinguished derived desires. These are desires derived cognitively (from other desires and beliefs about what will bring about their satisfaction).

There is one immediately puzzling point about Dretskean desires. He initially appears to claim that while desires are intentional states, they are not representational. Dretske describes the intentional property possessed by desires as "being *for* something" or as "being *for* some outcome R". It is not clear how much weight is to be placed on what Dretske commitment to his claim that a desire's intentional "*for*-ness" is a non-representational property. He writes, "Desires, though they are not representational states, do have an object, something they are a desire *for*, that gives them special status." (Dretske 1988a, p. 127) However, this comment is immediately qualified in a

footnote. Referring to the claim that desires are not representational, Dretske comments as follows:

> Not, at least on the face of it. However, Stampe (1987) describes desires as something like perceptions (and hence representations) of needs. (Dretske 1988a, p. 127)

It is eventually Dretske's view that, like a belief, a (pure) desire must involve some relation to an environmental condition[9]. A desire D cannot be for R without there having been a past occurrence of R's having been obtained (thereby relieving bouts of D). On Dretske's theory, brushes with instances of some condition F, given satisfaction of special indication conditions, gives an internal (belief) state an F-relevant content. It is not clear to me why one would not want to say the same kind of thing about intentional motivational states. Receptivity-reducing brushes with R (possibly in cases where there are some further restricting conditions to be satisfied) yields a desire-state D which has some *content* relevant to R, and hence is representational.

Here is one plausible explanation of this curious aspect of Dretske's discussion. To be representational, a mental state has to have H-indicated (i.e. had a history of its tokens indicating) some condition. It has to be a state past tokens of which were information-bearing, and which came to be part of a behaviour in virtue of that history. Receptivity states are not in this way information-bearing. At least tokens of some receptivity-state are not information-bearing in relation to that which they are *for*. A token of state D does not indicate an outcome R (or a condition F). How could a D-token (an as-yet unsatisfied receptivity) bear a physical dependency on the outcome R that would satisfy it, or the condition F in which R is likely to be obtainable?

Nevertheless, there is a way of construing receptivities as kinds of indicators. Taking the suggestion in Dretske's footnote reference to Stampe as a guide, one could think of these as indicators of conditions of the organism itself. My current receptivity state (which I happen to

recognize as being of a type that would be satisfied by food) is an indicator of my physiological condition (being short of nutrients), for example. Of course taking this approach is somewhat inconvenient from a Dretskean perspective, since the proposed case of indication does not involve a relation between an organism's internal state and some external condition of its environment.

In a 1990 symposium, Dretske responds to a contribution by Stampe on desires[10]. He does not, however, avail himself of the opportunity to clarify his seeming view that desires are non-representational. Nor does he express any particular criticism of the view of desires as perceptions of needs. Perhaps the answer is to read Dretske as saying that if one conceives of desires as perceptions of needs, then they are representational states, and that he does not object to conceiving of desires in this way. However, one is then left to wonder why Dretske should shroud the account in mystery by suggesting in the first place that desires are not representational.

The main features of Dretske's views of motivations are usefully summarized near the end of his explication of the intentionality of desire (Dretske 1988a, pp. 130-131). My paraphrase of these points is as follows:

(1) Desires can cause movements which do not bring about what the desire is for,

(2) A desire can be satisfied by something that is not its object,

(3) A desire for A (or A's) may not be a desire for B (or B's) even if A is B (or A's are always B's),

(4) Potential objects of an organism's desires are limited by the organism's discriminative capacities.

The opacity of desire intended in point (3) is of a familiar kind. A subject S may have a desire for R. Bringing about a given R may have some additional consequence. It does not follow

that S wants this additional consequence. For example, I may want to join the army without wanting to get a haircut or scrub toilets. This is possible in Dretske's view even if there is no possible world in which joining the army does not entail scrubbing toilets.

The claim that instances of a type of desire can be desires for A and not desires for B even if all A's are B's requires some view about the determination of the specific object of desires. This is an area of Dretske's theory which is not without its problems. Dretske eventually settles on a combined mechanism for determination of a desire's "*for*-ness". The view in *Explaining Behavior* is ambiguous on this point, but the view Dretske later opts for is that a desire is *for* A if past instances of that type of receptivity state have been satisfied by A (thus reinforcing a certain kind of behaviour), and the system S has discriminative capacities sufficiently fine-grained to discriminate A's as such. In explicating desires, Dretske sometimes uses an example of his friend's rabbit PJ learning to lick the spout of a water bottle (Dretske 1988a, p. 128). One could use PJ the rabbit to illustrate the role of discriminative capacities as follows: PJ cannot have a desire for Nanton Spring Water if the rabbit cannot discriminate between Nanton Spring Water and ordinary tap water.

Point (1) refers to the property of desires that corresponds to a false belief's being able (on a given occasion) to explain a behaviour. PJ's poking at the spout of an empty water bottle can be explicable in terms of a desire for water. On the occasion in question, however, PJ's desire goes unsatisfied.

Point (2) suggests that a given desire can be alternatively satisfied. PJ's desire may be relieved by carrot juice without that desire being a desire for carrot juice. In connection with derived desires, this is an obvious point. I may have a certain receptivity and a belief about what

will satisfy it. I think that I want sparkling apple juice. Someone slips me a glass of Seven Up and I discover that I am just as happy with that. The object of a derived desire is fixed by other desires and beliefs about what will be satisfying. What will in fact satisfy a given receptivity by no means need be limited to that which the agent believes will satisfy it. In some cases, it turns out that our beliefs from which a derived desire is derived are mistaken. The condition one thought one wanted does not satisfy, some hitherto unwanted condition *does*. Archie, as a result of other cognitive states, derives the desire to take Veronica to the prom, but cannot because she is already going with Reggie. Archie settles for Betty (who asks him) and has a marvellous time while realizing that, had he gone with Veronica he would have been miserable. Such a situation is made possible because a derived desire inherits its content from other intentional states. These other states (any relevant beliefs particularly) may fail in various ways to capture the conditions with which Archie will be satisfied.

It is less obvious that pure desires could similarly be *for* one condition and satisfied by another. Suppose that a pure desire of PJ's is satisfied by carrot juice. Could that desire nevertheless be a desire for water and not carrot juice? The answer turns on the issue of the mechanism by which ground-level (pure) desires acquire their content. If a pure desire's object is simply whatever *would* satisfy it, then it cannot be *for* one thing, yet satisfied by something else. If, on the other hand, a pure desire's object is, say, some environmental condition that has satisfied (past instances) of that desire, then a pure desire could be satisfied by something other than that which it is *for*. Dretske (as Bratman notes) appears to entertain both of these views, although he eventually settles on a version of the latter. This is an issue that will dealt with in more detail directly.

(ii) The Drive/Desire Distinction

Desire is essential in Dretske's view to explanations of behaviour in terms of reasons, so must be in some way essential to behaviour-structuring. Pure desires[11] for things are the bedrock of a system of intentional motivations. On Dretske's classification scheme, drives, having no semantic properties (or at the very least no effective ones), cannot contribute to an explanation of behaviour in terms of such properties Derived desires depend on prior intentional states. Dretske allows as well that with respect to simpler organisms, one may be reluctant to use the term "desire". The general points Dretske makes applies to any system with motivations or receptivities, whatever one wants to call such states.

In his critique of the account of desires in *Explaining Behavior*, Bratman suggests that Dretske is variously endorsing two different theories. He comments:

> Let us focus on what Dretske calls "pure" desires. A desire for X is a pure desire, roughly, if its ability to explain action does not depend on beliefs about the further impacts of X. Dretske typically treats the desire for water as pure in this way, so let us follow him here. We need to ask two questions: First, what makes my desire to drink water a desire to *drink water*? What gives this pure desire its content? Second, how can the content of my desire contribute to causal explanation of action? The problem is that Dretske suggests two rather different answers to the first question. (Bratman 1990, p. 797)

Bratman refers to the two answers to the question of how pure desires get their content as Theory I and Theory II:

> In the case of belief the history of the belief-state plays a double role: It helps to constitute the content of the belief by way of shaping what it is the function of that state to indicate. And it also is the source of the relevance of that content to the causal explanation of action. In the case of pure desire, in contrast, Dretske seems ambivalent between two ways of seeing the relevance of the individual's learning history. On the first theory -- call it theory I -- what the desire is *for* is not an historical matter: the content of the desire is determined by its present power to make certain things reinforcing. But we must still appeal to the individual agent's history when we try to

> account for the explanatory relevance of what is desired. What is desired can help to explain present action only when there is a history of learning in which what the desire was for played an appropriate role in recruiting the desire as a cause of certain bodily movements and/or upshots. On the second theory -- call it theory II -- what the desire is for is determined by the same learning history that is, on the theory, the source of explanatory relevance. (Bratman 1990, p. 798)

The difference between Theory I and Theory II is that it is only on Theory II that receptivity states' coming to be *for* some condition parallels the account of information-carrying states' acquisition of belief-contents. A "genuine" desire *for* R is a motivational state which can explain behaviour in virtue of what it is for. Such a desire is the product of a developmental solution. By contrast, Theory I does not require a developmental solution for desires.

Bratman's identification of two questions is useful here. Question (1) is the question of how desires are supposed to figure in the explanation of a behaviour, theories I and II are substantially the same. It must be the case, even on Theory I, that a behaviour pattern (B+D)'s causing M (C's causing M) has been structured in an organism, and that behaviour would not have been established (reinforced) if at least some past instances of D had not been satisfied by a given outcome R (that is relevant to condition F).

Instances of PJ's continued licking of this spout (instances of a type of behaviour) is explained partially by past receptivities (thirst-instances) being satisfied by movements with respect to the water bottle spout. These movements delivered water. In an associated footnote, Dretske writes,

> When we are dealing with simple desires, those that are *not* cognitively derived, there *must have* been past tokens of R in order for the present desire to explain behavior. PJ may never get water, but if he licks the spout because he wants water he must, sometime in the past, have received water for behaving in this way. If PJ never received water, then the receipt of water could not possibly explain D's (an internal state that makes water reinforcing) recruitment as a cause of M -- could not possibly

explain, therefore, current behavior (D's causing M). (Dretske 1988a, p. 128)

It is important to remain clear about what is at issue in this passage. It concerns a D-state's role in behaviour-structuring and does not concern the fixing of that state's intentional properties. Another point worth drawing attention to is the conditional claim in the passage that "[i]f PJ never received water, then the receipt of water could not possibly explain D's... recruitment as a cause of M..." What the passage says is that the obtaining of a specific environmental outcome R cannot explain S's acquisition of a behaviour if S has never previously obtained R. This is a straightforward logical point: That S has previously obtained R cannot explain S's performing a behaviour if it is not the case that S has previously obtained R. A less straightforward inference one might be inclined to draw from this is that S's having a desire to obtain R cannot explain S's performing a behaviour if it is not the case that S has previously obtained R. Of course Dretske does not say this. This second claim is false in relation to derived desires and at least questionable in relation to pure ones.

Question (2) is the question of what gives a desire its content. It is here Theories I and II are importantly different. Theory II is that a motivational state D is a desire for R if R has normally been the outcome in virtue of which D has been a contributor to (B+D)'s causing M. Past D-tokens were satisfied by outcome R, and D would not have been "recruited" as a partial cause of C's causing M (that is, (B+D)'s causing M). A developmental solution has to come about, it seems, before tokens of the type D become desires for R. If, prior to a developmental solution, D is not a motivational state that is *for* R, it is certainly worth wondering what kind of thing it is such that D both (1) happens to be satisfied when R is obtained, and (2) changes the probability of a movement being repeated in some condition F.

As an example of the possible commitment to Theory II, Dretske states that what a desire is for is determined "...by the sorts of results it *tended* to produce in the animal's past experience" (Dretske 1988a. p. 129). Again using PJ the rabbit, he continues,

> If we now, after learning, put beer in PJ's bottle, the fact that licking the spout now tends to produce nothing but beer, is irrelevant to understanding why PJ licks the spout. He licks it because he wants water even if licking no longer tends to get him *water*. And even if PJ likes the beer he gets, even if it quenches the thirst that motivated him to lick, he still licked it because he wanted water. The rabbit may, *hereafter*, lick in order to get beer. He may, that is, develop new, more discriminating, desires. But until that happens, the desire that explains the behavior is a desire *for* whatever past result figured as a structuring cause of the behavior. (Dretske 1988a, p. 129)

The last sentence is a fairly clear statement of Theory II: an outcome R that happened to satisfy a receptivity D, such that a behaviour with D as a partial cause came to be reinforced, is what D comes to be a "desire *for*". Even here, though, there are possible stronger and weaker readings of Theory II. One reading is a kind of "baptismal" view which states that a receptivity must have actually been satisfied by some outcome in order for that receptivity to be a desire for that outcome. This is suggested by Dretske's comment about PJ and the beer: PJ may have a desire for beer after having obtained some by licking the spout (but not before). So a minimum necessary condition for having a desire for R is that one has at some past time had a receptivity satisfied by R.

The stronger view is that a mere satisfying encounter with R is not enough. A piece of one's behaviour must be modified (or an existing behaviour must be sustained) by encounters with R. Then and only then could one have a desire for R. So if PJ's water is systematically replaced with beer, and PJ continues over some period of time to continue to go to the spout, then PJ has acquired a desire for beer. Alternatively, if after being "introduced" to beer, PJ is sometimes offered a

choice between beer in a dish and water in his water bottle and he acquires the behaviour of (sometimes) going to the dish for a lick of beer, then PJ has acquired a desire for beer. The stronger view is the view which would parallel the Dretskean account of belief. The "for-ness" of a desire is the product of a developmental solution to the Design Problem. It is this stronger view that both Bratman and Dretske seem to have in mind by "Theory II"[12].

Theory I is a weaker view than either interpretation of Theory II. A current token of D is a desire for R if obtaining R would increase the probability of recurrences of the process which brought about the obtaining of R. No claim is made about state D acquiring its status as a desire for R requiring that at some prior time an instance of that type of motivational state was in fact satisfied by (obtaining) R[13]. It need only be the case that the current D-token would be satisfied by R. This is the general view that Dretske claims he (mistakenly) endorsed in *Explaining Behavior*, when (he says) he in fact ought to have opted for Theory II:

> Bratman has convinced me that I waffled once too often. Or not often enough. Anyway, despite some ambiguity in the text, I ended up with Theory I when what I really wanted, what I really needed is, Theory II. It is Theory II that makes my account of desire exactly parallel with my account of belief. It is Theory II, and *not* Theory I, that, making the content of desire into a historical (developmental) property of the desire, makes the fact that the desire has this content explanatorily relevant (as a *structuring* cause) to the behavior the desire is invoked to explain. (Dretske 1990, p. 834).

The quote, interestingly, implies that having the account of desire parallel the account of belief is a desirable goal. Nowhere that I have found does Dretske raise the question of whether it is indeed a good thing to have an account of desire parallel an account of belief. One may well wonder if the nature of belief and desire and their respective roles in shaping behavior makes a parallel account reasonable. Moreover, does Dretske's view of belief as states with an

information-carrying function make such a parallel reasonable?

One reason for thinking that the accounts ought not to be parallel is that doxastic states, on Dretske's view, are directly dependent on the external condition which will (on promotion) determine their intentional content. Motivational states are not similarly dependent. A token indicator of water depends on a non-accidental relation with actual water. If there is no water, then it cannot be an indicator. What is the analogy with a motivational state? Prior to D's becoming *for* water, must tokens of D bear some receptivity relation to water? Certainly, water need not be present (nor obtained) in order for the token in question to be a receptivity that *would be* satisfied by water. It is the fact that tokens of some type of receptivity state D would be satisfied by water (regardless of whether they have been so satisfied or have occurred in the presence of water) that determines whether D could be a desire for water, not the reverse.

In trying to sort out Dretske's theory of motivational states, two important questions arise which are different aspects of the same issue:

(a) What is the mechanism by which a desire's intentional object is determined?

(b) Where is the line between desires and drives?

There are definite tensions in Dretske's attempt to differentiate desires and drives while accounting for the intentionality of desire in a way that fits with his behaviour-related framework. Drives in *Explaining Behavior* seem to be receptivities that are "wired in" to selectional behaviour patterns. The blowfly and its feeding responses are an example. The drive in this case is a receptivity-state that, when engaged, is a partial cause of a movement which occurs whenever the organism's knees contact the right sort of liquid.

Thinking of drives this way makes for a tidy explanation of why drives are non-intentional:

the outcome of tokenings of behaviour caused (partly) by the drive-state have no effect on the organism's subsequent behaviour. Hence, the drive's being satisfied[14] by outcome R is irrelevant to what the organism does. This in fact coincides with Dretske's reasoning for the irrelevance of selectionally acquired representational states. A certain kind of high-frequency sound triggers a B-token in a moth which causes random darting and fluttering flight movements. This is a tropistic behaviour inherited from ancestors who survived bat attacks. In the particular moth, B-tokens' indicating the presence of a predatory bat (or not) is (causally and explanatorily) irrelevant to the moth's behaviour. It is so because the probability of the moth's continued execution of the movements caused by B is immune B-tokens in that moth indicating (or not) the presence of a bat (Dretske, 1988, p.93).

In the later (1990) reply to Bratman, the category of drives has undergone an expansion:

> Such things as thirst (for water), hunger (for food) and sexual urges are best understood, I think, not as *desires*, not even as *pure* desires, but what I (in the book) call *drives* (for water, food, and sex). One can be thirsty before one has had water, one can be in an internal state that makes water reinforcing, but one cannot have a desire for water (at least if this is understood as a *pure* desire) until one has *had* water. (Dretske 1990, p. 835)

(The qualification in the last sentence, that one cannot have a *pure* desire for water without having had water is to make clear that, by means of other beliefs and desires, one can *cognitively derive* a desire for an R that one has never in fact obtained.)

The views expressed here differ importantly from the view of drives I have sketched from comments found in *Explaining Behavior*. The key point is the claim that a receptivity state can make water reinforcing without being a desire for water. According to what is expressed in the foregoing quote, as yet unsatisfied thirst cannot be a desire at all. In *Explaining Behavior*, a

receptivity's being reinforcing is one way of drawing the line between drives and (pure) desires. A blowfly's probiscis extensions are partially caused by a drive-state (and not a desire-state) because the behaviour is impervious to reinforcement. A drive is not for any outcome, but is simply a receptivity that will set off a movement (invariably) in a certain set of stimulus conditions. A desire, on the other hand, by being *for* some outcome R, will support modifications of behaviour. Those which (although they may have tended to yield R in the past) cease to yield R over time are repeated with less and less frequency. Behaviours which tend to yield R become entrenched. This appears to be the 1988 (*Explaining Behavior*) distinction.

Let us summarize the two views of "drive" at this point.

(1) The 1988 view: A drive is a non-reinforcing, non-intentional movement-causing receptivity state of a system S.

(2) The 1990 view: A drive is either a state as defined in (1), or is

a) a receptivity state of S which is reinforcing, and is such that,

b) although that state has not been a structuring cause of any learned behaviour in S in which obtaining R was relevant, if S *were* to obtain R, the receptivity *would* be satisfied.

I add the proviso b), that the drive has not been a structuring cause, in order to maintain a *prima facie* distinction between drives and (pure) desires.

The drives of view (1) are a proper subset of the drives of view (2). I see no evidence in Dretske's writing of a change in categorization which would make, for example, the blowfly's receptivity to sugar-water something other than a drive. View (1) is basically the idea that drives are receptivities which are parts of selectional Design Solutions. That they are parts of inherent behaviour-packages is what accounts for their imperviousness to reinforcement. On view (2) a

drive may either be this kind of state, or it may be a receptivity state that could make conditions reinforcing, but has not yet acquired the kind of history that Dretske requires for it to be a type of state which is *intentionally for* R.

The shift in Dretske's view of drives from view (1) to view (2) is a consequence of a shift in his view of desires from (Bratman's) Theory I to the Theory II. On view (2) the category of drives expands to fill the hole left by insisting that even pure desires must be the product of a developmental Design Solution. The expansion, however, makes drives a heterogeneous class. The rationale for the grouping is not easy to justify.

First, there are the original non-reinforcing, non-intentional drives. I will call these "n/n drives". Second, there is a kind of drive which appears to function as a sort of "prospective desire". These I will call "p-d drives". That P-d drives are receptivities which would be satisfied by certain world conditions allows them to have a causal/explanatory role in the structuring of behaviour. Unlike n/n drives, p-d drives' being satisfied can bring about changes in an organism's behaviour.

In the 1990 work, Dretske allows that drives could include states that make certain conditions reinforcing. They could therefore, at least in the sense of making some outcome R reinforcing, be said to be *for* some outcome R. A "thirst (for water)", one of Dretske's later examples of a drive, is quite obviously *for* something in *some* sense. It is for what would satisfy it. One might reasonably suggest that we look to what a type of receptivity would make reinforcing, its conditions of satisfaction[15], following Bratman's description of Theory I as determining what that receptivity is *for*.

Dretske's move from Theory I to Theory II has an immediate counter-intuitive consequence that can be described as follows. When PJ is still young and is (plausibly considered to be)

undergoing the learning requisite for him to feed himself, *something* must lead him to try the water-spout again. Until he has "trained up" to a water-obtaining behaviour, this something, an internal receptivity-state of some kind, cannot be a desire *for* water. In fact, if he has not already learned how to obtain any potable liquid, it cannot be a desire *for* a potable liquid of any kind. So PJ has no desire *for* a potable liquid, but somehow, via the obtaining of such liquids, PJ will manage to acquire a water-obtaining behaviour. There is, therefore, no relevant desire available to account for the establishment of PJ's water-obtaining behaviour. The structuring of this behaviour explains, rather than being explained by, PJ's having a desire *for* a potable liquid.

There can be no accounting for PJ's acquiring this behaviour unless something it brings about (water, presumably) reinforces the behaviour. What it has tended to eventuate, on the way to becoming an established routine of PJ's, has to be rewarding. There would be no difficulty if one could say that PJ desires water and that is what reinforces the spout-licking movements. Since, according to the 1990 Dretske, this desire *follows from* the spout-licking routine being reinforced, it is not available as part of the explanation of why that routine is reinforced. This problem is avoided by the initial Theory I. On Theory I, PJ's acquiring of water-obtaining behaviour can be explained in part by appeal to its having a receptivity which (in virtue of its conditions of satisfaction) is *for* water (or at least is *for* drinkable liquid). The state has the effect it has on the structuring of behaviour in virtue of its being of a type such that it has the conditions of satisfaction that it does.

One might counter that a receptivity state of any kind has some condition of satisfaction. There is some condition (existing or not) which (if that condition could come about[16]) would satisfy it. Even n/n drives can be construed as having conditions of satisfaction. Thus an n/n drive will

also be *for* something. The blowfly's drive that triggers probiscis-extension when it perceives sugary liquids is satisfied. Despite this, surely we do not want to attribute an intentional motivation in an explanation of a blowfly's tropistic feeding behaviour.

My answer is that there is no need to do so. The line for intentional "*for*-ness" can be drawn in a way that draws on Dretske's comments on the general irrelevance of intentional states in explaining the structuring of an individual system's behaviour. A state connected to a tropistic behaviour in Dretske's view (the selectional version of C's causing M) has a function to indicate. Tokens of such a state can indicate perfectly well, but what tokens of C of an individual S do indicate is irrelevant in accounting for what S does. If C-tokens, in selectional solutions, cease to indicate what they had acquired the function of indicating, there will be no effect on S's behaviour. So C's semantic properties are irrelevant to its behaviour.

The same move is available for receptivities. N/n may be said to be *for* something: there is some outcome that will "switch off" an onset of the state. Whether the movement to which the state is connected (selectionally) is rewarding has no effect on the probability of the movement being caused by a future onset of the receptivity (and the usual perceptual stimulation.

However, it is not genuinely intentional, and thus not reinforcing, because its actual satisfaction (or non-satisfaction) in various conditions cannot affect the blowfly's behaviour. Actual satisfaction of the receptivity is irrelevant to the blowfly's (attempted) feeding behaviour. At the other end of the motivational scale, derived desires are receptivity states whose conditions of satisfaction have been derived from other receptivities and beliefs. For example, my desire to be wealthy (a condition I have certainly never attained) has a satisfaction condition (being wealthy) which is derived from a desire to be popular and a belief that wealthy people are popular.

Another potential objection to Theory I, on which p-d drives would be re-classified as (pure) desires, might be that receptivities not yet linked to a behaviour are in a sense too general. Consider an organism with a receptivity which turns out to be relieved by water. It does not follow from this that the receptivity must have been *for* water because it was of a sort that would be relieved by the obtaining of water specifically. There are substances the obtaining of which would have had the same effect. A receptivity that has at some time never been relieved by water cannot (in general) be specifically *for* water.

One development of the claim that an organism's receptivity state cannot be specifically *for* water (prior to having at least been satisfied by water[17]) is that there may well be other outcomes which would have had the same effect (the obtaining of some other drinkable liquid, for example). Again, what follows from this? It certainly does not follow that the receptivity (prior to having at least been satisfied by water) cannot be *for* anything. Nor still does it follow that the receptivity, once connected into a behaviour routine in virtue of being satisfied by some specific outcome, thereby gains specificity (i.e of being *for* that outcome).

For simplicity, let us suppose that the receptivity type in the given organism has never been satisfied[18]. Its conditions of satisfaction include water but include other outcomes as well. So the conditions of satisfaction are more general than the particular outcome "obtaining water". There is therefore no particular reason to identify the receptivity as being *for* water. This is true, but neither is there a reason here to claim that the receptivity is not *for* anything. What it is for is simply more general than water. Perhaps there is a convenient term to describe the general sort of outcome that would satisfy this state, such as potable liquid, or cool, non-poisonous liquid. The point is that, if a receptivity state is such that it would be satisfied by some conditions and not

others, those satisfaction conditions, general though they may be, provide some basis for identifying what the receptivity is *for*. If Dretske requires of desires only that they be receptivity states that are *for* some outcome, then it is not clear why any receptivity state with satisfaction conditions would be a desire, regardless of whether or not it has in fact been satisfied.

Additionally, the fact that the potable liquid which satisfied an organism's receptivity was always water does not license the inference that the content of the resulting desire is "water". That the receptivity was so satisfied may make it reasonable to exclude the possibility that the resulting desire is *for* beer or *for* carrot juice. The receptivity's having been satisfied by water does not exclude the possibility that the desire is a desire *for* some potable liquid, a state that is just as general as the "p-d drive" from which it was derived.

Consider an abstract case in which a subject S has some receptivity state (a p-d drive) that would be satisfied by R_1, R_2, or R_3. For convenience, let us suppose that there is some general term under which these outcomes fall, Rg. The receptivity state, in other words, would be satisfied by Rg. S, as it happens, acquires a piece of behaviour which is, during the structuring period, reinforced because an internal state C's causing M when a condition F was present (over the learning period) brought about R_1. Does this make the resulting state a desire *for* R_1? Well, that the obtaining of R_1 is what reinforced the behaviour is certainly an argument *against* claiming that the acquired desire is a desire *for* R_2. Even so, it is no argument at all against claiming that the acquired desire is a desire *for* Rg.

So the mere fact that a receptivity state (with some general satisfaction conditions which include water) has been satisfied by water does not allow the inference that it therefore comes to be desire that has "water" as intentional content. Nor is this inference allowed by the receptivity's

participation in behaviour modification. If it could do this job simply in virtue of being a state satisfiable by a drinkable liquid, then, on the face of it, the state ought to be "promoted" to a desire with that content. It *may* yield a type of state one might be inclined to characterize as a desire (specifically) *for* water. That, I suspect, would depend on the subject's other beliefs and/or desires, but we are here interested in what would make a (non-intentional) receptivity into a pure desire, not a *derived* one.

There are as well some positive considerations for thinking that what I have called p-d drives really belong in the category of desires. The first is that Dretske's reasoning for excluding inherent behaviour-governing states (those internal states connected selectionally to movements) do not apply to p-d drives. P-d drives do effect behavioural changes in virtue of their having conditions of satisfaction. Conditions of satisfaction can be reasonably understood as determining the "*for*-ness" of receptivity states. A given receptivity state's having the satisfaction conditions that it does determines the outcomes that it will make reinforcing. The outcomes it makes reinforcing determines the effects it has in shaping (learned) behaviour.

Second, one may well ask how it is that a p-d drive manages to have the effect that it does. N/n drives do their jobs in a way that renders their being satisfied irrelevant (in a particular organism). N/n drives are in particular individuals already wired to movements. P-d drives have to *become* wired to movements (in order to become effective in persisting behaviour processes). How do the satisfaction conditions manage to make a difference such that a p-d drive could become a part of a Dretskean Design Solution? The answer appears to be that it does so by making some type of obtained result rewarding to the organism. It is a conscious difference. The way things feel to a system during the onset of a receptivity (thirst, for example), and/or the way things feel when

a certain result is obtained is what makes that outcome rewarding.

Third, a reasonable supposition about conditions such as thirst or hunger is that they are receptivities for which the satisfaction conditions are linked to the physiology of the organism in which they occur. That a certain type of receptivity of mine turns out to be of a type that would be relieved by water, but not by saltine crackers, has nothing to do with what I have in fact encountered in my environment. It has everything to do with my physiological condition: that receptivity is the one which is activated by a deficiency in body fluids. Some range of outcomes, including the obtaining of water, turn out to be the ones which "deactivate" an onset of the receptivity by restoring necessary fluids. The key point is that what makes a certain type of general receptivity thirst is how that receptivity is (inherently) connected to the physical condition of the organism. That is what determines the conditions under which instances of the receptivity will occur, and as well determining what conditions will (or will not) satisfy the receptivity.

From these observations, I suggest that p-d drives can be thought of as a subject's inherent awareness of its own internal condition. A p-d drive is not going to be effective in shaping behaviour unless it has some kind of phenomenological impact. The basis of this, the conditions for the onset of a given type of such a receptivity and for its satisfaction, is one's own physiological condition. A p-d drive can thus be construed as *representation* of one's internal condition. A state, such as thirst, being thus directed at the alteration of one's physiological condition is prior to that state's participation in a *developmental* Design Solution.

P-d drives, such as thirst, are sure to make trouble for Dretske's theoretical categories. On the one hand, thirst is after all a receptivity *for* something. It may not be *for* a particular substance, but it does have satisfaction conditions. However, the reasonable suggestion that its intentional

content be determined by its satisfaction conditions leads to a state with active "meaning" properties that is not the product of the right form of the Design Solution. On the other hand, the states that I have called p-d drives can be classed together with hard-wired internal controllers of tropistic behaviour. There is no obvious reason, at least none that is not motivated by Dretskean theoretical concerns, for this classification.

The Theory (of Bratman's two proposed possibilities) to which Dretske commits himself is Theory II. I have argued for the greater plausibility of Theory I. Theory II, by requiring a developmental solution, says that a pure desire's content is determined by the satisfying condition that has in fact reinforced some behaviour. This version raises difficult questions. Bratman writes[19]:

> Dretske is clear that his account commits him "to denying... that a pure desire can figure in the explanation of behavior if one has never experienced a gratification of that desire." (p. 148) On Theory II Dretske must also deny that one can even *have* a pure desire for that which one has not yet experienced. Consider now a teenager who has as yet had no significant sexual experience. Since he has not yet engaged in significant sexual activity, he has never gratified a desire for such activity. On either Theory I or Theory II it follows that we cannot explain any of his action by appeal to a pure desire for sexual activity. On Theory II it even follows that he can as yet *have* no such pure desire. Are these plausible results? (Bratman 1990, pp.799-800)

The two questions Bratman raises are:

(1) Is it plausible to suppose that a receptivity-state, prior to having been satisfied by some condition cannot explain an organism's behaviour?

(2) Is it plausible to suppose that, prior to a receptivity-state's being a partial cause of the entrenchment of a learned behaviour routine in virtue of being satisfied by R, instances of that type of state cannot be an intentional motivational state (a desire) *for* R?

Question (1) is a problem that confronts any version of Dretske's account, for he insists that

a motivation can only explain behaviour through its being satisfied by some condition which makes the behaviour reinforcing. Dretske's account of desires, to co-opt Bratman's example, precludes the possibility of explaining a teenager's "acting out" in partial terms of sexual frustration or tension prior to that teenager's having had actual sexual experience. The plausibility of this result is certainly debatable.

More important, in my view, is question (2). There seems to be a tension in Dretske's Theory II over the need, on the one hand, for a reinforcement mechanism and, on the other, Dretske's historical requirement on states' having intentional properties. Something must explain what moves PJ to explore its environment and what reinforces the movements which result in water being delivered from the water bottle in order for PJ to learn to go back to the water bottle. A state that PJ is in from time to time must be such that obtaining water is a kind of result that is rewarding (i.e. reinforcing) *before* one can expect the spout-licking behaviour to be reinforced. The answer seems to be that the state that PJ is in is of a kind that would be rewarded by obtaining water, and its being that kind of state has nothing to do with PJ's behaviour or with what PJ has in fact obtained in the past. Moreover, the idea that PJ is rewarded when it does obtain water suggests some apprehension of its own internal condition. That PJ's state (a state that we call "thirst") would be satisfied by some conditions and not others suggest that is already has directedness. The state's being for some outcome is a function of what would satisfy it.

Summary

Dretske is initially committed to a commonsensical, and fairly reasonable account of motivational states and a distinction between intentional and non-intentional receptivities. This is Bratman's Theory I. Bratman has convinced Dretske to opt for a theory, called Theory II, which

has the account of desires "paralleling" that of belief[20]. Dretske's need for this shift is motivated by the overall approach he is taking to an account of intentional states. Theory II, however, is not an attractive theory. It posits receptivities with satisfaction conditions that affect an organism's behaviour but which fail to be *for* anything. On Theory II, such states can be "promoted" by a developmental solution to the status of a desire. Their status prior to such a promotion is unclear. What accounts for PJ's being conditioned to lick the spout on the water bottle if PJ does not, prior to this development solution, already have a relevant intentional motivation? Perhaps the state is not specifically a desire for water (PJ may *never* have such a desire), but it is hard to see how a behaviour could be reinforced without an existing receptivity for something (a potable liquid) which water (among other things) would satisfy. It is hard to see why one would deny that such a state was not a desire[21]. If, as Dretske and Bratman both suggest, Dretske's general theory of intentionality needs Theory II of desires, then so much the worse for Dretske's general theory.

c. Motivational States and Behaviour: Sniffy Gets His Oats

With the addition of motivational states, Dretske is able to fill out his picture of the structure of behaviour and of explanation by reasons. In essence, a type of behaviour is a type of movement M which has come to be jointly "governed" by a type of internal belief-desire complex (B/D). B/D has come to cause M in the presence of some type of external condition, F, because this has brought about R, and R is reinforcing. Dretske goes on to elaborate that what a piece of successful operant learning brings about is a higher order procedural belief that doing M when F is present will bring about R. Dretske's intended view is that when an organism whose behaviour with respect to some outcome type R is explicable in terms of reasons, the explanation takes the form: S did (behaviour)

E because it believed that F, wanted R and believed that doing M in the presence of F would bring R about. The higher order procedural belief that doing M when F is present will yield R can be either implicit or explicit. Which it is determines the breadth of the range of behaviour that the procedural belief could support. As the procedural belief is crucial to the explanation of behaviour in terms of reasons, it is crucial to the causation of behaviour by intentional states.

In this section, I wish to examine the way in which Dretske assembles his model of behaviour with the inclusion of motivational states. This is somewhat puzzling for a couple of reasons. The first is how exactly to impose the components of Dretske's general model on the particular type of example (that of a rat being trained to press a release bar) that he employs. What is the content of the rat's internal representation? The second is over the introduction of the sort of "procedural belief" mentioned in the preceding paragraph. For the most part, Dretske appears to be taking what I call a "deflationary" approach to intentional states. A current C-token, for example, being a representation of F just comes to its being of a type which has H-indicated F (and has come to cause M because of this). The discussion of procedural beliefs comes very close to what I will call an "inflationary" view of representation: from the learning situation, there emerges a representation of the form "doing M in the presence of F will bring about R".

(i) Sniffy

In making his general structure more concrete, Dretske makes frequent use of the example of rats being trained to respond to some stimulus-condition in order to obtain food. The following is my version of this kind of example: Sniffy the rat comes to press a food-release bar when a red light is illuminated. There is a hopper which can hold food and the food is released by the pressing of the bar. When and only when a red light that the rat can see is illuminated, the hopper contains

food. After a certain number of trials, Sniffy acquires the appropriate sort of discriminative behaviour: pressing the release bar when the light is illuminated but not otherwise. In order to be able to do this, on Dretske's view, Sniffy needs an indicator of the relevant external conditions -- certainly a sensitivity to the light's being illuminated or not -- and he also needs a receptivity for the food in the hopper. B is an internal physiological state caused by Sniffy's perception of the illuminated light. D is Sniffy's receptivity for food. M is the bar-pressing movements which release the food. When Sniffy sees the red light on, there occurs in him a B-token. This produces nothing if Sniffy is not hungry, but if Sniffy is also in state D (a receptivity state for food), these two states together come to cause movement M. As Dretske puts this at one point, B and D's "recruitment" makes them "necessary parts of a sufficient condition for M" (Dretske 1988a, p. 113)

(ii) Imposing the Model

There is some question about how exactly to impose Dretske's model on this specific example, however. As a result of such a learning process, B-tokens should come to be representations of some external condition. It is not clear what condition it is that B acquires the function of indicating. Suppose that, during training, the experimenter places grain in the hopper each time that the light is illuminated. In the experimental set-up described, B-tokens would be indicators both of the red light being illuminated and of food being in the hopper. So Dretske could be thinking of B as acquiring either one of two functions to indicate: the function to indicate that the light is on or the function to indicate the presence of food in the hopper.

One element of Dretske's theory has a link to teleology. The content of a state with an acquired indicator function is determined by that which would be of some benefit to the organism. On the developmental solution a "benefit" is a receptivity-satisfying outcome, or so one must

suppose. After all, an outcome that conditions behaviour has to be a rewarding (or reinforcing) one[22].

Dretske, in discussing an example of a dolphin being trained to respond to cylindrical objects (from distances of fifty feet), writes:

> But a dolphin that can infallibly identify, detect, recognize, or discriminate (use whatever cognitive verb you think appropriate here) cylinders from this distance should *not* be credited with the ability to identify, detect, recognize, or discriminate, say, red objects just because all (and only) the cylinders are red. If the fact that all (and only) cylinders are red is a coincidence, of course, then something can indicate that X is a cylinder without indicating that X is red. This follows from the fact that an indicator could exhibit the requisite *dependence* on the shape of X without exhibiting any dependence on the color of X. But even if we suppose the connection between color and shape to be more intimate, we can, because of the different relevance of these properties to the well-being of an animal, imagine a detector having the function of indicating the shape of things without having the function of indicating color. (Dretske 1988a, p. 76-77)

The way I read the end of this passage is as suggesting that the issue of content-determination is partly settled by appeal to what condition, in a Design Solution, is relevant to the well-being of a system. (Where this involves learning, the condition cannot merely be something that would in fact promote the interests or needs of the system. It has to be some condition which satisfies a receptivity.)

There then appear to be two factors in content determination from a developmental Design Solution. One factor is connected to indication. Type B must H-indicate a condition F over the solution period. The second factor is connected to receptivity satisfaction, to what has some purpose for the system. So if tokens of Sniffy's internal B-states, over the training period, indicate the presence of food in the hopper and also that the red light is on, one must infer that Sniffy's B-state, after training, has something akin to "food in the hopper" as its acquired representational

content. It is surely not because a red light is illuminated that Sniffy develops the habit of pressing the bar when the light is on. Sniffy would not have acquired this behaviour if the pressing of the bar under the specified conditions did not result in food. On the other hand, if there were a non-accidental correlation between food and the illumination of a green light, Sniffy would presumably then learn to press the bar in the presence of the green light.

However, Dretske makes comments which suggest that he is thinking of B as coming to function to indicate that the red light is illuminated. In going on to describe the next stage of his theory Dretske writes: "So, it seems, wanting food (D), and knowing the light is red (B) are *not* sufficient to produce bar-pressing movements." (Dretske 1988a, p.116) The identification of B with "knowing the light is red" invites the interpretation that Sniffy's internal state B comes to function to indicate the light's being illuminated, rather than coming to function to indicate the presence of food.

So there are two ways (at least) of depicting the rat example in relation to Dretske's framework:

(1) R is the obtaining of food (from the hopper). F is the food being in the hopper. B is a physiological state (caused by Sniffy's seeing a red light illuminated) which indicates that food is in the hopper, and D is a receptivity for food.

(2) R is the obtaining of food (from the hopper). F is the red light's being illuminated. B is a physiological state (caused by Sniffy's seeing a red light illuminated) which indicates that a red light is illuminated, and D is a receptivity for food.

In (1) it is the dependency between illuminations and food in the hopper that makes tokens of B indicators of food in the hopper. These are also indicators of red light illuminations.

Now how one fits the Dretskean schema onto this example determines whether Sniffy comes to do M in the presence of some condition F because Sniffy is in an internal state which indicates the presence of food, or one which indicates that a light is on. What actually does the "structuring", the changing of the probability of Sniffy's internal (complex) state C's causing M, is surely the relationship between the occurrence of B-tokens (the informational components of C) and the presence of food. Sniffy is hungry and he does not especially care about lights. At least there is no reason to think that he would. Of course the dependence between food in the hopper and a certain physiological state *goes through* the illumination of the red light. The illumination of the light is dependent on food being in the hopper (because experimenters make it so). The occurrence of B in Sniffy is dependent on (Sniffy's perception of) the red light being illuminated. So occurrences of B in Sniffy are dependent, by transitivity, on food being in the hopper.

So it would seem that the way in which Dretske's general scheme should be imposed on the case of training a rat to release food would be according to (1). B's indicating the red light does not explain why Sniffy, who occasionally wants food, learns to press a bar in order to release food. This is just what is suggested if F is understood as denoting the light's being illuminated. However, Dretske's discussion makes it clear that he is thinking of Interpretation (2). Explaining the difference between a trained rat ("the first rat") and its untrained companions, Dretske writes:

> The first rat, as a result of training, must know, or at least believe, that pressing the bar when the light is red will bring food. The other rats don't know this. So the complete, or at least *more* complete, explanation of the rat's behavior is that he knows the light is red, wants food, and thinks that by pressing the bar when the light is red he will get food. (Dretske 1988a, p. 116)

Perhaps as Dretske imagines the example the condition of food's being in the hopper is not, during Sniffy's training, H-indicated by the internal B-states. Perhaps some tokens of B occurred

in the absence of food in the hopper. This possibility, and the fact that it would alter the resulting content of B is telling: There are indeed two content-determining factors on Dretske's theory. Moreover, Dretske sees a state's having the appropriate indicator history as the dominant one. What is indicated during training is different from what tends to reinforce a behaviour. We need only imagine that the experimenter occasionally turned on the red light in Sniffy's presence when there was no food in Sniffy's hopper. When that which is indicated during training is different from that which (in the training situations) is of benefit to the organism, Dretske takes H-indication to be the factor which determines a states content. So if non-beneficial light-illuminations are H-indicated (every B-token occurs non-accidentally in the presence of the light's being illuminated) but the *rewarding* condition of there being food available is not H-indicated, Dretske's theory decides content in favour of light-illuminations. This is a point which will lead to trouble, as we shall see in the next chapter, especially in section f. There I examine the two content-determining factors in more detail in connection with Godfrey-Smith, who argues that Dretske's view gives the content-determining weight to the wrong criterion.

(iii) Procedural Beliefs

Previously, I cited Dretske as saying "So, it seems, wanting food (D), and knowing the light is red (B) are *not* sufficient to produce bar-pressing movements." (Dretske 1988a, p.116). This statement sets the stage for the introduction of the procedural belief. The implications of what Dretske says about state B are interesting. B is identified as denoting a belief (or some other doxastic state appropriate to rats) that the light is illuminated (*knowing* that the light is red). In the context of the quote, the rat in question has not yet learned to press the bar when the light is on in order to obtain food. How did B, which is in fact a type of physiological state (of a rat in the case

at hand) come to have a representational content describable as "that the light is illuminated"? Dretske's account of how ground-level beliefs come to be beliefs is tied directly to the formation of behaviour. There is no more primitive mechanism than the developmental Design Solution for the formation of doxastic states with representational content. What "red-light-relevant" behaviour accounts for B's having acquired the *function* of indicating that the red light is on[23]? It cannot be bar-pressing behaviour, for we are here concerned with rat-states prior to its learning this behaviour. There is no suggestion in Dretske's discussion that his rats have acquired any *other* behaviour which has come about in virtue of the physiological state B being an indicator of a red light being illuminated.

Possibly this is a simple lapse, that Dretske was speaking loosely in describing B (prior to the learning of bar-pressing behaviour) as if it were an intentional state with a specific content. He should have described an untrained rat as having a receptivity for food and being in a state of perceptual stimulation which indicates that a red light is on. The state in question does not have the *function* of indicating this. It will turn out, however, that the problem here is not so easily handled: at the next stage of the Dretskean account, B's being an intentional state is essential.

Dretske connects the formation of a new behaviour with the acquisition of a "procedural belief". Sniffy's training, the correlating of food in the hopper with the illumination of the light when Sniffy is hungry, results in Sniffy coming to press the hopper release bar when he sees the light is illuminated and not otherwise. Sniffy, according to Dretske, now has a "belief" which explains why he presses the bar when the light is on (and not otherwise). This "belief" is absent from an untrained, hungry rat that can see illuminated lights. The belief is that pressing the bar when the light is illuminated will deliver food.

> In describing [a trained rat] as wanting food and thinking the light is red, we have not distinguished him from his cousins who know and want the same things but behave quite differently. It seems we need something in the *way* of a background belief about the way to achieve one's ends, some belief to the effect that producing M in conditions F will bring about R. The first rat, as a result of training, must know, or at least believe, that pressing the bar when the light is red will bring it food. The other rats don't know this. So the complete, or at least a *more* complete, explanation of the first rat's behavior is that he knows the light is red, wants food, and thinks that by pressing the bar when the light is red he will get food. (Dretske 1988a, p.116)

"Procedural belief" of the sort that Sniffy acquires may not be sophisticated enough to warrant attribution of full-blown goal-directed behaviour in Dretske's view. However, he does claim that such background states are necessary components of "goal-directed, purposive" behaviour. It is only where behaviour is purposive that it is explicable in terms of reasons. So it should follow from this that only such behaviour as it supported by procedural beliefs is explicable in terms of reasons. This is what Dretske has to say about background beliefs and purposive behaviour (his reference to "this objection" is a reference to the point that a belief that the light is illuminated and having a desire for food is insufficient to explain a rat's bar-pressing behaviour):

> This objection raises an important point about the way background beliefs, beliefs about the efficacy of means for bringing about ends, figure in the explanation of goal-directed behavior. To explain someone's performance it isn't enough to point out that, in the conditions he believes to obtain, his behavior leads to results he desires. One also wants to know whether the agent *knows* (or at least believes) that his behavior *will* lead to results he desires. Was his getting a desired result a mere fluke, or did he do what he did *in order* to achieve that result -- with the belief that such action would, or would likely, get him what he wanted. (Dretske 1988a, p.116)

Dretskean procedural beliefs may be implicit or explicit. Dretske says that they can be thought of in terms of dispositions (Dretske 1988a, p.117). He describes the implicit ones as "single-track" dispositions[24], rules or, in computer science terms "productions". Explicit procedural beliefs are termed "multi-track" dispositions. The idea is that the difference between implicit and

explicit should be a difference in the behaviours that a background belief will support. Where a belief of the form "doing M when F is present will bring about R" is implicit, it will only yield instances of M when the organism believes that F is present and is in the appropriate motivational state. It will only bring about instances of C's causing M. An explicit procedural belief is more flexible in the range of behaviour-types in which it may play a causal role. My implicitly believing that carrying an umbrella when it is raining will keep me dry supports only my carrying an umbrella when it is raining (given that I want to keep dry). My explicitly believing this may bring about a variety of behaviours in conjunction with other motivational states. If I do not want a friend to get wet, I may suggest that she takes an umbrella along as she leaves to go outside. I may hold my umbrella over a companion with whom I am standing (given that I want to make a kind gesture). I may hide my brother's umbrella as he prepares to go out in the rain (given that I do want my brother to get wet). Dretske's idea is that an explicit procedural belief allows some range of inferences to other behaviours given different desires. An implicit procedural belief does not allow such inferences and thus only disposes an organism to execute a specific type of behaviour under certain circumstances.

Dretske also states that procedural beliefs are defined over intentional elements. This is his basis for a distinction between dispositions or rules which are beliefs and those which are simply physical processes that tend to have a certain type of result although they are not in any way intentional. We do not want to be forced to ascribe procedural states to sugar lumps in order for them to dissolve or to cars in order for them to start when their ignition key is turned. In his discussion of implicit procedural beliefs, Dretske writes:

Rather, an implicit belief is a disposition or rule that describes the relationship among

> entities that are already intentionally characterized (beliefs and desires, for instance) or among such intentionally characterized entities and movements. A rule or disposition that says, simply, "Produce movement M when conditions F and D obtain" is *not* an implicit belief. It is, at best, a mere regularity of some kind. My car "follows" the rule "Start the engine when the key is turned on, the battery is charged, and there is gas in the tank." But this is not a belief. In order to qualify as a *belief*, the disposition or rule must be defined over intentional elements (beliefs and desires) *qua* intentional elements. (Dretske 1988a, p.118)

So the account of the structuring of behaviour, the formation of behaviour by the developmental solution appears to be as follows: Tokens of some internal state B of an organism S indicate an external condition F. R is an outcome relevant to the presence of F to which S has a receptivity (from time to time at least) D. S, during the course of exposures to F, and when D is present, does M. This results in R. M is repeated when F is present (more accurately, when the F's presence is indicated by a B-token) and D is present. S thus learns to do M when tokens of B and D occur. So the types B and D come to be "recruited" as causes of M in the presence of F. Thus S acquires the procedural belief that doing M in the presence of F will bring about R. This rule constitutes a procedural belief because it is "defined over" the intentional states B and D, and in virtue of their being intentional states. The twin components of the new cause of a movement M are D, tokens of which are desires for outcome R, and B, tokens of which are beliefs or other doxastic states to the effect that F is present.

In the framework for behaviour structuring here described, it seems that we have more than one acquired doxastic state. To return to Sniffy's training, the result of Sniffy being trained to press the bar when the light is on is a procedural belief that pressing the bar when the light is on will deliver food.

In the description of the process by which Sniffy acquires this procedural belief, what is

denoted by "B" has shifted from being such that its tokens are F-indicators to being an F-representation. Otherwise, the acquired disposition to press the hopper bar when B and D are jointly instantiated will not be "defined over intentional elements". This acquired disposition will therefore fail to qualify as a procedural belief, or as anything more than a simple, non-intentional disposition.

Now what is supposed to account for B's being a representational state (rather than a mere indicator) is a process of behaviour-structuring, B's being part of a complex internal state C, tokens of which come to cause M. It cannot be Sniffy's acquisition of discriminatory bar-pressing movements which accounts for B's promotion to "food-in-the-hopper" representation. It is the fact that B-tokens are "thinkings" that there is food in the hopper and that these representations of external affairs figure into the structuring of Sniffy's bar-pressing behaviour which accounts for Sniffy's acquisition of the further (implicit) procedural belief. So by means of what piece of behaviour-structuring does Sniffy's indicator of the light's being red become a belief that the light is red? There are no obvious candidate suggestions. Unless, however, it is by means of some piece of behaviour-structuring that Sniffy acquires a belief that there is food in the hopper (or even that the red light is illuminated), then there must be some way, other than by a (developmental) Design Solution that organisms' internal states can come to be basic perceptual beliefs. It is unclear exactly what is built in to the notion of "being defined over elements that are already intentionally characterized".

The discussion of procedural belief also takes a view of the representational properties of belief states that is rather more "inflationary" than is usual in Dretske's exposition. In the present context, I mean by an "inflationary" view an interpretation of representational properties in which

these are distinct properties, over and above a cognitive system's physiological properties and the indicator relations those phsyiological properties bear to aspects of the system's environment. In the foregoing discussion of procedural belief, one is nearly tempted to think of these as cognitive items which emerge from an appropriate conditioning situation, and which may be explicitly entertained, as, for example, a "sentence in the head" or explicit thought. A current belief that doing M in the presence of F will bring about R is certainly not obviously constituted by past tokens indicating F.

Dretske's view of representational properties is more usually "*de*flationary", as exemplified in the following quote:

> To say that C_n means F is to say from whence C_n got its causal powers -- powers it is (presently exhibiting by helping to produce M_n. It is to say that C_n acquired its causal efficacy (the sort of efficacy relevant to the shaping of output) from earlier tokens of this type *indicating* (carrying the information) F. This does not make the appeal to C_n's meaning (in the explanation of why it is causing M_n) circular. For the causal job being explained is what C_n is doing (causing M_n -- current behaviour), and the property that is invoked to explain it, C_n's indicator function, is a property constituted by what earlier tokens of this type did. (Dretske 1991a, pp. 116-117)

(The subscript "n" is used in the notation here to indicate tokens.)

In this passage, the representational content of a current internal state comes to its having the property of being of a type past tokens of which were indicators of a condition F. The "deflationary" reading is reinforced by Dretske's subsequent claim that "...meaning is an historical property..." (Dretske 1991a, p. 117). This reading has a present function of indicating F being constituted out of past token having actually indicated F. The "inflationary" picture has a doxastic state with a "procedural content" somehow emerging from a developmental Design Solution.

Of course there is no reason, on the one hand, to suppose that Dretske thinks of non-natural

meaning properties as distinct or in any way emergent ones. On the other hand, his usual "deflationary" view makes something of a mystery of a distinction between *explicit* and *implicit* procedural beliefs. What would explain that a given procedural belief was explicit rather than implicit? Surely the explanation is *not* that the belief, as a dispositional state, is "multi-tracked" rather than "single-tracked". The direction of explanation should (if anything) be precisely the reverse. A belief's being explicit (whatever that comes to) is what accounts for its being, as a dispositional state, "multi-tracked".

Summary

This concludes my outline of Dretske's addition of motivational states. Throughout the development of his theory, Dretske has consistently worked to fuse an account of how the intentional states (*qua* intentional states) are explanatorily relevant to behaviour with an account of how certain of our internal, physiological states manage to have representational or intentional properties. Some difficulties have begun to emerge. It is not clear how such properties could arise through a developmental solution to the Design Problem without some states already being representational. There are questions about timing. Dretske appears at times to want representation (and "*for*-ness") to emerge from behaviour-modification when it is unclear how behaviour-modication could come about without such properties. As well, the official view is that indicators and receptivities work as *structuring* causes. They thus appear to have an historical relation to a piece of current behaviour. As we shall see, this has led critics to worry about whether Dretske's theory manages to assign the *right* causal/explanatory role to intentional states.

NOTES

1. The inclusion of "or some variant" here is to acknowledge that Dretske allows the possibility that some natural systems may have doxastic and/or motivational states, but the systems may themselves be too simple to be regarded as having states that properly qualify as beliefs or desires.

2. I will refer indifferently to "motivational state", "receptivity", or "receptivity state". I do not intend any difference between these terms.

3. There is also an ambiguity in the notion of a beneficial outcome (generally). Does one mean an outcome that is in fact of some benefit to the organism, or an outcome which is perceived by the organism as being of some benefit?

4. As we shall see later it is important to be clear that on Dretske's view belief and desire states are *not* themselves structuring causes of behaviour.

5. A belief (or belief-like state) is what arises from an application of the developmental solution. Internal states *indicating* a condition is what is required for the solution.

6. The crucial section this topic is Chapter 5, section 5.3 of *Explaining Behavior* (Dretske 1988a, pp. 122-127)

7. Possibly this could be more carefully put by saying that in a particular organism which exhibits a piece of inherited behavior, a drive-state which has a causal role in the behaviour process does not have that role in virtue of being for some goal.

8. There is some question, as will emerge shortly, about whether one should describe Dretske's view by saying that desires "*are* receptivities available for behaviour-structuring" or "*were* receptivities available for behaviour-structuring".

9. As will become clear in the discussion of the drive-desire distinction, this is a point to which Dretske is only ambiguously committed in *Explaining Behavior*.

10. Dretske 1990, pp. 836-839

11. I realize that there may be difficulties in identification and individuation of pure desires: a kind of "motivational state depth problem". Why is there not simply one pure desire, namely, the desire to be relieved of the phenomenal sense of wanting or receptivity? Suppose I now go to the kitchen, pour and consume a glass of water. Given that this is explicable in terms of reasons, what is the pure desire involved? Is it a desire for water? Arguably, my desire for water is derived from my believing (implicitly or explicitly) that water is a potable liquid and my desire for that. This desire is derived from my desire to relieve my thirst and my belief that a potable liquid will do the job, but the desire to relieve my thirst is derived from a desire to feel better, to be relieved of the receptivity. I desire not to desire. The desire for relief from the phenomenal sense of discomfort, of lacking something is the one which appears to be truly *pure*. However, if this is where I finally

uncover a desire which really is not derived from other cognitive states, how is it distinguishable from the pure desire that moves me to take a walk, fix a sandwich, or any other behaviour?

12. There is one last side point about the previous passage. It is an interesting (but obviously inconclusive) thought experiment. What if one were to reverse the order of PJ's training? Let us suppose that PJ's water bottle was initially filled with beer for some period of time for PJ to become conditioned to lick the spout when PJ was in some receptivity-state. Later, PJ's beer was replaced with water. In the original example, it sounds perfectly reasonable, even after an encounter or two with beer, to say that PJ's spout-licking is because he wants water. On my variation, even after a couple of encounters with water, one should have to say (analogously) that PJ's spout-licking (which yields water) is because PJ wants beer. This is not as obviously a reasonable explanation of PJ's behaviour. At least it does not strike me as being as intuitively obvious as explaining PJ's behaviour which yields beer by an appeal to a desire for water.

13. This is very general and deserves some qualification. Dretske, in endorsing such a view, would want to make it clear that the outcome R, or the object F of that outcome, would have to be desired under some description on which it was reinforcing. It must be distinguishable from potential alternatives. One cannot have a (pure) desire that is specifically for bottled water if one lacks a capacity to distinguish bottled water from tap water, for example.

14. That is, the deficit-state being, as it were, turned off.

15. The idea for "conditions of satisfaction" comes to me from John Searle, although it is not clear, because his purposes are different, that he means to use it as I have employed it here: "Conditions of satisfaction are those conditions which, as determined by the Intentional content, must obtain if the state is to be satisfied." (Searle 1983, pp. 12-13)

16. Could one not have desires for that which is logically impossible?

17. Of course Theory II requires more than the receptivity's merely having been satisfied.

18. This is impossible for water, of course.

19. The internal reference to Dretske in this quote is a reference to Dretske 1988a, p. 148.

20. I should note that Bratman does not himself endorse either of these views. There is no evidence to suggest that Bratman pushes Dretske in the direction of Theory II because he (Bratman) likes that theory. In fact the textual evidence would suggest just the opposite. Bratman concludes his critique of Dretske by pointing to implausible consequences of Theory II. I rather suspect that Bratman has argued that Dretske "needs" Theory II in the sense that this is the account which is consistent with Dretske's the theoretical commitments Dretske establishes in his account of representation and belief.

21. Leaving aside the issue of whether organisms such as rabbits are sophisticated enough to have their doxastic and motivational states legitimately referred to as beliefs or desires. The point is that

if PJ qualifies as a desirer at all, it looks as if it is one of his desires that reinforces the spout-licking behaviour.

22. Of course some conditioning outcomes condition by negative reinforcement, but the general point is the same. The general point is that "benefit" here is a reference to the experience of the organism, and not necessarily to what is, in some objective sense, beneficial.

23. Prior to a relevant solution to the Design problem, tokens of B of course can indicate a red light's being illuminated, but in Dretske's theory indicators are very clearly to be distinguished from representations and also from beliefs.

24. This is a reference to terminology initially employed by Gilbert Ryle in *The Concept of Mind* (Ryle 1949).

CHAPTER FIVE

a. Introduction

In this chapter, I will review and develop important criticisms that have been levelled against Dretske's approach to intentional states. The tight relationship between the two significant components of the theory, how representation arises from Design Solutions and the role played by internal states which come to be representations, is cause for concern and confusion. Has Dretske indeed shown that representations (or other intentional properties) do have an explanatory role in relation to behaviour?

The central points in the discussed criticisms are as follows:

(1) If Dretske does claim that on his theory some state's being a representation (or the fact that it is a representation) is causally relevant, then his theory is circular.

(2) Dretske's theory fails to assign the sort of "here and now" explanatory or causal relevance commonsensically assigned to beliefs and desires cited in explanations of behaviour.

(3) If the theory avoids the circularity in (1), then it does so at the expense of conceding that a state's being a representation is not causally relevant.

(4) Indication and H-indication are not in general requirements of operant learning, and so cannot be cited as structuring causes of learned behaviour.

Dretske's response to (4) diverges from most of his elaboration of his theory. He suggests that what he is doing is providing conditions for concept-formation. This appears to imply that in order for states to come to have non-natural representation, indication conditions need only be *satisfied* during behaviour-structuring: those conditions do not need to be causally relevant in behaviour-structuring. I will argue that such a view would be too weak for Dretske's purposes.

These issues are discussed in order of their listing here, although I address (2) and (3) together. Section b of this chapter concerns issue (1). Issues (2) and (3) are the topics of section c. Issue (4) is the topic of section d. This is followed by my elaboration of the implications for Dretske's strong version of indication and finally a look at the possibility of adopting an "indicational" theory of representation based on weaker requirements for indication.

b. Baker's Circle

In this section, I explore the charge that if Dretske's theory claims that some state's being a representation (or the fact that it is a representation) is causally relevant, then his theory is circular. This criticism is due to Baker. I will argue that her version of it fails. Nevertheless, we shall see in chapter 6 that Dretske's theory may be contain a more general, pervasive sort of circularity.

Baker has argued that Dretske's theory of mental semantics either fails to assign a causal role to semantic content or is circular[1]. The main part of Baker's article, however, is given over to showing how Dretske (allegedly) falls into circularity. A number of other critics have repeated the alternative charge that Dretske's theory assigns no role to semantic properties[2]. I will begin by focussing on the question of circularity, as raised by Baker.

I believe it is possible to answer this criticism as it stands because Baker's reasoning rests on a misunderstanding of the explanatory role Dretske intends to assign to content. However, considering it will clarify how we are led to the other horn of the dilemma: that Dretske's informational theory fails in its initial promise to assign a causal role to the semantic properties of intentional states.

Baker claims that Dretske must be committed to three claims:

(1) X's explanatory role is X's causal role
(2) A state C has an explanatory role in virtue of having meaning
(3) A state C has meaning in virtue of having a causal role. (Baker 1991, p.100)

(1) is a generalization over any phenomena but it is enough for purposes of the issue at hand that Dretske be committed to claiming that an intentional state's (and its content) explanatory role is its causal role. This equation of explanation and causation, combined with (2) and (3), yields the conjunction "A state has a causal role in virtue of having meaning, and has meaning in virtue of having a causal role." (Baker 1991, p.100).

Baker's argument proceeds through a series of definitions. Most of this series constitutes a step-by-step reconstruction of the Dretskean account of representation and, in particular, his account of developmentally acquired representations. The key steps highlight the connection between representation, functioning to indicate, indication, and structuring causes. The focus is on what Baker refers to as *natural* representation, representational systems, and functions to indicate. Baker says that she means the term "natural" to coincide with Dretske's type III systems (see chapter 3, section b.). This claim could be a little distracting because Baker's actual use of her term coincides with Dretske's type III representational systems where the representations have been acquired developmentally. A type III system may acquire its representational states either by learning or by inheritance through natural selection. In the latter case, it is Dretske's view that the resulting representational state does not have a causal or explanatory role in virtue of the representational properties. Baker's natural representational systems are those in which representations are causally active. Of course in practice typical examples of type III representational systems will have some type III representations acquired developmentally and

some acquired selectionally. So the kinds of systems (humans, for example) which exemplify Dretske's notion of type III will also exemplify Baker's notion of a system of natural representations.

Baker sets up her critique of Dretske by noting a concern that Dretske does not always follow the indicate/function to indicate distinction in practice. I think it is easy to slip, and the distinction between representation and indication is one about which a reader of Dretske must be vigilant, for in ordinary usage, "represents" and "indicates" are often synonymous. Baker voices her concern about the maintenance of the difference between "indicating" and "functioning to indicate" by highlighting a selection from *Explaining Behavior*:

> That Dretske does not always observe the difference is revealed in remarks like this: The "causal relationship between C and M, if it is going to be explained by the fact that C *indicates, or has the function of indicating*, how things stand elsewhere in the world." (EB, 84,; emphasis mine.) However, the difference between indicating and having the function of indicating is crucial in the context of explicating representation or meaning. (Baker 1991, p. 108)

We have noted previously (in chapter 3, pp. 99-101) an example in which "what C says about external conditions" appeared to be used variously as referring to representing and as referring to indicating. Another possible example of a slide between a state's simply indicating and its representing some condition to be found in Dretske's discussion of motivational states:

> Consider an untrained rat, or this (trained) rat *before* it was trained. Or think about a rat that was trained to do something *else* when the light was red in order to get food. Such rats might want food, know that stimulus conditions are F and *not* press the bar. So, it seems, wanting food (D) and knowing the light is red (B) are *not* sufficient to produce bar-pressing movements. (Dretske 1988a, p.116)

This passage leads up to Dretske's introduction of "procedural beliefs". Of particular interest here is the clear implication that an untrained subject (a rat) might have a belief that the red light

is on (state B of an untrained rat is identified with "*knowing* the light is red"). The *only* account given in Dretske's theory of how an internal *indicator*-state comes to have the semantic properties required for belief is through behaviour-structuring. One is left to wonder what behaviour, prior to the rat's being trained to connect the illumination with the presence of food, has been structured by the illumination of the light. If the answer is "none", then Dretske has drifted illegitimately from "indication" to "representation". The rat cannot be in a belief-state (thus cannot know) that the light is illuminated unless its internal state can represent this. It cannot represent this unless it has an internal indicator which has been "promoted" through a Design Solution.

Of course Baker is quite right that it is essential to maintain the indication/representation distinction. Indication is a natural foundation upon which, in Dretske's theory, representation is based. Elsewhere, Dretske has been very explicit about the importance of the distinction (where 'M_f' and "meaning$_f$" stand for his notion of functional or "non-natural meaning):

> Whether or not M_f represents any progress in our attempt to naturalize meaning (and thus understand a system's non derivative power to misrepresent) depends on whether the functions in question can themselves be understood in some natural way. If these functions are (what I shall call) *assigned* functions, then meaning$_f$ is tainted with purposes, intentions, and beliefs of those who assign the function from which meaning$_f$ derives its misrepresentational powers. We shall not have tracked meaning, in so far as this involves the power of misrepresentation, to its original source. We shall merely have worked our way back, somewhat indirectly, to *our own* mysterious capacity for representation. (Dretske 1986, p. 22)

One should recall, in looking at this passage, that misrepresentation is a crucial dimension that distinguishes indication from representation. One should also note that the difficulty we have in accounting for representational systems (in Dretske's view) is a difficulty in accounting for type III representational systems, systems whose power to represent is intrinsic. So states of such a system ought to be able to be representations in a way independent of other intentional agents.

Some states of this sort of representational system at least ought to be able to be representations in a way that does not presuppose other purposes, intentions, and beliefs of the system itself. Otherwise, we will still have succeeded only in working back "to our own mysterious capacity".

Natural meaning (indication) clearly needs to be kept separate from non-natural meaning (representation). This is essential in order to be able to try to show how, via natural functions, representation could be derived from some natural, non-intentional relation.

To return to Baker's central criticism of Dretske, the account she gives of "structuring cause" is as follows:

> (D0) A state C's having extrinsic property P_e is a *structuring cause* of some behaviour E if (i) C has been recruited to produce E and (ii) C would not have been so recruited if it had not had P_e. (Baker 1991, p.101)

Now for C to have meaning, it must (minimally) be the case that (past) C-tokens have indicated F. The additional required ingredient is that C-tokens having indicated F must be a structuring cause of C's causing M. Thus:

> (D3') A state of type C has the function of indicating F iff the following: (i) State C is of a type that indicates F, and (ii) C's indicating F is a structuring cause of behaviour [for] some behaviour E. (Baker 1991, p.107)

In the definition[3] (D3'), C's having had its tokens be F-indicators (C's having H-indicated F) corresponds to the "extrinsic property Pe" in (D0). So type C acquires the function to indicate condition F if both C has been recruited to produce E, and C would not have been so recruited if tokens of C had not been indicators of F. C's (now) being a representation of something as F depends directly on C's having H-indicated F being a structuring cause of behaviour E (= C's causing M).

Thus Baker commits Dretske to her (3). C has meaning in virtue of C's indicating F being

a structuring cause of C's causing a movement M. The "causal role" referred to in (3) is the role of (past) tokens of C indicating F as a structuring condition for C's coming to produce a movement M that is appropriate to the presence of F.

Now how do (1) and (2) complete the circle that Baker alleges? (1) connects the notions of causality and explanation. (2) then is translatable into a claim about causal role and meaning. Substituting "causal role" for "explanatory role", (2) becomes,

(2') A state C has a *causal role* in virtue of having meaning.

Circularity does not follow directly, however. What is being referred to by "a causal role" in (3) (C's indicating F structuring behaviour) has to be the same as what is referred to by "a causal role" in (2'). In the final paragraph of her circularity argument, Baker sets about accounting for this:

> We can also see that the causal role referred to in (3) and the explanatory role referred to in (2) are one and the same. The relevant role for both is the role conferred by the fact that C's indicating F is a *structuring cause* of some behavior E. Let 'S' stand for 'the fact that C's indicating F is a structuring cause of some behavior E.' Then, putting it awkwardly, by (2), C is such that S in virtue of having meaning; but by (3), C has meaning in virtue of S. (Baker 1991, p.107)

The formulations are awkward indeed, so I will translate them out in full. Substituting the exact phrase for 'S':

[From (2)]

a) C is such that the fact that C's indicating F is a structuring cause of some behavior E in virtue of having meaning

[From (3)]

b) C has meaning in virtue of the fact that C's indicating F is a structuring cause of some

behavior E

b) is unproblematic: this is the heart of Dretske's theory. The attribution of a) to Dretske is not unproblematic. Dretske certainly does not mean to claim that C's causal role in structuring the given behaviour is due to its having some *non-natural* meaning, but the meaning referred to in b) is non-natural: a representation of F's being present. Dretske does not claim that C's indicating that F is present is a structuring cause of E (C's causing M) due to its meaning "that F is present" in the non-natural sense of representing something as F.

There is an alternative here. C has a structuring causal role in C's causing M. It is in relation to the "recruitment" of C as a cause of M that C has a causal (hence explanatory) role in virtue of its semantic properties. C's H-indicating F is a partial cause of C's being recruited as a cause of M. Through the Design Solution, C (as a physiological state) thus also acquires a causal role in that C tokens come to be[4] triggering causes of M tokens. So C tokens come to *function* to indicate F in virtue of acquiring the causal role of causing M in a condition (F) that (in the ideal case) tends to have some F-relevant outcome R. On this alternative, the circle alleged by Baker does not hold since, contrary to her intended conclusion, the causal role referred to in (2) is not the same as the causal role referred to in (3). The causal role referred to in (2) is the role that past occurrences of C (in virtue of indicating F) play in structuring a behaviour type (C's causing M). The causal role referred to in (3) is the role that C plays in a case of a C-token's causing M, and the *meaning* that C has in virtue of this causal role is that of being a representation (not an indicator) of things as being F. As I understand Dretske's reply to Baker, this is in fact the tack that he takes:

> So, to say that C_n means F is to say that earlier tokens of this type indicated F and that this fact (that they indicated F) was responsible for a causal re-organization, a shaping, of control circuits so that later tokens of this type (including, of course, the current

> token C_n) would have causal duties (in the determination of motor output) they did not previously have. To attribute meaning to a token internal state is, on this account of meaning (and belief), to describe the *source* of its causal efficacy. It is to say what gave it a voice in the determination of output, what led to its installation as a control structure. To say that C_n means F is to say from whence C_n got its causal powers -- powers it is (presently) exhibiting in helping to produce M_n. It is to say that C_n acquired causal efficacy (the sort of efficacy relevant to shaping output) from earlier tokens of this type indicating (carrying information) F. This does not make the appeal to C_n's meaning (in the explanation of why it is causing M_n) circular. For the causal job being explained is what C_n is doing (causing M_n -- current behavior), and the property of C_n that is invoked to explain it,C_n's indicator function, is a property constituted by what earlier tokens of this type did. (Dretske 1991a, p.117)

The general idea is that C_n's being such that it causes M_n is accounted for by the structuring effects of past tokens of C having been indicators of F. Once the structuring has taken place, subsequent tokens of C have the function of indicating F. So H-indication (having a history of indication), via physiological entrenchment of the (type of) process "C's causing M", is the structuring cause which brings about C's having its function to indicate an external condition. Representation follows from, and is not the cause of, behaviour-structuring[5].

Note that the parenthetical comment in the last quotation, that C_n's (this being a current token of C) "acquired causal efficacy" is "the sort of efficacy relevant in shaping output", is puzzling. What is meant by this sort of efficacy? Is it the power that C_n has to cause a certain type of movement? Of course a given token of something may exist over some extended period of time, and over that time its properties may change. Does Dretske mean that a given token C_n, over some extended period of time, comes to have an output-shaping power which that given token initially failed to have? It is hard to see where this suggestion would fit in the context of Dretske's theory generally or in the context of his response to Baker. Dretske's mechanism for "shaping" causal properties, the Design Solution, only allows that a type C may acquire some causal efficacy in the

sense that current tokens have effects that past tokens did not have. Also, what of the reference to "shaping output"? Could this mean simply C_n's power to cause a token of the movement type M? This appears to be what Dretske has in mind, for the claim that the current token gets the power to cause M from past tokens of C having indicated condition F certainly fits with explication of the Dretskean theory in other of his writings[6].

To summarize, the circularity suggested by Baker requires C to be a structuring cause of behaviour E in virtue of its (non-naturally) meaning that F. However, the account provided by Dretske only claims to require that C's having H-indicated F be a structuring cause. Dretske claims that no stronger, meaning-relevant relation is required in order for C's reliably indicating F to structure the behaviour pattern E (C's causing M). His way out of the circle, then, is to distinguish between C's having H-indicated F and C's representing something as F.

Consideration of Baker's critique raises some additional questions. One question is whether Dretske can succeed in generating non-natural meaning out of a purely non-intentional indicator relation. Is indication by itself enough to bring about behaviour-structuring? If the answer is no, are the additional resources required themselves free from appeals to the explanandum (semantic content, especially representational, of mental states). Dretske's theory may commit a more general kind of circularity if the acquisition of non-natural meaning properties of mental states can only be accounted for in part by appealing to the non-natural meaning properties of other mental states. This is an issue I return to in chapter 6. The second question is one which Baker and others have pursued. If the *representational* properties of doxastic states are not active as structuring causes of behaviour, what causal role exactly are they assigned on Dretske's theory? Once representation and indication are sharply distinguished in the theory, it is not clear whether a state C, *as a*

representation, has any causally explanatory role in behaviour.

c. What is the Causal/Explanatory Role for Semantic Properties?

Recall that Baker's specific complaint with the Dretskean information theory is a dilemma. The first horn is that Dretske's theory is circular. The other horn of the dilemma is that the theory makes representational content epiphenomenal (Baker 1991, pp. 107-110, Baker 1995, p. 62). This should be a devastating objection to Dretske's view, for the promise of the theory should be precisely to give a naturalized account of intentional states which preserves the idea that these are causal/explanatory in virtue of their being beliefs and desires. To preserve this idea, one would suppose, would require showing that such states are causal/explanatory in virtue of the meaning properties which distinguish them as intentional states (i.e. representation). The "soprano problem" (see Introduction) is that some phenomenon can both be meaning-bearing and can be a cause of some effect without it being the case that the meaning is in any way causally relevant. Dretske's promise is to give an account of the (*non*-natural) meaning of intentional states such that the meaning is not like the meaning of the soprano's high "C" in relation to the breaking of the glass. So showing that on the theory intentional states had no worldly effects in virtue of their being representational clearly would undermine the very point of the account.

In addition to Baker, others have expressed concerns about the causal/explanatory role that Dretske's theory gives to semantic properties, including Kim (Kim 1991, p.65), Horgan (Horgan 1991, pp. 85-86), Noordhof (Noordhof 1996, pp.220-222), and Robert Cummins (Cummins 1991, pp. 105-106). Among these critics, it is possible to identify two different sorts of objections concerning the causal role of semantic properties in Dretske's theory. Dretske does not always

seem clearly aware that there is more than one type of challenge to his claim to have made belief and desire, *qua* belief and desire, matter. Although both of the types of criticisms are difficult for Dretske, the one which can be drawn out of Baker poses a greater problem than the other. It would be a mistake, therefore, to run the two together and simply concede "the point".

Let us begin with the criticism which is not drawn out of Baker's discussion. This sort of problem, which is an objection based on temporal currency, I will label "Objection 1":

The causal role of that Dretske's theory assigns to semantics, if there is any, is in the past. The semantic properties of a current token of C cannot causally explain the current instance of the behaviour C's causing M. The commonsense picture of explanation in terms of rationalizing reasons seems to be a case of appealing to present instances of thoughts and their semantic contents. Dretske's theory does not place the explanatory weight on these present states. So the theory does not really vindicate the "ordinary picture." Dretske is of course not absolutely committed to such a vindication, but he is fairly committed: he does not want to abandon striking "commonsense" features of mental states if he can avoid doing so.

One version of Objection 1 appeals to Dretske's view of behaviour as a reason for saying that current semantic properties cannot be causally/explanatorily relevant. The following passage is Horgan's statement of the objection:

> Consider a statement expressing a rationalizing explanation, say, "Fred went to the fridge because he wanted a beer and believed that there was a beer there." As we would ordinarily understand this statement, it is true that only if Fred's act of going to the refrigerator was *caused* by the cited belief and the cited desire. Although the existence of such a causal relation presumably is not sufficient for the statement's truth (since the properties *being a desire for a beer* and *being a belief that there's beer in the fridge* also must be explanatorily relevant to the action), it does seem to be at least a necessary condition.

> Yet Dretske is committed to denying this. For, on his view the action is a causal transaction which includes the reason (i.e., belief-desire combination) as a component part. Since the causal process cannot be caused by a part of itself, he must maintain that reasons do not in fact cause the actions they rationalize. This goes contrary to our pre-theoretic understanding of the "because" statements that typically express rationalizing explanations. (Horgan 1991, p. 85)

The force of the criticism comes into clear focus in the last couple of sentences. Horgan focuses on Dretske's conception of behaviour as a process. The important underlying concern is that Dretske's theory assigns to semantic properties an explanatory role which is importantly different from that reflected in "rationalizing explanations": Dretske's theory fails to assign the sort of "here and now" explanatory or causal relevance commonsensically assigned to beliefs and desires cited in explanations of behaviour.

It is an essential part of Dretske's view that behaviour be conceived as a process, of which internal physiological states are a part. So a present instance of behaviour is an instance of a C-token's causing an M. As the behaviour is itself C's causing M, the current instance of C is not a cause of the current instance of behaviour. It is not, after all, a cause of something of which it is a proper part. Supposing that the current C-token is a cause of the current behaviour instance is the same as imagining that the current instance of doorbell pressing is a cause of the instance of the doorbell pressing's causing a ringing. The current doorbell ringing is part of the caused instance of the process, not a cause of it. Similarly the current physiological state caused by Sniffy's being visually stimulated by a light is part of Sniffy's current bar-pressing behaviour, not a cause of that behaviour.

Of course the current physiological state is a cause of Sniffy's bar-pressing *movement*, but this has nothing to do with what Sniffy's internal state *means*. The movements are caused by the

neurophysiological processes of which the internal state is comprised. The bar-pressing *behaviour*, as opposed to the movements, is the internal physiological state's causing the movements. That is the view of behaviour upon which Dretske insists.

A second objection is that the information theory, appealing as it does to a state type's history of indication, only succeeds in giving a causal role to natural semantics. It fails to make non-natural semantics causally active. Both Baker and Cummins (Cummins 1991, p. 106)[7] offer versions of this objection.

Baker's version is as follows:

Objection 2: On the Dretskean picture, the only obvious role given to any semantic properties is the role of a structuring cause of a behaviour process. It is Dretske's view that a given internal state of a certain type acquires a non-natural meaning, a function to indicate, as a result of a structuring process. That is, C comes to cause M and would not have come to cause M if past C-tokens had not indicated F. One cannot appeal to C's non-natural semantic property of functioning to indicate F on pain of circularity: If one were to appeal to C's non-naturally meaning F, then C's functioning to indicate F would both explain and be explained by C's coming to cause M. So in order to avoid this circle, Dretske must say that it is C's being an *indicator* of F, not C's being a *representation* of F, has the causal/explanatory role in structuring behaviour.

As we now see, this proposed solution to Baker's circle is not exactly without its difficulties, though. Dretske, recall, sometimes appears to take an "inflationary" view of representation: a representational content is a distinct property resulting from a Design Solution. In light of the solution to Baker's circle, however, it is not at all clear that representation, as a distinct property, has *any* causal/explanatory role in behaviour. A behaviour is *structured* (a physiological state C

comes to cause M) because of what C has H-*indicated*. Once the structuring has taken place, is there any reason to think that a given C-token's causing an M-token is not fully explained as a physiological reaction to external condition (without reference to what C represents)? Of course, Dretske far more consistently takes a "deflationary" approach to representation. That is, his usual view is that C's (now) representing things as F just comes to C's having come to cause M in virtue of past tokens of C having been indicators of a condition F. The deflationary view is certainly in evidence in Dretske's previously quoted response to Baker (Dretske 1991a, p.117). However, if a C-token's representing that F just comes to its being of a type which causes M because past tokens of its type were F-indicators, then it still looks at bottom as if the only meaning property available to do any work is indication. "Representing F", on the deflationary view, seems to be just a way of referring back to past C-token's roles as F-indicators. There does not, then, appear to be a distinct causal role for C's being a representation (over and above the role that C has had in H-indicating).

Objections 1 and 2 are distinct. Objection 1 claims that the semantics of a current C-token is irrelevant to the current behaviour-instance. This may raise questions about how well Dretske succeeds in vindicating *ordinary* belief-desire explanation. Objection 2 says that non-natural semantic properties, past or present, are never relevant. Objection 1 only tends to suggest that semantic properties have a different role than we had previously supposed. Dretske at times appears to be happy with this. To some extent, he argues explicitly for such a view in his theory through his idea of a structuring cause and his concept of behaviour. Objection 2 is that the target properties, the very one whose causal role we have been attempting to define, do not have a distinct causal role.

Dretske apparently does not distinguish between the two types of criticisms, as can be seen in a relatively recent exchange between Dretske and Noordhof. Noordhof also takes up a version of the charge that Dretske does not succeed in assigning a causal role to representational states (qua representational states). (This is only part of Noordhof's strategy but it is the part of interest for my purposes.) We may call this criticism (after a section of Noordhof's article) the "Present Effects" argument. Dretske's response to Noordhof is especially puzzling and interesting: it is clear that Dretske is prepared to acknowledge Noordhof's charge, but he appears not to regard it as problematic. However, as I previously noted, failing to show that representational content, as such, has a causal role is to fail in what was the originally stated goal of Dretske's programme.

Noordhof takes his objection to be a variation on one offered by Gabriel Segal and Elliot Sober. He writes:

> By making the present [C-token] a part of the behaviour, Dretske seems to rule out the present [C-token's] content from being causally relevant to the behaviour in *any* sense (Segal and Sober, 1991, pp.23-4). The account he offers of the causally explanatory role of contents seems to talk only of the present effects of [C's] *previously* indicating [Fs], not the present effects of [a C-token's] *present* possession of content. (Noordhof 1996, p.220) (Lettering substitutions in square brackets are mine)

Noordhof cites a passage in which Dretske seems to respond to this point by identifying the representational content (non-natural meaning) of the present C-token with that natural meaning (information) in past occurrences of C's which explain the present C-token's causing M. (I have already suggested that such a response is problematic.) Noordhof's reply is that Dretske's response is too weak:

> What Dretske seems to be arguing is that he has shown that the content of the current [C-token] is causally relevant because [Cs] were recruited as a cause of Ms as a result of past [C-tokens] having the corresponding natural meaning. But, at best, what this shows is that he has established that something with the *same semantic properties* as

> the content that is currently tokened is causally relevant. What he has not done is establish the causal relevance of the current token of content with these semantic properties. (Noordhof 1996, p.221) (Lettering substitutions in square brackets are mine)

Noordhof's proposed problem for Dretske's theory is that only semantic properties (of *some* kind) of previous C-tokens are relevant in the explanation of the current C-token's causing a given M. This can be seen as a version of objection 1. As Noordhof says about Dretske in the immediately preceding quotation "What he has not done is establish the causal relevance of *the current token of content*..." (my emphasis). Earlier in this passage, Noordhof also gives a version of objection 2 when pointing out that Dretske is trying to get causal relevance of a current token through an appeal to the type C's being recruited as a cause of M in virtue of past C-tokens "having the *corresponding natural meaning*" (my emphasis).

Dretske's response to Noordhof reveals how the two objections come to be run together:

> If it isn't the fact that the present token of type [C] represents F, but the fact that past tokens of this type indicated F, that explains (structures) current behaviour, doesn't this show that beliefs do not explain behaviour? It all depends. It depends on whether an appeal to beliefs is understood as an appeal to the information the belief is supposed to carry or to the fact that it is supposed to carry it. According to the first way of looking at things, *my* way of looking at things, explanation of behaviour by appeal to what someone believes is a way of explaining their behaviour in terms of the influence that this information-carrying structure (type) had *when* it was doing its job, *when* it was carrying the information it is now its function to carry. This, I suppose, is merely another (grudging) way of conceding Noordhof's (and Sober's and Segal's and Baker's and Cummins' and Horgan's, etc.) point -- that, according to this theory, the content of belief is causally relevant to behaviour, yes, but not *as* the representational content of current token, but *as* the informational content of previous tokens. Appeals to a belief in explaining behaviour are, therefore, appeals to the information (carried by previous tokens of this type) that the current token has the function of carrying. They are not appeals to the fact that the current token has the function of carrying it. (Dretske 1996b, p.229) (Lettering substitution mine)

By the end of this passage, it is quite unclear whether Dretske is responding to objection 1

or objection 2. The roster of critics to whom he says he is "conceding" a point include both kinds of objectors. Thus, the "point" he concedes could be read as a concession to either type of critic. Note, for example, that Dretske says that "according to this theory, the content of belief is causally relevant to behaviour, yes, but not as the representational content of current token, but as the informational content of previous tokens." Is this an acknowledgement, then, that only previous tokens' semantic properties are relevant, that only natural semantic properties (indicatings) are relevant, or both? Dretske here also seems persistently inclined to blur the distinction which was initially of vital importance to his theory, at the heart of it in fact: representation versus indication. In the sentence of Dretske's that I re-quoted three sentences above, what is the reference of "the content of belief"?

Ambiguity that appears earlier in the passage makes the issues difficult to sort through. Consider the following sentence:

> According to the first way of looking at things, *my* way of looking at things, explanation of behaviour by appeal to what someone believes is a way of explaining their behaviour in terms of the influence that this information-carrying structure (type) had *when* it was doing its job, *when* it was carrying the information it is now its function to carry. (Dretske 1996b, p.229)

The first question one should ask is "what exactly is meant by 'this information-carrying structure'". It is not a reference to a belief. Whether a type of internal state is or is not a belief, on Dretske's theory is *irrelevant* to the state's being an information-carrying structure. An information-carrying structure is an indicator. A C-token's being an indicator does not make it a belief-state. This comes about when C gets promoted (via the appropriate Design Solution) to *the job of causing M* in the presence of F. Prior to the promotion, what (if anything) C-tokens might *represent* is irrelevant. After the promotion, what C-tokens might *indicate* is (one would have

thought) irrelevant. This is the point of distinguishing between representation (non-natural semantics) and indication (natural semantics). It is supposed to allow the semantics of belief (a kind of non-natural semantics) to be grounded in natural semantics in such a way that token belief-states can *mis*represent. One wants C-tokens to be able to have "that F is present" as their content when F is not present. That is what promotion to an indicator-function is supposed to do. So of course one must continue to be careful not to slide from indication to representation. Beliefs are not supposed to be equated with "information-carrying structures". So, under the terms of Dretske's theory, by finding a causal role for an information-carrying structure, one is not thereby identifying a causal role for beliefs (*qua* beliefs).

A bit of usage that again might encourage confusions here is Dretske's reference to the information-carrying structure "doing its job". When a behaviour process comes to be structured, the type of internal state C is described as getting "promoted" to an indicator function by being given a "job". A bimetallic strip connected to a furnace functions to indicate room temperature, activating the furnace so as to regulate the temperature. So in the previous sentence from Dretske, what job is being referred to in the phrase "the influence this information-carrying structure (type) had when it was doing its job"? Prior to entrenchment of the C-M structure, C does not (yet) have the job of causing M (under appropriate conditions). This is the job that (type) C is promoted to as a result of a Design Solution. Dretske cannot mean *this* job unless he means to refer to a history of C after that state has come to be a cause of M in the presence of F. However, what happens -- after a Design Solution -- with C's is quite irrelevant to getting any explanation in terms of semantic properties of what the current C-token does. Until C becomes part of an appropriate behaviour structure, it does not have the job of indicating that F. The phrase in question invites one to

conflate the property of indicating with the property of having the job (function) of indicating. In terms of semantics, the only *job* that Dretske's theory allows for C, when it was carrying the information that later became its function to carry, is C's structuring of the C-M process by virtue of C-tokens indicating F.

In reading Dretske's response to Noordhof, then, it is in the first place very difficult to see the appropriate distinction between indication and representation. Dretske in fact seems to slip between the two, appealing to belief in a context in which the representation-semantics of belief would be unavailable as an explanatory resource for the job that an "information-carrying structure" has. In the second place, it is difficult, through much of his reply to Noordhof, to determine what Dretske really means to concede. It is not clear whether he sees the difference between the previously distinguished objections 1 and 2.

Conceding objection 1 may only amount to conceding that the causal role that semantic properties have is different from the usual conception. This is awkward, but perhaps not insurmountable. One need only note that, contrary to the unreflective commonsense view, the semantics of intentional states play a different causal role, one that makes them historically relevant to the kinds of behaviour a cognitive being presently exhibits.

Conceding objection 2 amounts to far more: this is to concede that mental representations, *qua* representations, are causally (and therefore explanatorily) inert. The opening of Dretske's (p. 229) passage suggests that there is a simple difference in viewpoint between Dretske and his critics over what causal/explanatory roles of beliefs could be. Whether or not an explanatory role for belief survives Noordhof's objection supposedly "depends on whether an appeal to beliefs is understood as an appeal to the information the belief is supposed to carry or to the fact that it is

supposed to carry it". Tucked in this phrase is the distinction between indication (information-carrying) and representation (functioning as an information-carrier). The terms of success for Dretske for giving belief an explanatory role is quite clear: Belief must be given a role *qua* belief. To give belief an explanatory role *qua* belief is to give it a role in virtue of its semantic content. A belief's semantic content is a representation. Therefore to give belief, *qua* belief, an explanatory role a theory must make it relevant in virtue of its being a representation. If the theory gives no role to representation then, no matter how one looks at the issue, the theory fails to make belief *qua* belief explanatorily relevant.

Having previously expressed disagreement with his critics' viewpoints, and having been systematically ambiguous over which issue he is in fact addressing, Dretske then does appear in closing to make the concession to objection 2:

> That is as close as I could get to non-natural meaning doing some causal work in the world. Its work is done by the natural meaning out of which non-natural meaning is created. (Dretske 1996b, p.229)

Critics, as far as I can see, are not simply missing a point. At least some are alleging that Dretske does not succeed in the task with which he is originally confronted. The original task was to show how a belief-state, in virtue of its a *representational* content, could be a cause of behaviour. What Dretske appears to be suggesting here is that it cannot be. So the account fails to do what we should expect of a representational theory of intentional states: it fails to make those states causally explanatory in virtue of their representational contents.

d. Does Natural Meaning Have a Causal/Explanatory Role?

The next step is to ask whether Dretske's approach at least delivers on this modified promise of making *natural* meaning relevant (in virtue of a causal role) in behavioural explanations. Natural meaning, if relevant, is relevant as a structuring cause of behaviour (according to Dretske's theory). For natural meaning to be relevant, then, C's H-indicating F and C-tokens being *indicators* of F need to be causally explanatory in the structuring of behaviour. (To remind, a state C of a system (organism) S H-indicates a condition F over some period of time t if and only if, over t every token of C in S indicates that something is F. H-indication over a period of development is what Dretske requires in order for the state C in S to come to be an (active) representation of condition F. S must come to do M in condition F, and it must be the case that S would not have come to do M in condition F if C had not H-indicated F (over the behaviour-structuring period). Dretske needs indication and H-indication to be effective in operant learning. I will contend, however, that indication and H-indication are not in general requirements of operant learning, and so cannot be cited as structuring causes of learned behaviour. (This is point (4) from the introduction to this chapter.)

The initial problem here is that Dretske's satisfaction conditions on his notions of indication and H-indication are very strict. In fact, they are much stricter than is required for, say, a correlation that could produce successful results in cases of operant conditioning. This is a point which is made very effectively by Slater in a recent article. After reminding us of Dretske's stated goal of giving an explanatory role to mental meaning, she introduces her concern:

> Dretske proposes that what he calls *functional meaning* or *meaning*$_f$ meets this demand for explanatory relevance (Dretske, 1988a). I shall argue that whatever its theoretical

> merit in this regard, we have been given no reason to believe that functional meaning is anywhere actualized in natural systems. Dretske's confidence that it must be so stems from a conviction that relations of indication explain changes that constitute actual discriminative learning. I shall argue that this conviction is misplaced. There is no reason to think that indication, as Dretske understands it, can legitimately be invoked to explain changes in the causal hook-ups that mediate sensory inputs and motor outputs of actual organisms. If I am correct, Dretske cannot both insist upon explicating functional content in terms of a preferred notion of indication and, at the same time, depend upon discriminative learning to guarantee the instantiation of such content. (Slater 1994, pp. 163-164)

Why does this suggest a problem? The developmental design solution is based on behaviour being acquired through operant conditioning. However, Dretske's strong take on indication drives a wedge between those conditions required for operant conditioning to occur and those conditions for the satisfaction of indication. It does so in a way that allows that behaviour structures can (and regularly do) form in the absence of an indication relation. This in turn will lead me to question whether indication (or H-indication) is *ever* causally necessary in the structuring of behaviour. This is a crucial point, for it is after all Dretske's view that for C to represent something as F, it must have come to cause M *because C has H-indicated F.* This is where semantic properties are supposed to gain their relevance. It cannot merely be the case that, through operant conditioning an organism's state C comes to cause M in the presence of F and, as a matter of fact, C happened to H-indicate F during "structuring". C-tokens having been indicators has to do the work.

My final point differs from the way in which Slater draws out the implications of her points. She only concludes that even a weakened version of Dretske's indication relation (this will be explained presently) is in general neither necessary nor sufficient for behaviour modification. Dretske, in responding to Slater, makes important concessions, including that Slater is right in suggesting that operant conditioning does not in general require indication. Yet, he takes her

criticism to be answerable by saying that indication must be satisfied in order for operant conditioning to bring about concept formation. I will argue that Dretske's response is inadequate precisely because he needs indication to do the causal work in some cases of operant conditioning and his answer does not show that indication does do this. I will proceed by outlining Slater's objections, then by showing how my own criticisms extend from these.

One very nice feature of Slater's discussion is that it brings out the fact that there are two distinct sources of extraordinary strength built into Dretske's notion of indication, relative to simple correlation. I will call these:

(1) The necessity requirement for a C-token to be an indicator: a C-token does not indicate F unless occurrences of C are dependent on F such that, within the normal constraints of a given environment, a C-token would not occur in the absence of a token of F. So, not only must an F-tokens be present when the C-token occurs, but, in some sense of necessity, the C-token must bear some dependency on the presence of F. The necessity Dretske requires turns out to be a fairly limp sort[8].

(2) The (stronger) "learning requirement" for C to be recruited as an F-representer: During the learning period, while C is being recruited as a cause of a movement, there cannot be C-tokens which in fact occur in the absence of F. That is, during the learning period C must H-indicate F.

Slater explains Dretske's necessity requirement as follows:

> According to Dretske, X's indicating Y is a nomic relation, one that amount to what he has elsewhere referred to as X's carrying information about Y. For C to carry information about or to indicate F, Dretske holds, there must be more than a merely 'accidental' coincidence between them: in the absence of 'objective constraint', even perfect correlation between C and F does not constitute indication (Dretske, 1986, pp. 18-19). (Slater 1994, p. 168. The Dretske reference included in this passage is to his paper entitled "Misrepresentation".)

This requirement of non-accidental dependency, as well as the learning requirement have already been developed in my explication of Dretske. It is not difficult to find passages which commit Dretske to either of these requirements on indication, and in fact he does not deny either of them in his response to Slater.

Slater's specific criticism, however, rests on holding Dretske to the view that indication is necessary (generally) for operant conditioning. It is this generalization which Slater is anxious to attack. She initially holds Dretske to a "strong" argument for the role of indication which she supports with pieces of Dretskean text, then arguing that even a compromised view of the role of indication will not work. Here is the "strong" argument to which Dretske is supposed to be committed:

> The argument just canvassed from the claim that (1) indication relations are necessary for operant discriminative learning, to the claim that (2) indication therefore plays an explanatory role in operant discriminative learning, to the conclusion that (3) the existence of operant discrimination learning guarantees the realization of meaning$_{I}$. Let me call this Dretske's strong argument. My first goal will be to argue that the strong argument is unacceptable because indication, as Dretske understands it, cannot play the explanatory role required by (2). (Slater 1994, p. 167)

Her reason for rejecting (2) has to do with the necessity requirement: organisms are not sensitive to the necessity or contingency of correlations. Suppose that a token of C occurs in the presence of condition F. Whether or not that token is an F-indicator depends on a probability assignment: if the probability of F being present given a C-token is less than one, the C-token still fails to be an F-indicator. The distinction between a given C-token which indicates and one which just coincides with F is not a distinction to which organisms are sensitive. So the particular feature, the necessity of C's occurring in the presence of F, is not something which could explain C's coming to cause M.

Slater's next step is to consider a modified view in which there is an "indication" relation which is not grounded in necessity (C-tokens simply never occurring in the absence of F) such that, when a) "non-necessary indication" is satisfied, b) it happens to be the case that C-token's dependence on F is nomic, and c) C is recruited as a cause of M, then C still comes to represent F. The idea is that the original necessity-based indication plays a "vicarious" role. It is simply the exceptionless correlation of C and F that does the conditioning, but the existence of the exceptionless correlation happens to be explained by a nomic dependence between C and F. So by a principle of "explanatory transitivity" which Slater invokes for the sake of argument, C-tokens indicating F explains the behaviour structuring because C is exceptionlessly correlated with F (and the exceptionlessness of the correlation is explained by C-tokens' dependence on F).

The "non-necessary" form of indication is labelled "indication*". This weakened relation is then used to define a possible "principle of operant conditioning:

> [I*] A discriminative response M to an environmental event F can only be acquired if M is put under the control of some internal state C that indicates* F. (Slater 1994, p. 170)

[I*] still must meet the learning requirement: there cannot be instances of C occurring in the absence of F (although this can now be a matter of contingency). So Slater proceeds to argue that this is also too strong a requirement for operant discriminative conditioning. She does so mainly by appeal to (well known) empirical studies which show that what is required in order for a correlation to condition behaviour depends on factors other than mere strength of a correlation or the absence of counter-examples. The point is made forcefully by Slater's appeal to studies in which operant conditioning is shown clearly to be successful in the absence of principle [I*][9]. However, commonsense observation really ought to be sufficient to conclude that conditioning a

subject to a given stimulus does not require the reinforcing condition to be introduced every time the stimulus is present. It is obviously unlikely that Dretske would think that operant conditioning in general required perfect correlation between a stimulus and a reinforcing condition.

Slater's central point is that some property independent of indication* is relevant to the conditioning of behaviour[10]:

> Empirical studies of learning leave little doubt that a crucial factor in establishing control over a response is, indeed, contingency -- what Egon Brunswick called the *ecological validity* of a cue and what has elsewhere been referred to as the *predictive* or *inferential value* of a stimulus... Contingency is often indexed by a statistical measure of association whose value increases as the probability of a behaviorally relevant state of affairs given the presence of a signal exceeds its probability in the signal's absence. In an influential early study, for example, Rescorla found that rats' acquisition of a classically conditioned emotional response (CER) to a tone (CS) that was accompanied by shock (US) diminished dramatically as the probability of shock in the presence of the tone -- p(US/CS) -- came to differ less and less from the probability of shock in the absence of the tone -- p(US/~CS). (Slater 1994, pp. 174-175)

The upshot of this kind of example is that the simple frequency of shock-tone pairings is not the determining factor in establishing the CER. So a stronger simple correlation between a shock and the sounding of a tone is not necessarily more effective in conditioning the rats. Rather, what is important is comparative frequency. If the frequency of shocks accompanied by tones is higher than the frequency of shocks unaccompanied by tones, then the rat will become conditioned to the sounding of the tone.

This seems a little more complex than is really necessary. Surely one need only note that conditioned responses can certainly be established with less-than-perfect correlation. A rat, shocked often enough (but not every time) in the presence of a tone, and shocked much less frequently in the absence of the tone, will acquire the CER. In a summary of her description of

Rescorla, she does make this kind of point:

> It would appear that M comes under the control of an external C only when some behaviorally relevant state of affairs, F, is more probable given C than in its absence. Contingency can, however, hold in the absence of indication*: even if the probability of shock given a tone falls far short of 1, there will be an association between shock and tone if the probability of shock in the *absence* of the tone is smaller yet. Contingency in the absence of indication* is perfectly capable of supporting learning (Rescorla, 1988a, 1988b). (Slater 1994, p. 176)

Observation of operant conditioning in action thus suggests that conditioned responses can be learned even when Dretske's learning requirement of exceptionless correlation between C-tokens and F is not satisfied. Principle [I*], to put it Slater's way, does not appear to be a necessary principle of operant learning. The necessity of principle [I*] was the remaining way of making the original (necessity-based) version of indication explanatorily necessary (by a kind of "explanatory transitivity). Therefore, Dretske's original notion of indication fails even in having an explanatory role through [I*]'s having an explanatory role.

Dretske's response is to deny that he means indication to be necessary in general for operant conditioning. He wants to claim only that it is required in order for operant conditioning to result in "concept formation", or, put in more familiar terms, formation of non-natural meaning. So the claim is that operant conditioning can take place perfectly well without indication, but this does not result in a representational content. However, if operant conditioning takes place, and C comes to cause M when F is present, and both the necessity and the learning requirement are satisfied, then and only then does C come to be an F-representation.

I do not think this meets the most important issue. By noting that correlations, even imperfect ones, can ground behaviour formation by operant conditioning, Dretske opens the way to question the explanatory job which C-tokens indicating F is supposed to do on his theory. There

is just no reason to suppose that, even when H-indication is satisfied, that C-tokens' *indicatings* of F are causes of the resulting conditioned behaviour. I will describe a set of simple imaginary examples to illustrate.

Imagine the following: I want to train Kelly the weasel-hound to chase weasels that come around the chicken-yard. The chickens panic and squawk loudly when they sense a weasel. Since weasel-chasing is something that Kelly likes doing, a piece of behaviour that led to weasel-chasing would be reinforcing. So my job is to get Kelly to perform the appropriate behaviour when the chickens raise the alarm. Now let us imagine three variations:

Case 1: Nothing around here makes chickens squawk the way that the presence of a weasel makes them squawk, and nothing around here sounds to Kelly the way that chickens sound when spooked by a weasel[11]. So when the ruckus starts, the probability of a weasel being present is 1. During Kelly's "training period", I run Kelly out to the chicken-yard 50 times and each time Kelly finds a weasel to chase. After this period, Kelly does not need me to show her out to the chicken-yard. In this case, indication is satisfied. Moreover, Kelly has acquired a new "behaviour structure": she runs out to the chicken-yard when a certain type of physiological state, the kind which during training was caused by chicken-squawkings, occurs. I have tried to tell this story in such a way that conditions for a physiological state of Kelly's to indicate the presence of a weasel near the chicken-yard are satisfied. If I have succeeded, and if one accepts Dretske's view, the upshot is supposed to be that this structuring is explained by natural semantics: what Kelly's chicken-squawking detector indicates (the presence of a weasel).

Case 2: Something around here besides weasels make chickens squawk the way weasels make them squawk. Chickens do not have a very powerful sense of smell. They detect weasels

by means of smell, but they cannot tell the olfactory difference between a weasel and a mink. Down the road lives Farmer Bloggs, who happens to be a mink farmer and so of course smells like mink. Bloggs likes to take walks cross country when he can, sometimes passing by my chicken-yard. Except for this change to the story imagine all other details of story are the same as case 1. I train Kelly in the same way. Fifty trials yield fifty weasels. The only difference is that now occurrences of the auditory state C in Kelly that is caused by the chickens' "weasel-is-present" squawking are not weasel indicators because the probability of a weasel given a C-token in Kelly is not 1. Even though there was a weasel in each of the trials, it could have easily turned out that Farmer Bloggs had been passing too close to the chicken-yard and this caused the chickens to sound off as if there were a weasel (causing a C in Kelly in the absence of a weasel). That this did not happen during Kelly's training was accidental.

However, if there is any reason to think that Kelly could be trained according to the details imagined in case 1, there is equal reason to believe that Kelly could have been trained still in case 2. However, in case 2, the necessity requirement on indication is not met. So if training of the kind imagined in case 1 is possible, it does not depend in case 2 on tokens of the auditory state C in Kelly being weasel-indicators. Hence there is no reason, even with respect to case 1 (where the necessity requirement is satisfied) to claim that this Kelly's C-states being *indicators* of weasels helps to explain the structuring of Kelly's behaviour. Put in terms of my example, Kelly does not care about the modal strength of the correlation.

One could also imagine the following variant, **case 3**:

Everything is the same as in case 2, except that during Kelly's "training period", I run Kelly out to the chicken-yard 50 times and 38 times Kelly finds a weasel to chase. In spite of the twelve

disappointments, when Kelly finds only Farmer Bloggs, whom she can not chase, Kelly nevertheless acquires the desired behaviour.

In case 1, both the necessity and learning requirements are satisfied. There is a (locally valid) dependency that accounts for the correlation. In case 2, only the learning requirement is satisfied. There is an exceptionless but accidental correlation. In case 3, neither H-indication requirement is satisfied. There is an accidental correlation with exceptions. Between case 1 and 2, there is nothing which could conceivably make a difference to Kelly: the experiential difference between a series of exceptionlessly rewarded accidences and exceptionlessly rewarded instances of a necessary relation are nil. Dretske, responding to Slater, writes as follows:

> So, she asks, how can the fact that A indicates B (requiring as it does, some type of nomic dependence between A and B) play an explanatory role in learning as it is supposed to do on my account of concept formation? A mere (non-nomic) constant conjunction between these event can have the same effect.
>
> This is Hume's problem, and I do not have a solution to Hume's problem. If laws (nomic relations) are something more than contingent relations (of constant conjunction) what more could they be and how could we ever know there were any? All the evidence we have concerning the nature of the relation between A and B consists of evidence available in the actual world while differences between nomic and universally contingent relations are evinced in other worlds -- *counterfactual* worlds. It is quite true that if indication is a modal relation, then learning of the sort I describe will not respect the difference between the effects of contingent and necessary co-occurrence of F with C. (Dretske 1994b, pp. 203-204)

Dretske agrees on three important points with Slater:

a. Indication is modal[12],

b. Whether or not indication conditions are satisfied is a matter that is inaccessible to a subject of conditioning,

c. Operant conditioning can occur even when the necessity requirement is not met.

Despite agreeing to these points, Dretske draws a conclusion that is somewhat mystifying.

He claims that Slater's observations leave him right where he wants to be; and he claims the observations leave him with "...indication explaining only some cases of discrimination learning". (Dretske 1994b, p. 204)

Possibly the argument as developed by Slater permits such a view. However, I would argue that the position in which he is left is *not* that indication (or H-indication) only *explains* some cases of discrimination learning. Rather, the position in which he is left is that indication is only *satisfied* in some cases of discrimination learning. It remains to be explained why one should think that indication *per se ever* explains cases of discrimination learning. What is it about my case 1 that explains Kelly's learning to run out to the chicken-yard when she hears a characteristic type of chicken-vocalizing? Why not (comparing cases 1, 2, and 3) appeal to Mill's method of agreement and say that it is the correlation between past instances of Kelly's auditory state C and the reinforcing condition of weasel-sightings? This "constant conjunction" is present in case 2, and Kelly learns to chase weasels. The similarity between case 1 and case 2 is the correlation between C-tokens and weasels.

That there are, as a matter of fact, behaviour modifications without even perfect correlations (actual cases like case 3) only strengthens the case for correlation, and not indication, always being the explanatory mechanism in discriminative operant conditioning. In terms of my own simple example, Kelly learns to run out to the chicken-yard when she hears the chickens squawking because she's been conditioned to expect weasels and Kelly loves to chase weasels. In case 3, the learning came about under conditions such that Kelly did not find a weasel every time, but the pay-off is high enough that she learns to respond to the chickens' squawking anyway. It is the correlation between hearing the chicken noises and the subsequent weasel-findings which does the

work.

In summary, my argument is as follows:

Premiss 1: There are in fact examples of behaviour-structurings via operant conditioning which fit the salient conditions of each of cases 1, 2 and 3 (1 -- H-indication conditions are satisfied, 2 -- the necessity requirement is not satisfied, 3 -- neither the necessity nor the learning requirement is satisfied).

Premiss 2: In all three types of cases, what is common is some relative correlation between tokens of an internal state and a reinforcing outcome.

Premiss 3: H-indication is not satisfied in cases 2 and 3, but there is a correlation in all three cases between Kelly's auditory states and weasels.

Premiss 4: The correlation between auditory states and weasels is what explains Kelly's new learned behaviour in cases 2 and 3.

Therefore: Correlation explains Kelly new learned behaviour in case 1.

There is a further consideration here, for one might conceivably claim that, as long as there are cases in which behaviour-structuring comes about via operant conditioning and the indication-conditions are satisfied, C's having a history of indicating F still has some crucial role. Possibly, this is the move which Dretske does try at this point. He claims that, while indication is not required for operant conditioning, it is required for operant conditioning to result in acquisition of concepts:

> I was not, however, interested in discrimination learning in general, but a certain kind of discrimination learning -- concept learning -- the kind of learning in which an organism acquires the capacity to hold beliefs of the form: this is an F. For this kind of learning indication is necessary. The question is not whether some internal C can come to cause M in conditions F without C indicating F. Yes it can. All that is

> necessary is an appropriate contingency between C (the internal cause of M) and F, and this can exist without C indicating F. The question, rather, is whether C can come to *mean* that F exists without C indicating F during the process in which it comes to mean this. Contingency between C and F may be enough for an animal to learn to respond with M in conditions F, but is it enough for meaning, for C (the internal cause of M) to come to represent something as F? (Dretske 1994b, p.204)

This response will not do. All along, Dretske's goal has been to give mental states an explanatory role in *behaviour* in virtue of their semantic properties. It turns out that the semantic property which is supposed to have a role on his theory is a natural meaning property, namely, indication. Thus, the semantics of Kelly's current perceptual belief-state, the one caused by chicken-squawkings is relevant in virtue of being a type of state, past instances of which have indicated weasels. What this natural semantic property of being a weasel-indicator is supposed to explain is the structuring of Kelly's behaviour such that Kelly goes on a weasel chase when state C is tokened. However, that operant conditioning can occur without indication, but cannot occur without "contingency" (to use Slater's term), suggests that it is the latter which explains behaviour-structuring even when indication conditions are satisfied. As I have argued, there is no reason to invoke indication (or H-indication) as an explanatory principle in operant conditioning even when it occurs during a case of conditioning.

On Dretske's theory, representation or other non-natural semantic properties have no explanatory role of their own. This is a point that came clear in section c, a point that appears to be conceded by Dretske himself in his response to Noordhof. So appeals to beliefs and desires could be genuinely explanatory only provided that the natural meaning property of indication had a causal role. Its causal role was supposed to be in relation to behaviour-structuring. Now Dretske appears to have shifted to the suggestion that indication's causal/explanatory role is as a partial

cause/explanation of acquired concepts. By concepts, he clearly means "elements of non-natural meaning". So how are mental states such as beliefs supposed to *explain behaviour* in virtue of what they mean? The current non-natural meaning of a current token of a belief-state does not cause anything. The natural meaning it has (in virtue of past tokens having been indicators) evidently only explains that state's having a certain non-natural meaning. So an appeal to belief is not, *qua* belief, explanatorily relevant for behaviour.

There is a further problem in defending this weaker view on which indication (or H-indication) is a requirement for representation without its being causally relevant in structuring behaviour. The problem is that on such a view there is no basis for claiming that the content of a representation is specific in a way that the "content" of an indicator is not. As noted at the end of chapter 2, indication and H-indication are quite non-specific. If C H-indicates F over some period, and there is a dependency of F on G such that, over that period, every F-token indicates an instance of condition G (F H-indicates G, in other words), then C H-indicates G as well. So state C could not be promoted as a representation of F without being promoted as a representation of G as well. However, the point of distinguishing indication and representation is that the latter is supposed to be specific in just this way: that a system could acquire a state which represents things being F without necessarily acquiring a state which represents things being G even if the dependency of that system's internal states on F is equal to its dependency on G. The way in which this is supposed to be possible is through the way in which C gets "promoted" as a cause of M. If its being promoted is *causally explained* by C H-indicating F, then C is (subsequently) an F-representation. If C would not have been promoted had it not H-indicated G, then it is (subsequent to promotion) a G-representation. To remove the requirement that indication be part of the causal explanation

for C's coming to cause M is to remove the possible means in Dretske's theory of accounting for specificity of representation. Dretske has set out his theory in such a way that indication, C H-indicating F or C-tokens being indicators of F, needs to do the work of structuring an organism's behaviour. I have argued that it does not do this work.

e. Are We Swamp-Persons?

Dretske's strong notion of indication as the grounding for representation leads to an uncomfortable consequence beyond that which has already been mentioned: on this interpretation of Dretske's theory, we (cognitive beings) are far more like Swamp-persons than we could have imagined.

"Swamp-person" is an allusion to a thought-experiment character called Swampman, originally introduced by Donald Davidson[13]. The issue of Swampman-like characters and their implications for theories such as Dretske's which treat mental semantic properties as historical has received quite thorough discussion, notably in a forum to which Dretske himself contributes[14]. Critics raise some rather unpalatable consequences for Dretske using the example, the central one being that in theory a being could exhibit all the (apparent) synchronic behaviour that I now exhibit while failing to have any of my intentional (or other mental) states. Dretske responds by accepting the consequences and questioning their unpalatability. The literature on this topic is vast and I do not intend to treat the whole issue. What I will do is suggest, in light of my discussion of the explanatory role of indication, that on Dretske's view it may turn out that we are not so very different from Swamp-persons.

A Swamp-person is a spontaneously produced being that is a molecule-for-molecule duplicate of some simultaneously destroyed existing intentional being. From the outset of its

existence, Swamp-person will exhibit all of the apparent behaviour the "replaced" intentional being (RB) would have exhibited under the same conditions. Yet, according to any theory on which representational contents (and particularly causally active intentional ones) are external, Swamp-person has none of RB's mental contents. So Swampman has none of RB's mental state. So, the argument concludes, given an externalist account of content, a being could do everything that some intentional agent RB does without having any of RB's mental states.

Davidson's original descriptive passage is as follows:

> Suppose that lightning strikes a dead tree in a swamp; I am standing nearby. My body is reduced to its elements, while entirely by coincidence (and out of different molecules) the tree is turned into my physical replica. My replica, The Swampman, moves exactly as I did; according to its nature it departs the swamp, encounters and seems to recognize my friends, and appears to return their greetings in English. It moves into my house and seems to write articles on radical interpretation. No one can tell the difference. (Davidson 1987, p. 443)

Davidson goes on to claim that Swampman is in none of these intentional states:

> But there *is* a difference. My replica can't *re*cognize my friends; it can't recognize anything, since it never cognized anything in the first place. It can't know my friends' names (though of course it seems to) it can't remember my house. It can't mean what I do by the word 'house', for example, since the sound 'house' it makes was not learned in a context that would give it the right meaning -- or any meaning at all. Indeed, I don't see how my replica can be said to mean anything by the sounds it makes, nor to have any thoughts. (Davidson 1987, p. 444)

The view that Davidson expresses here on Swamp-person replicas is a view that Dretske shares[15]. After all, it follows from Dretske's analysis of semantic properties of intentional states as historical properties that nothing lacking the history appropriate to the acquisition via indication of a given intentional state could ever be in that state.

Some have concluded from the Swamp-person case and the claim that such a being would not have (initially at least) mental states that such states cannot be causally explanatory. The

reasoning is simply that if a being that is my physical duplicate could behave just as I behave while lacking all of my mental states, then those mental states must be irrelevant in explaining behaviour. This is a view that is taken by Steven Stich, at least with respect to psychological explanation of behaviour, for example:

> Suppose that someone were to succeed in building an exact replica of me -- a living human body whose current internal physical states at a given moment were identical to mine at that moment. And suppose further that while fast asleep I am kidnapped and replaced by the replica. It would appear that if the crime were properly concealed, no one (apart from the kidnappers and myself) would be the wiser. For the replica, being an exact copy, would behave just as I would in all circumstances. Even the replica himself would not suspect that he was an imposter. But now, the argument continues, any states or processes or properties which are not shared by Stich and his identically behaving replica must surely be irrelevant to psychology. (Stich 1983, pp. 165-166)

The reasoning is little more complicated than it may initially appear, for some states, processes, or properties not shared by Stich and the replica are relevant to some aspects of some behaviour. Performance of some action may hinge on legal status, being the husband of a particular individual or having ownership for example. To capture the point he is after, therefore, Stich defines a notion of "autonomous behavioural description" as "any description of behavior which satisfies the following condition: if it applies to an organism in a given setting, then it would apply to any replica of the organism in that setting" (Stich, p. 167). Stich then argues that psychological explanation is only explanation of behaviour under autonomous descriptions. If he is right, we have the following argument:

Premiss 1. Intentional explanation is a type of psychological explanation of behaviour.

Premiss 2. Psychological explanation of behaviour is explanation of autonomously described behaviour.

Premiss 3. States or properties or processes not shared by an organism and any replica are irrelevant to autonomously described behaviour.

Premiss 4. Intentional states are not shared by an organism and any replica.

Conclusion 1. Therefore, intentional states are irrelevant to autonomously described behaviour.

Conclusion 2. Therefore, intentional states are irrelevant to psychological explanations of behaviour.

Dretske agrees with a key premiss, premiss 4, but disagrees with the conclusions that Stich draws from it:

> It should be clear why I reject Stich's autonomy principle and his replacement argument (1983, p.165) against the relevance of the intentional explanations of behavior. A physical duplicate of an intentional agent, though it behaves the same, does not *yet* (not until it acquires sufficient experience to give the internal indicators requisite functions) behave that way for the same reasons. Although physically indistinguishable systems will behave the same way (C will cause M in both), there is no reason to suppose -- and if they have different *histories* every reason *not* to suppose -- that the explanation of *why* C causes M, of *why* they behave that way must be the same for both. (Dretske 1988a, p.105)

In the context of the present discussion, what Stich and Dretske disagree about is the implications of a thoroughly relational conception of (mental) representational content. Dretske takes this to be compatible with such content being explanatorily relevant even though a being and a physical replica of that being need not share representational contents. Stich disagrees that there could be such a compatibility.

What is important here is not whether Stich or Dretske is right, but only that Dretske's commitments be made clear. It is in his view unequivocally necessary that a system have the right history in order to be a possible subject of some mental state. It is also his view that a system may

share the same autonomously described behaviour with some replica, that an appeal to intentional states may well genuinely explain that being's behaviour even though the replica's corresponding behaviour cannot be similarly explained. A fully formed replica of me, a Swamp-Brent, could exhibit all of the outward signs, including verbal, of being intentionally driven, and driven by the same types of intentional states that apparently account for my behaviour, while having either no intentional states or none that correspond to mine. There is a very definite intuitive unpalatability in a theory of intentional states which has this as a consequence.

Dretske, however, is prepared to defend this as a consequence of what he takes to be the correct view of functions. It is not merely representational states which have this feature, but any functional state, process, or system. Thus has Dretske argued that a spontaneously created duplicate of his Toyota Tercel has none of the functional components of a Tercel or even an automobile[16]. It would seem that intentional states are at least in good company. Neither does Swamp-Brent, on Dretske's historical view of functions, have circulatory, respiratory, or reproductive systems.

Initial intuitions tend to run counter to that which is implied by Dretske's view of functions. He does offer reasons why others ought to put aside initial intuitions, though: different environmental pressures may give rise to the same apparent feature of a type of organism. That the environmental pressures which gives rise to it are different would suggest that in each case where there is a different source-pressure there is a different explanation of what gave rise to it. Dretske illustrates this point with an example of the Scarlet Gilia plant's behaviour of changing blossom colour in mid-summer:

One explanation, the one given by Page and Whitham (1985), the botanists from whom

> I take this example, is that the plant does this in order to attract pollinators. Early in the season hummingbirds are the chief pollinators, and hummingbirds are more attracted to red blossoms. Later in the season the hummingbirds migrate and hawkmoths become the principal pollinators. Hawkmoths prefer white blossoms. The flower changes color "in order to" exploit this seasonal alteration in circumstances. (Dretske 1995, p.155)

Dretske asks us to imagine a molecular twin of Scarlet Gilia (Twin Plant), evolved in a different environment, which exhibits the same colour-changing tendency:

> Consider now, a physically indistinguishable plant, one that is the molecular twin of the Scarlet Gilia. The twin evolved in a quite different environment, an environment in which, in the midst of its flowering season, rapacious beetles arrived that were attracted by the red blossoms. As a result of this selectional pressure, a slow change occurred: the plant evolved into a form in which it changed color, from red to white, in the midst of every flowering season. The beetles hate white blossoms. (Dretske 1995, p.157)

The moral drawn from this example:

> Although Twin Plant is physically identical to Scarlet Gilia, and although it therefore behaves in exactly the same way, the explanation of why it behaves in exactly the same way is quite different. Scarlet Gilia changes color to attract Hawkmoths. Twin Plant changes color in order to repel beetles. You could study these two plants under a microscope for years, know everything there was to know about their present physical constitution, and never realize that there were quite different "reasons" for their behaving the way they do. (Dretske 1995, p.157)

This appears to be a convincing example of the same behaviour (colour-changing being caused by some internal physical state) which in each case is grounded in a different functional property. In Scarlet Gilia's case, the internal state which causes colour-change in the blossoms has the function of attracting hawkmoths. In Twin Plant's case, the same type of internal state, causing the same colour change, has the function of discouraging beetles[17].

With respect to developed (as opposed to naturally selected) behaviour, Dretske's commitments certainly lead to trouble. First, the right (causal) history is absolutely required (as

we see in the previously described selectional case) for a system to have a particular function (or any function at all). C will not come to be an F-representation without the right history of C-tokens being indicators of F over the behaviour-structuring (learning) period. It *has* to be the case that C having H-indicated F is part of the explanation of why C came to cause M. In Dretske's view, an intentional explanation of why S (now) exhibits M, is an explanation of the form because S believes that s is F, wants R and believes that doing M when F is present will bring about R. The legitimacy of this intentional explanation depends on the explanatory efficacy of C having H-indicated F.

Second, Dretske is committed to the necessity of a strong notion of H-indication, and a strong notion of indication. The satisfaction requirements of H-indication (that every C-token be an F-indicator over the developmental period) is too strong to allow the notion of H-indicating to be a principle of operant conditioning. So the cases in which operant conditioning takes place and where the conditions of H-indication are satisfied is likely to be very small relative to the cases of operant conditioning in general. Therefore, one need not appeal to far-fetched thought-experiments in order to generate unappealing consequences for Dretske's theory. Each of us is going to be "Swamp-person-like" with respect to portions of our own behaviour. If, as seems very likely, there is a substantial gap between operant conditioning in general and operant conditioning with H-indication, that portion of behaviour which is "Swamp-person-like" will similarly be substantial. If my extension of Slater's criticism, that H-indication is never explanatory, is right, then we will be "Swamp-person-like" with respect to all of our behaviour.

There is a tempting, fast argument using Swamp-persons. This argument, however, is answerable. The "fast" argument is as follows:

Premiss 1: A Swamp-person either has initial intentional states sufficient to account for autonomously described behaviour or such states are irrelevant to (psychological) explanations of behaviour.

Premiss 2: Intentional states are relevant to psychological explanations of behaviour.

Premiss 3: Dretske's theory denies that Swamp-persons have any initial intentional states.

Conclusion: Therefore, Dretske's theory cannot be correct.

It is very difficult to see how to defend the first disjunct in premiss one without resorting to some sort of appeal to intuition. Dretske argues that an intuition, the internalist intuition, that would ground this claim is in fact a brute intuition. Instead of defending the "fast" argument, I will offer some disagreeable consequences generated by Dretske's commitments which do not depend on premiss 1 or variants.

The two elements of Dretske's view, the strong satisfaction requirements for indication and the necessity of the right history being explanatory for the formation of intentional contents, generate the follow consequences:

1. Probably most of my behaviour is merely "pseudo-explained" by appeals to intentional states,
2. Whether behaviours might be genuinely explained is a matter uncovered only by close scrutiny of the conditions in which my behaviour has been conditioned,
3. Whether my behaviours might be genuinely explained by intentional appeals is a matter which is not in general accessible to me,
4. A proposed explanation of a subject's behaviour in terms of intentional states and based on present observed behaviour is likely to be false. It is certainly at any rate unreliable.

These four consequences cut into the issue at the heart of Dretske's project: how to show appeals to intentional states can be explanatory in virtue of being intentional states. Consequence 1 says that, at least to some significant degree of probability, intentional states are explanatory in a minority of cases. They are certainly not in general explanatory of our conditioned behaviour even in cases in which a belief-desire explanation would be intuitively plausible. Here is a possible example:

My tennis partner has generally served wide from the "deuce" court when she bounces the ball an extra time (four instead of three) before serving. Suppose further that even while I was acquiring the compensatory behaviour, she occasionally served to the centre of the court, and also occasionally served wide even when only bouncing the ball three times. Nevertheless, I acquire the habit of leaning to the wide side in expectation of a wide serve when I hear the "pok-pok-pok-pok" of the four bounces. Let us also suppose that I play tennis with no one else and that I am not aware of that it is the hearing of the four bounces that inclines me to lean in to the wide serve. I believe that she will serve wide and that is what explains my leaning, or so a reasonable "commonsense" explanation might go. However there is nothing that functions to indicate "wide serve", or even "four bounces", because nothing H-indicated these conditions over the period during which I acquired the habit of leaning to my right.

The remaining consequences speak to the practical utility of intentional explanation on Dretske's conception rather than the theoretical issue of whether or not it has been shown that such explanation could be genuinely explanatory. Consequences 2, 3, and 4 suggest that intentional explanation is of rather limited usefulness, given that what is wanted from it is a genuine explanation of behaviour. Its genuineness hangs on facts not available in current observable

circumstances of the subject whose behaviour is to be explained. Certainly in the rough circumstances in which we often apply them, the aptness or accuracy of such explanations is sure to be hopelessly questionable. So if Dretske's account has managed to make the meaning properties of intentional states matter in behavioural explanation at all, a large part of what he set out to do, he has not made them matter much.

The conclusions of the previous two sections have been critical of Dretske's strong conception indication. The earlier section concludes that having the right indicator history is not a causal factor in behaviour structuring. This is going to be a problem for satisfaction of relevant counter-factuals: it will never be true that C has come to cause M and has a history of indicating F, and C would not have come to cause M if past C-tokens had not *indicated* F in cases in which C has not H-indicated F over the requisite learning period. Moreover, even where C does have this history, it is certainly debatable whether or not it is the having of this indicator history (as opposed to there being some correlation) which does the work of structuring the behaviour. This last section goes on to suggest that even if indicator history is relevant, the fact that it will only occasionally be satisfied threatens to restrict seriously the class of behaviours which have genuine intentional explanations.

f. Should the Indication Conditions be Revised?

One response to the challenges of the preceding sections is to concede that the notion of indication is simply too strong and to weaken it[18]. Dretske himself moves away from the idea that the relation between C and F need be nomic in the strict sense. Another suggestion would be to weaken the stipulation on H-indication that, over a developmental period, every C-token must be an indicator of F. There is certainly precedent for such a move, Lloyd's information-based theory

of representation being one example[19]. This would in turn open the way for a history of indication to be satisfied in cases of imperfect correlation. Kelly's internal detector of chicken-squawks could then still be promoted to an indicator-function of a weasel's being present even if it is occasionally set off by chickens squawking in the presence of Bloggs the mink farmer. One need only lower the requirement for a state C to have the appropriate indication history by saying that C must be reliably correlated with F. So C gets promoted to an F-representation provided that past C-tokens have tended to indicate F (or in the past C has been reliably correlated with F), C has come to cause M, and C would not have come to cause M if past C's had not tended to indicate (had not been reliably correlated with) F. A weakening of the conditions of indication will turn out not to be helpful, however.

Dretske's theory really incorporates two distinct components. One is the statistical notion of information-carrying (or indication). Any variant on indication is going to be tied to some notion of reliability. The only way, statistically, of distinguishing F from other external conditions with which C may be non-accidentally associated is in terms of Cs being reliably correlated with Fs, either *simpliciter* or relative to other associated external conditions. The supposed uniqueness of Dretske's view, though, lies in the attempt to fuse an informational approach with teleology. C, in coming to cause M when F is present (or in F-like conditions), thereby acquires a *function*. Part of the story of how this function comes about is that a movement M in the presence of F has some F-relevant result R which is of some benefit[20] or of some (selectional) advantage to an organism. However, in order for the theory to be information-based, indication needs to occupy the central role in fixing content and there is a serious question as to whether or not it can.

A way of extending the difficulties already posed for the Dretskean account would be to

argue that the problem is not with Dretske's strong version of a "reliabilist" (indication) criterion, but that any criterion of reliability (C, by some measure, being reliably correlated with F) is misplaced as a general criterion for determination of representational content. This is a way of seeing Godfrey-Smith's arguments against Dretske's approach. Because Godfrey-Smith considers dependencies of C-tokens on an external condition F which are substantially weaker than Dretske's official notion of H-indication, Godfrey-Smith's arguments have application beyond the Dretskean development of an information-theoretic account. He allows the possibility that C-tokens are merely reliably correlated with F[21], not the much stronger requirement that there be no "false positives". One may be a reliabilist, holding on to a variant of a Dretskean theory while maintaining only that the conditional probability of an F given a C must be relatively high. "Relatively high" could even be a value of less than .5. In my terms, C H*-indicates F over some period iff at some relatively high number of C-tokens indicate F over that period.

Godfrey-Smith's arguments are meant to have application still even for a theory which links a notion of indication to relative reliability. An example of such a variant would allow the following: Suppose that there are three mutually exclusive conditions F, G, and H, any of which may be present when C occurs. Suppose further that the conditional probability of F given C is .3, of G given C is .2, and of H given C is .05. On relative reliabilism, C H#-indicates that condition with which it is more reliably correlated that any other (over some period). So C would, if anything, H#-indicate F in the described case (supposing that C does co-occurs with F with greater frequency than any other condition over some period of time)[22].

The reasoning the Godfrey-Smith employs exploits the fact that there are two components to Dretske's view for content determination. Put in counterfactual terms the view is: C represents

something's being F only if C has come to cause M in the presence of F, and C would not have done so if past C-tokens had not indicated F (or C had not H-indicated F over t). There is in this criterion what Godfrey-Smith calls a built-in assumption of "weak adaptationism". This is the idea that the behaviour structurings produced by the Design Solution are of some benefit or advantage to the organism[23]. On the developmental solution, C will not acquire the function to indicate F unless there is some result with respect to F which makes C-caused behaviour that tends to have that result reinforcing. On the selectional solution, C will not acquire the function to indicate F unless there is some result with respect to F which makes C-caused behaviour that tends to have this result advantageous. F, or rather some F-relevant outcome R, should in this sense be of benefit to the organism.

Also built into the counter-factual is the indication condition, which requires that C's coming to cause M be preceded by a run of F-indicators. As already noted, on Dretske's view it is not enough simply that past C-tokens have occurred in the presence of F. Past C-tokens must have indicated F. More specifically, over the "learning period" during which a new behaviour is acquired (call this t), C must H-indicate F. To summarize, the two implicit conditions for "C represents F for S" are:

(a) C has H-indicated F over the period t in which C came to cause M,

(b) an outcome R with respect to F is beneficial to the organism, such that had past C-tokens not been dependent on F, C would not have come to cause (F-relevant movement) M.

The counter-factual at (b) is different from the counter-factual stated previously in that there is no specific mention of the relation between C-tokens and F being one of indication. I have summarized the two criteria in this way so as separate the "reliabilism" (indication, criterion (a))

and the "adaptationism" (criterion (b)) in a way that makes it easier to see what Godfrey-Smith argues. The idea is that there are plausible cases in which the two criteria come apart. C might come to cause M and would not have done so were it not for some association with a beneficial (or advantageous) condition F, where the association in question fails to meet the standards for C having H-indicated F over t. There are factors in virtue of which C might be recruited as a cause of M which are independent of the relation between C and F being one of C's having been reliably (and non-accidentally) correlated with F.

Godfrey-Smith pursues two critical points. The first is that, on an intuitive notion of representation, it will turn out that there will be cases in which C has come to cause M in the presence of F-like conditions such that it is plausible to say that C has come to represent F but C will lack an appropriate reliability relation with F. The second is that C's H-indicating F over t fails to occupy a central role it needs to occupy in structuring explanations in order for C come to have the function of indicating F. The result is going to be the same for weaker reliabilist alternatives to "H-indication" because the problem lies not in the relation but in the description of the function to which C is promoted. Here is Godfrey-Smith's version of this point:

> The important point is that Dretske needs indication to be significant enough in recruitment explanations for it to make sense to say that inner states like C have the *function* to indicate, that indication is what a certain inner state is *for*. But though Dretske is clear about the fact that functions somehow derive from explanations, his caution about functions prevents him from telling us how *much* explanatory significance is needed. On the face of it, the claims of explanatory prowess Dretske makes are very strong. To say simply that C was recruited because it indicates F, as he does, is not to claim that this fact of indication alone explains the recruitment; but this apparently commits Dretske to the claim that indication is doing the bulk of the explanatory work. (Godfrey-Smith, p. 301)

In order for the bulk of the explanatory work to fall to indication (or some reliabilist variant)

it needs to be the case that what makes recruitment of C as a cause of M *adaptive* is C's reliable association with F. Consideration of a range of cases suggests that this is implausible. While C's being reliably correlated with F might be an important factor in a subclass of behaviour-structuring situations, it will not in general be the principle in virtue of which internal states come to cause movements. The conclusion that Godfrey-Smith recommends is that we should drop criterion (a) (or any other "reliabilist" replacement) as part of the general criterion for content determination. The weight of determination should rest on some development of (b). Accepting Godfrey-Smith's recommendation is to repudiate information-theoretic accounts of representation.

Godfrey-Smith considers a series of examples[24] involving an unidentified predator and various strategies by which it might be conditioned to strike at prey. There are three distinguishable types of cases that Godfrey-Smith employs:

(1) An internal state C is recruited as a cause of a movement, by an adaptationist criterion because of a relation with F, under conditions in which neither C nor any "competing" potential causes of M is reliably correlated with F,

(2) An internal state C, because of factors independent of reliability, is recruited as a cause of M even though some other internal state C* is more reliably correlated with F,

(3) In cases of "borrowed" indication[25], where C is reliably correlated with an external condition G and G has some relation with F, reliabilist and adaptationist considerations may give different answers to the question of what C comes to represent. Moreover, the problematic answers appear to be the reliabilist ones.

Godfrey-Smith's description of the first kind of case is as follows:

There is a variety of cases in which predatory animals will recruit as causes of

> aggressive M states which are poorly correlated with prey. The simplest kind of situation is that where the environment and the prey's evolution have combined to produce a highly cryptic (camouflaged) prey population, and the predator's sensory and cognitive abilities are such that there is no inner indicator of prey available to recruit. There is no inner state such that when it is tokened, the chance of there being prey at hand is better than even, or better than the chance of any other state obtaining. In this situation the predator might recruit as a cause of M the state best correlated with prey, but because it has not recruited an indicator, Dretske would disqualify the state as a representation of prey. (Godfrey-Smith, p. 298)

This class of cases is problematic. It would appear that, were one actually studying the behaviour of the organism, the exhibited movement M was directed at capturing prey. So it would be reasonable to suppose (assuming that behaviour is grounded in representational states) that some state of the organism which represents external conditions as prey is the cause of the exhibited movements. Yet, where the requisite reliability relation is absent, the organism is a kind of "Swamp-predator" with respect to its prey-catching movements. The aggressive M, caused by C can be directed at obtaining prey, but an instance of this exhibition cannot, by Dretske's theory, be explained by appeal to intentional states. It cannot be so explained even if the predatory behaviour is conditioned behaviour.

This first kind of case is still not the most difficult. One could still offer a modified "reliabilist" criterion: C must be more reliably correlated with F than is any other internal state of the organism. Such a move is rather unmotivated, however, and it cannot help with the second kind of case. This second kind of example makes the point that, once adaptationist considerations creep in, reliability in any form will often be beside the point.

We consider a series of examples involving an unidentified predator and various strategies by which it might be conditioned to strike at prey. Reliability between a prey detector and the actual presence of prey, Godfrey-Smith argues, is not the most important consideration in

behaviour-structuring. Given two potential detectors C_1 and C_2, the one which is most unreliably correlated with prey might well be the one which gets recruited as a cause of prey-catching movements. C_1 may generate a lot of "false positives", which is to say it may often occur in the absence of prey. C_2 may generate far fewer "false positives". Suppose, though, that prey is scarce and C_2 generates far more "false negatives". That is C_2 often fails to occur when prey is present. If prey is scarce and C_1 rarely fails to occur when prey is present, then C_1 may well be the internal state recruited as a cause of M. In this case, the conditional probability of F given an occurrence of C_1 is lower than the conditional probability of F given an occurrence of C_2. If anything is an indicator of prey it is C_2. However, in actual cases C_1 is more likely to be recruited as a cause of M.

Other considerations lead to the same result. The many false positives generated by C_1 are harmless whereas the few false positives generated by C_2 are not. Suppose that in the predator's environment, there is its prey, (present in) F_1 and two close relatives, (present in F_2 and F_3). The many false positives generated by C_1 are F_2 conditions. The organism present in condition F_2 is inedible but harmless. The few false positives generated by C_2 are F_3 conditions. The organism present in condition F_3 is not only inedible but administers a very painful sting when attacked. What we would expect in actual cases of this type is that the predator will become conditioned to exhibit predatory movements when C_1 occurs? One may speculate that C_2 will eventually cause avoidance, probably after very few encounters with condition F_3.

Another factor might be reaction speed. C_1 may be poorly correlated with F_1, but when it occurs it can trigger the attacking movements more quickly than would C_2. So Godfrey-Smith asks that we suppose C_1's causing M has a high success rate when F_1 is present but C_2's causing M has

a much lower success rate. In the second case, M occurs more slowly and the prospective prey has time to escape. In such circumstances, even though C_2 is more highly correlated with F_1 than is C_1, C_1's causing M is more likely to be reinforced since it is relatively more likely to bring about the reinforcing condition.

Examples such as these are easier to press *a priori* in cases of natural selection. Assume that natural selection generally operates by "selection" of members of a population with some behavioural advantage. Species members with a C_1-M package may well have an advantage over those with a C_2-M package even when C_2 is more reliably correlated with prey than is C_1. This could be because C_1 occurs far more often, thus generating a high rate of "false positives", whereas C_2 generates very few false positives but many false negatives (failures to occur when prey is present). Thus the organisms with the C_1-M package obtain prey more often despite the relative unreliability of C_1's correlation with the presence of prey. The cost of individual missed strikes may be much lower for C_1 than C_2 (suppose the few false positives generated by C_2 result in the predator eating a poisonous organism). The reaction-time for C_1-M may be lower or C_1 may allow the detection of prey while the predator is in hiding[26].

The examples Godfrey-Smith uses generally exploit the fact that Dretske's theory is asymmetrical: it is committed to head-world reliability. This is to say that Dretske places great weight on false positives and no weight on false negatives. So one might reasonably suggest that an amendment would be to insist on "two way" reliability. To some degree of reliability, both Cs must generally occur when F is present and Cs must generally occur only when F is present.

Godfrey-Smith has a short but reasonable way with this sort of suggestion. Essentially he argues that amending conditions of reliability will not solve the problem because it is

considerations extraneous to reliability of correlation that gives rise to the problem. An internal state C with the highest degree of head-world reliability will sometimes be recruited as a cause of M. However, whether it is so recruited will depend on its general success when there are other competing Design Solutions. General success is determined by considerations of cost versus benefit rather than considerations of reliability. Highly reliable F-indicators (a C with high mind-world reliability) make sense where the cost of false negatives compared with false positives is relatively high. Imagine an environment in which C_1 and C_2 may occur either in the presence of prey F_1 or some poisonous organism F_2. C_1 generates many more false positives than does C_2. A predator in which C_1 causes M is much more likely to obtain F_2 than is a predator in which C_2 causes M. In a selectional solution, over successive generations, "C_2 predators" are more likely to proliferate. In a developmental solution (supposing that our predator survives poisoning), an individual with states C_1 and C_2 occurring with some frequency in the presence of prey is more likely to have C_2, the state that minimizes poisonings, promoted to a cause of M. In short, where the cost of false positives is relatively high, what is likely to be "recruited" is a state with relatively high head-world reliability. However, where the reverse is true, where it is the cost of missed opportunities (false negatives) that is high, a state with relatively low head-world reliability may well be the one which is recruited as a cause of M. The consideration at work here are relative costs and benefits to an organism. This consideration is not addressed at all by adding to the existing reliability condition. Those cases, as Godfrey-Smith puts it, in which joint head-world and world-head reliability are important are those in which the cost of false positives and false negatives are relatively equal.

Suppose that prey is quite scarce (so that it is rather important not to miss opportunities) but

the environment contains stinging non-prey (so that it is rather important not to strike at non-opportunities). One might then expect that a state that is in a joint sense more reliable will be recruited as a cause of M. Suppose in this case C_1 generates false positives 60% of the time and false negatives 10% of the time. Suppose that C_2 generates false positives 20% of the time and false negatives 30% of the time. Assuming that C_1-M is approximately as successful as C_2-M in obtaining prey (when prey is present), and assuming that "rather important" carries the same weight in relation to both false positive and false negatives, C_2 is likely up for promotion rather than C_1. C_2 causes M when and only when prey is present to a 50% degree of reliability, whereas C_1 causes M when and only when prey is present to a 30% degree of reliability.

The crucial point here is that reliability criteria are not generalizable. The importance of some criterion of reliability between the occurrence of an internal state and some state of an organisms environment depends on particular features of the case. Whether a certain type of state gets recruited rather than some other type depends especially on the costs and benefits involved. It also depends on what organism states are available for recruitment: none may be particularly reliably correlated with a relevant state of the environment but the development of a certain behaviour may nevertheless be advantageous. In the case of selection, it may be advantageous enough to promote the survival of organisms in which the unreliable C causes a movement. In the case of development, it may be sufficiently strongly reinforcing that an organism is conditioned to do M when C occurs even though C is very unreliably correlated with F.

These results are generated by playing on the tension within Dretske's criterion for content determination. As previously noted, he depends on some (more or less undefined) adaptationist component. It is the fact that obtaining F is of benefit which grounds counterfactuals of the form

"C would not have come to cause M had F not been present when C occurred". Dretske's actual version of this counterfactual additionally builds the reliabilist notion of indication into the counterfactual, specifying that this is the way that C must occur in the presence of F. That is, C must H-indicate F. The tension is then generated by pointing out through examples that a "C-M" behaviour package which tends to have obtaining F as its result can be beneficial even when there is C has not H-indicated F. Godfrey-Smith's proposed solution is to keep the adaptationist component and let go of the reliabilist component, since reflection on possible cases suggests that it is the adaptationist component which is of over-riding importance in behaviour-structuring. As this component is the over-riding criterion in behaviour-structuring, one might then reasonably infer, a theory which conceives of representations as a class of functions should adopt that criterion as the main content-determiner.

Here is an adaptation of a final example of Godfrey-Smith's, this one involving "borrowed" indication, which suggests resolution of the tension in favour of adaptation. Chickens S_1 and S_2 exist in environments in which some external stimulus, a certain sort of shadow G is H-indicated by an internal chicken-state C. In S_1's environment, to some criterion of reliability G's occur when condition F is present. F in this case is the presence of a hawk. In S_2's environment the connection between G and F is quite unreliable. So in S_1's environment, let us say that since both C "indicates" G and G "indicates" F, that C "indicates" F[27]. In S_2's environment C "indicates" G, but G does not "indicate" F, and hence C does not "indicate" F. Let us suppose further, so that the case at hand is clearly about learning, that S_1 and S_2 are each veterans of several hawk-attacks and have thus been conditioned so that, when they perceive an instance of the shadow denoted by G (ie. G causes a C-token), the triggered C-token causes hawk evasion movements. Now Dretske does allow the

sort of "stand-in" promotion that arises in the case of S_2. However, if we maintain the reliabilist component of content-determination, the content of S_1's C-states is different from the content of S_2's C-states. If an S_2 C-state represent anything it represents "that G". An S_1 C-state is going to represent "that F". Surely the (negative) reinforcing condition which explains why S_1 and S_2 flee when they see an overhead shadow is the same condition: some of those shadows come with hawks attached. So the grounding of the counter-factual in each case should be the same. S_1 would not flee overhead shadows if they had not (sometimes) preceded hawk-attacks. This is also the case for S_2. However, if one insists that C cannot have representational content "that F" unless C has "indicated" F, then S_1's C-states represent that F while S_2's C-states, lacking the appropriate "indication" history, do not. The difficulty vanishes if one drops the insistence on the reliabilist criterion. Both S_1 and S_2 have C-states which have come to represent hawks because (since it is avoiding hawks that is "advantageous") C has come to cause evasive movements and would not have done so were it not for the association (however unreliable) between C and the presence of a hawk. That is, it is the (occasional) presence of a hawk that has caused C to cause evasive movements.

One of the lessons of this discussion is, again, that the strong formulation of indication upon which Dretske relies figures into operant conditioning (at best) rarely[28]. In the face of such points, a Dretskean may stand by the strong notion of indication as the grounding for representation or not. The easy path looks to be the latter: simply concede the points against the strong notion and adopt a much more liberal basis for representation. Adopting a weaker reliabilist component does not alleviate the problems. In fact, this path amounts to the abandonment of the information-theoretic theories of representation and intentional content.

The central point is that the two components to Dretske's view of the basis of representation can conflict. The account of representation states that C represents F iff C has come to cause M and 1) C would not have come to cause M if past C's had not occurred when F was present, and 2) C has H-indicated F. 1) makes no claim about reliability, only that a movement in the presence of F is reinforcing and that on some past occasion doing M when C occurred was reinforced. 2) (on Dretske's or even some weaker "reliabilist" reinterpretation of indication) is the condition which stipulates that some criterion of reliability must be satisfied in order to get representation from an instance of behaviour-structuring. Godfrey-Smith suggests that as part of the account of how representational content is determined 2) should be dropped.

Thus even to concede that the reliabilist criterion should be adjusted to accord better with the adaptationist criterion is to concede that the content-determining criterion is really the adaptationist one. C represents whatever it is in the organism's history that accounts for C's coming to cause M. This is whatever satisfies the counter-factual "if past C's had not occurred in the presence of __, then C would not have come to cause M". Once the door is open in this way to question the general effectiveness of anything like indication in behaviour-structuring, there are convincing reasons for denying that it, or any reliabilist alternative, is an important content determinant. However, it is precisely the reliabilist component of Dretske's view that makes it information-theoretic. Godfrey-Smith's proposed change is therefore a rejection of the information-theoretic approach.

Moreover, rejection of the indicator relation, and of C's H-indication of F (over t) as representation determinant as the primary workers in the developmental solution to Dretske's Design Problem raises an interesting question. What does the work when a system undergoes a

developmental design solution? The answer in connection with the selectional solution is that the work of structuring is managed by the differential effects on the reproductive rates of members of a population with different genetic packages. The environment "selects" by process of elimination (very roughly) those packages which afford the organisms more opportunities to reproduce. By being reproduced more often, the selectional behaviour (a trait) proportionally increases in a population over generations. Operant conditioning is the basis of the developmental solution and it works on reinforcement. Very little is said by Dretske on what exactly this entails. Certainly the success of operant conditioning, the success of a behaviour being reinforced depends very little, if at all, on indication or H-indication.

NOTES

1. This criticism has been repeated in a less detailed form in Baker's more recent book *Explaining Attitudes* (Baker 1994, pp. 56-62).

2. I should qualify this point a little: the worry, particularly in Kim's case, may not be exactly that content has no role to play at all, but that an occurrence of a behaviour at a given time, a C-token's causing M at t, is not explained by the C-token's having a semantic property "that F" at t. The currency of the role of a semantic property in an instance of a behaviour appears to be Horgan's problem as well. This is not far enough for Baker's taste: she contends that if Dretske manages to avoid her circle, he does so at the expense of giving semantic content any causal/explanatory role *at all*.

3. This definition should actually be amended at (i) to say "state C is of a type past tokens of which have indicated F". The amendment recognizes Dretske's contention that a state does not necessary indicate what is its function to indicate. It does not affect Baker's point, though.

4. Perhaps it would be more careful to say "have come to be", rather than "are".

5. I realize that this view leaves wide open the question of what work, if any, representation itself does. This is an issue I intend to address shortly.

6. There is an additional difficulty in this passage that is certainly worthy of note. Generally, an instance or token of a behaviour would be a token of the process C's causing M. Dretske's description of the "causal job being explained", that which C_n is doing, is described as "causing M_n -- the current behavior". How now ought one to read the symbol "M_n" elsewhere in this passage?

Does "M_n" otherwise represent a token of the movement component of a behaviour process or the behaviour itself? The distinction is important for at least two reasons. First, on the "process" interpretation C_n is itself part of the behaviour and what is wanted in Dretske's view is an account of how reasons (that is, intentional states) can be explanatorily relevant. So C_n, *qua* reason would have to explain a token of a process when it is in fact part of that process. As we shall see a little later, the temporal problems involved in this have been a source of concern for a number of prominent critics. Second, it is supposed to be the idea of behaviour as a process which (for Dretske) makes it susceptible to the sort of explanation in which semantic properties are relevant. Is C_n, then supposed to be the cause of a token *behaviour process* M_n, or is C_n part of a token process which is explained by what past tokens of C_n's type have indicated? The usual line for Dretske is the latter.

7. Cummins has both of the objections I am discussing here. He lists four "consequences" of Dretske's view:

> 1 Only learned behaviours have explanations in terms of semantic content;
> 2 not even learned behaviors can be said to be explained by representations;
> 3 it isn't current meaning that explains current behavior;
> 4 whether current behavior is properly explained by the semantic content of an internal state depends essentially on the system's history. (Cummins, p.105)

Number 3 is the same point that is made by Horgan (and Noordhof). Number 2 is the point that Baker presses.

8. This is an issue that was covered in Chapter 2, section b.

9. Obviously, showing that principle [I*] is not a principle of operant conditioning was not the original point of these studies.

10. There is an additional qualification made, but it is a relatively minor one. Empirical psychology (in cases that would be relevant here) is typically concerned with external conditioned stimuli and how the affect behaviour. So the equivalent of Dretske's "C" would be a discriminable feature of the environment rather than an internal state of the subject. Slater takes note of this by adding a reasonable assumption:

> If, on the other hand, experimental studies suggest that indication* of an external F by an external C is not necessary to put M under the control of C, we shall have no reason to say that indication* of an external F by an internal C is necessary. (Slater 1994, p.174)

11. I may be taken here as exploiting Dretske's inclusion of "locally valid" dependencies. Let us suppose that chickens have a characteristic squawk when they sense a dangerous predator, a type of call that is distinguishable from other chicken-vocalizings, and that there are no other dangerous predators around my farm besides weasels. So there is at least a localy valid dependency between chickens' weasel-sensing and their characteristic predator-squawk. Moreover, nothing else but this type of squawk generates a certain auditory state in Kelly.

12. This is not clear from the cited quotation, but is clear from a statement later in the same paragraph: "This is the way modal relations -- including indication -- come to figure in the explanation of certain developmental processes (learning)." (Dretske 1994b, p.204)

13. This example occurs in Davidson's "Knowing One's Own Mind" (Davidson 1987).

14. This Forum is found in Mind and Language 11 (1996): 69-132. Contributions to this forum are included in my bibliography. The issue of Swamp-persons and its implications for theories of mind has generated a substantial body of literature on its own. I do not intend in any way to give a significant treatment of the "Swamp-person" issue in general.

15. One can of course agree with Davidson on certain of his points without agreeing with the general conclusions (eg. "no thoughts"). Certain mental states clearly do require some relation to things outside the cognitive being's head. Success-oriented states are definitely of this kind. Remembering my grade one teacher requires some relation to another particular individual. A spontaneously created duplicate would lack that relation to the individual, and so might, in Derek Parfit's terms "q-remember" (Parfit 1971,pp. 14-18) his grade one teacher. "Q-remembering" is not remembering. It is, however, a mental state. It is still less clear that a "Swamp-person" could not "q-remember", though one may argue that such a being would fail to have appropriate relations to the linguistic community for actual understanding of notions such as "grade one" or "teacher". Perhaps one could still suppose that a Swamp-person would "q-understand" the concepts required for "q-remembering" his/her grade one teacher. Would "q-understanding" or "q-proposition entertainment" be mental states a Swamp-person could have?

16. See Dretske's contibution to the Mind and Language forum on Swamp-persons (Dretske 1996a) as well as his *Naturalizing the Mind* (Dretske 1995, pp.141-151). The article is a very slightly modified version of what appears in the book.

17. This example is not problem-free. Certainly it raises the question of whether the identified function properties could, qua functional properties have any causal role in the plant's behaviour. That an internal trigger of colour change has the function of attracting hawkmoths is certainly not as such a reason for a current example of Scarlet Gilia to behave this way. This is a point which Dretske himself notes about selectional solutions of course. What is explained by the function, if anything, is merely why a genetic package including the mid-season colour change flourished within the Scarlet Gilia population. On the one hand, one may still deny that there is any causal difference between Twin Plant and Scarlet Gilia: there is after all a common, general explanation of why the colour-change was propogated. In the actual case, there were conditions present in Scarlet Gilia's environment which promoted the reproduction of colour-changers so that over successive generations plants with this trait became more common. This is exactly the same general explanation for Twin Plant. Its environment promoted the reproduction of colour-changers. What is different of course is the more specific feature of the environment in each case in virtue of which this was so. One could say that colour-changing was propogated because this trait had the function of allowing colour-changer pollen to be spread more commonly than non-colour-changer strains of the plant. Twin plant's colour changing has this function as well and

at bottom it is specifically this fact which has led to the propogation of colour-changers. Botanists of course have different interests than do philosophers. So they may well be interested in what specific factors in Scarlet Gilia's environment accounted for the success of the colour change. Botanists are interested in knowing what mechanisms might affect the relative success of plant pollination. That does not guarantee that there must be an interesting metaphysical difference between the function of attracting hawkmoths and the function of repelling beetles. So one might well argue that it is not clear that "being a hawkmoth attractor" and "being a beetle repellor" are important causal properties, properties that constitute a difference between Scarlet Gilia and Twin Plant.

On the other hand, notice, if any of these environmental properties are going to figure into the structuring explanation of either plants' behaviour, how those properties figure in. In Scarlet Gilia's case, let us suppose that the explanation of the propogation of colour change is to be explained specifically in terms of hawkmoth attraction. If the *function* of attracting hawkmoths enters anywhere in this explanation, surely it enters into the explanation of the spread of this trait. The trait becomes more common in successive generations because colour change to white flowers attracts hawkmoths which increases the length of the breeding season for plants with this trait. In increasing the population of colour-changers, colour changing is already doing the job (whether we know it or not) of attracting moths. If one includes "attracts hawkmoths" as part of the explanation of the propogation of this trait, that is, it is that the trait has the *function* of attracting hawkmoths which explains the propogation. So the selectional structuring of my current example of a Scarlet Gilia, the explanation of why it undergoes colour change, is given in partial terms of its being of a type in which colour change has functioned to attract hawkmoths.

This is a non-parallel with the developmental solution, an important one. Selectional structuring is explained by the function a trait has (which explains why it is advantageous). Developmental structuring cannot, on pain of circularity, appeal to *functions* (to indicate).

18. Alternatively, one could adopt a weaker notion intended to capture the same general spirit as indication, for example "reliable correlation".

19. This theory is developed in Lloyd's Simple Minds. His commitment to a much lower standard of reliability for his equivalent of Dretske's notion of indication is especially apparent in the second of his three conditions for representation:

> A natural event r is a representation if it meets these conditions:
>
> 1. *The multiple channel condition* There is a set of at least two events, $\{v_1,v_2,...,v_n\}$, such that r is dependent on the concurrent conjunction of at least two events in the set...
>
> 2. *The convergence condition* Events v_1 through v_n are further subject to the constraint that there is a set of single events $\{u_1,...,u_n\}$ (the single mutually effective stimuli), such that all of v_1 through vn depend on each element of $\{u_1,...,u_n\}$. The object of representation is the element of $\{u_1,...,u_n\}$ with the highest conditional probability, given r...
>
> 3. *The uptake condition* Event r has the capacity to cause either another representation or a salient behavioral event. (Lloyd 1989, p.64)

The key line is "[t]he object of a representation is the element of $\{u_1,...,u_n\}$ with the highest

conditional probability, given r". What r is a represention of, then, is whatever (of those elements appropriately dependent on r) is more (conditionally) probable given r than any competing "representation candidates". This, as Godfrey-Smith points out, is still a kind of reliabilist criteria but a very weak one. The conditional probability of F given C could be very low but still be higher than the conditional probability of any other condition G (upon which C is dependent) given C.

20. Something's being of benefit is to be understood in this context as receptivity-satisfying, not that which is in some objective sense beneficial.

21. In so doing, Godfrey-Smith is also allowing for a much weaker dependency between a C-token and a token of condition F: the conditional probability of an F-token given a C-token can be much less than 1. He does not differentiate these notion of a given C-token indicating F, and (type) C's H-indicating F (for some S, for some duration) as I have. I made this point in Chapter two. Thus, on Godfrey-Smith's way of looking at Dretske's theory, weakening the dependency of a C-token on condition F (characterized in terms of conditional probability) is to weaken the historical requirement for C to be promoted to an F-representation.

22. I am of course aware that the sort of reliabilist view I have just sketched is a hopelessly naive view as it stands. It is obviously subject to an objection in the form of a disjunction problem: the conditional probability of F or G or H given C is higher that the conditional probability of F (alone) given C, so why would not the representational content of C be the disjunctive content "that F or G or H"? It is not my purpose here to outline viable alternatives to Dretske's concept of indication. I am merely giving a sense of the breadth of application that Godfrey-Smith intends for his criticism. The breadth of application is to any view which builds some notion of reliable correlation of C-tokens with F into the criterion for content-determination.

23. In fact it is implicit in the very idea of design problems and their solutions. The premiss of the design problem as borrowed from engineer is "how does one assemble a system which services the needs of some representation consumer". How does one get chickens to hide from hawks? How does one get a rotating cable connected to an axle to tell us about vehicle speed? These are problems whose solutions confer some advantage.

24. One should note that Godfrey-Smith does not differentiate between developmental and selectional solutions to the Design Problem. Dretske himself makes a great deal of the distinction between selection and development. Godfrey-Smith downplays the difference. He sees it simply as development doing at the individual level what natural selection does at the species or population level. Some might not be so quick to dismiss this difference as a simple difference. I will interpret his examples so that they apply specifically to the developmental solution. It is this solution to the Design Problem which is of central interest.

25. This is a kind of special case allowed by Dretske where a given external cue G which is H-indicated but which is of no special benefit to an organism might be exploited because it of its association with some other ecologically important condition F. The organism may have no detectors for F, and thus borrow external cue G (for which they do have detectors) so that C will

come to cause M when F is (likely to be) present.

26. Imagine a species of ground-dwelling spider that is sensitive to prey via felt vibrations in the ground, something it could use from within its burrow and by sight (something it could only when exposed to the possibility of being seen by its prey. It is conceivable that a state caused by felt vibrations could be a more successful trigger of predatory movements that would be a state caused by visual contact with prospective prey. This could be true even if the vision-caused state is more reliably correlated with prey than the vibration-caused state, for the vibration-caused state may bring the prey much closer to the spider before it able to be aware of the spider's presence.

27. I am using (quotes) "indicates" here to avoid confusion with Dretske's use of the term. It is meant here simply to stand in for any agreed upon reliable correlation. This includes but is not exclusive to Dretskean indication. Of course one should also note that once the standard of reliability is weakened from Dretske's absolute criterion, the relation denoted by "indicates" is not necessarily transitive. I am supposing for simplicity of the example that the reliability of correlations between C and G and between G and F in S_1's environment is sufficiently strong that the C-F correlation meets the requirement as well.

28. Godfrey-Smith also maintains this with respect to natural selection. He sees little difference, even less than does Dretske, between behaviour produced by natural selection and that produced by learning.

CHAPTER SIX

Prospects for Dretske's Information-Theoretic Approach to Mental Representation

a. Introduction

In this last chapter, I want to further the discussion of Dretske's theory by arguing that commitment to the developmental solution to the Design Problem, certainly as Dretske himself conceives this, will not provide an intentionality-free base upon which to found an account of intentional states. The ways in which intentionality enter into the solution might have been incidental *if* the notion of H-indication had done the explanatory work that Dretske had hoped it would do. As I have already argued, it is essential to the claims of Dretske's Representational Naturalism that the indication relation bear the weight of determining representational content of intentional states. To do this, and to succeed in making the semantics of intentional states effective, a state type's having H-indicated an external condition needs to be causally explanatory. I have argued that this notion does not do the work required.

In his chapter "The Explanatory Role of Belief" (Chapter four of *Explaining Behavior*), Dretske anticipates the criticism that the model of (intentionally driven) behaviour he is elaborating is too simple to account for the sorts of sophisticated intelligent behaviour we would hope to understand. His answer (which refers to his chapters four and five) makes it clear that he sees himself as laying out the bedrock of intentionality:

> This challenge -- a very serious and understandable challenge, even among those otherwise sympathetic to naturalistic accounts of the mind -- will be confronted (with what success I leave for others to judge) in the final chapter. What we are after in the present chapter ["The Explanatory Role of Belief"] is something less ambitious: an account, however oversimplified and crude it might have to be, of the basic cognitive building blocks. What we are after in this chapter and the next ["Motivation and

> Desire"] are the elements out of which intentional systems, systems whose behavior can be explained by *reasons*, are constructed. (Dretske 1988a, p. 85)

Clearly, if one is endeavouring to provide the *basic building blocks* for a naturalistic account of the content of intentional states, it will not do to have those building blocks resting on unaccounted for phenomena which are themselves (arguably) intentional. If the information theory is to be naturalistic, the ground level (at least) of intentional states for which Dretske is trying to account should not appeal to *intentionality*. A more general sort of circularity than that envisaged by Baker (chapter 5 section a), then, could be shown by showing that Dretske's account of his building blocks already depends on intentional properties. In this closing chapter, I will revist the Dretskean view of how genuine semantic properties of intentional states are supposed to arise from the developmental Design Solution. I will suggest a couple of respects in which the solution itself may presuppose that for which Dretske is trying to account.

b. The Design Problem Again

To recapitulate, the move from indication (or information-carrying) to representation depends in Dretske's view on a solution to the Design Problem. The Design Problem is effectively a means by which a non-semantic physical structure is given an interpretation which exploits its causal powers. Both the engineer's solution and the selectional solution to the Design Problem are vitally different in two respects from the developmental solution. One is that according to Dretske the developmental solution (and only this solution[1]) is supposed to yield content that is "to or for" the organism in which the state is developed. The other two solutions do not. The second is that there is in each of the first two solutions a fairly clear mechanism or process which does the selecting of appropriate indicators. The selecting process in the developmental solution is not clear.

The mechanism for selection of indicators is most obvious in connection with the engineer's solution. The engineer is an intentional agent who brings to the Design Problem a body of knowledge/expertise about what sorts of potential devices have the requisite causal powers. The engineer needs to know how a prospective indicator reacts to various environmental conditions, and needs to know how it can be suitably connected to a mechanism which will perform the relevant output. One needs to know a great deal in order to get a cable connected at one end to an axle and at the other to a pointer in a numbered gauge to represent vehicle speed. So the engineer's solution gives an interpretation to representational structures by means which are of no help to a naturalist: the interpretation exploits (natural) causal powers, but it is nevertheless, *qua* representation, a conventional interpretation. What an engineered structure represents is whatever the engineer, relying on his or her knowledge of the structure's causal powers, makes it represent.

There is a "blind" equivalent in the selectional solution. Genetic mutation may result in different behaviours within a species population[2]. Such a behaviour is in a sense a completed package: there is nothing to be structured or developed within an individual organism which has the appropriate gene. The conditions that obtain in the environment then selects the prevalence of that package in successive generations by an elimination process. If the behaviour has some positive survival value, organisms with the gene reproduce with greater frequency than those which do not have it and the population of organisms which happen to have the behaviour trait increases. If the behaviour has negative survival value, organisms with the gene reproduce less frequently and the population of organisms so wired diminishes over time. Thus there is a blind, environmentally determined elimination process which, over time, establishes or diminishes such a genetically transmitted behaviour trait within a kind. The lesson which Dretske draws from the selectional

solution is that functions, in particular information-carrying, behaviour-producing functions are possible in the absence of intentional design (see above pp. 103-104). He similarly supposes that his developmental solution also is a source of natural functions.

However, the developmental solution, unlike the selectional solution, does not apply to kinds of organisms[3]. It is not a "solution" for species or populations. By the same token, the selectional solution is not a Design Solution which is solved on the level of individual organisms, for an individual is either born with the requisite code, in which case is has the behaviour, or not. If not, the individual does not have the behaviour, C's causing M in F-like conditions, and thus C does not represent something's being F.

Thus there is a remaining problem in the presentation of the three solutions of the Design Problem. At the level of individual representational systems, the selectional solution has no recruitment mechanism for indicators because it does not recruit indicators at the level of individuals. What is more, the mechanism at the phylogenic level by which the selectional solution solves the Design Problem without consciousness has no obvious equivalent in the developmental solution. The selectional solution operates on populations by trial and error (mutation) and a process of elimination. Those organisms which lack an advantageous behaviour are not able to reproduce as often so that the number of offspring with their genetic traits declines in a population. It is essential to this process that the mechanism which effectively "contains" the behaviour in encoded form in material passed on by reproduction. The engineer's and the developmental solutions are solutions to the individual Design Problem: how to get a particular behaviour-producing or regulating structure in a particular system. However, the engineer's solution is a conventional solution, but the developmental solution must be natural.

I suspect that Dretske, and others, think of development and selection as closely parallel. Perhaps trials in a conditioning period are supposed to be viewed as analogous to members of species. Behaviour (or movement) variants, sometimes performed under differing enviromental conditions, and/or when different internal states are present meet with differing degrees of success in relation to a reinforcing condition. The relative success of a given movement M_1 in the presence of a certain type of internal state C_1 entrenches the pattern so that C_1's causing M_1 becomes a regular part of the subject's behaviour repetoire. The structure "C_1 causing M_1" is reproduced because of relative success of past instances in yielding some beneficial result.

I am not suggesting that Dretske or anyone else has *explicitly* advocated exactly this view. Put in this explicit form, the analogy involves a clumsy and obvious equivocation over "reproduce". In connection with the selectional solution it refers to biological procreation whereas in this last context it simply refers to repetition. The speculative exercise does help to point up how the developmental and selectional solutions are different in a quite important way and that something is still missing from the developmental story. There is a glaring disanalogy here between development and selection: Unlike genes, behaviour-instances do not replicate. So an analogy between the developmental and selectional solutions breaks down at this point. We need an identification of the *mechanisms* in reinforcement, by which trials in operant conditioning establish a behaviour-type. Moreover, it is *prima facie* reasonable to suggest that reinforcement only works where certain kinds of mental, discriminative abilities are at work. Is it not (for example) Sniffy's desire for food, and his perception of token features of his environment as being of the same type that partially accounts for the entrenchment of bar-pressing movements?

Could one dismiss the prospect of such examples of "creeping intentionality" as unimportant?

Perhaps one could say that whatever cognitive capacities are necessary for reinforcement to do its job involve states of a rather low grade of intentionality and the role they play is not at all central in the content-determination of further intentional states. After all, Dretske has stated that he intends to defend his Representational Thesis "plus or minus a bit" (see Introduction). It may even be possible to argue that such processes as are necessary for learned, discriminative behaviour are really bit players since, according to Dretske, not all cases of operant conditioning are cases of indicator functions being acquired. The main base condition, on his view, is indication. Acquiring a state C whose function is to indicate F does not come about unless past C-tokens have all been F-indicators over the development period, and indication is a thoroughly physical, non-intentional relation. This, one might argue, should be enough to establish Representational Naturalism. Clearly, this move depends on defending indication, or some variant, as the central content-determinant. However, as we have seen, there are compelling reasons for questioning the centrality of indication. The issue of whether or not the developmental solution is bound to presuppose some representationality or intentionality cannot be disregarded.

Recall the predator examples of Chapter 5, section f. The general approach in these examples is to argue against Dretske and other reliabilists by suggesting that content-determination by some form of reliable correlation gives a view that is too simple to be generalizable. The way to fix this is to recognize that reliability is in fact the wrong criterion for content-determination[4]. It is the counterfactual criterion, says Godfrey-Smith, that ought to be upheld (C represents that F iff C comes to cause M and would not have come to cause M if Cs had not occurred in the presence of F). So it is the condition which satisfies the counter-factual that is the content of the state C.

The question that must then be addressed is "what are the grounds for such counter-factuals?"

The weak adaptationist answer to this question, an answer with a great deal of intuitive appeal, is that some outcome relevant to F is advantageous or beneficial to the organism. Dretske indeed does come to lean on the notion of "beneficial outcomes" in later parts of *Explaining Behavior*[5].

Godfrey-Smith suggests that the weighting of consequences are important in determining which internal states get promoted to certain functions. The promotion of C_1 rather than C_2 as a cause of M (and in order to obtain F) is partly but importantly determined by the relative severity of consequences, efficiency in obtaining the beneficial outcome, durability of the physical components, for example. In short, there are cost-benefit considerations.

How the cost-benefit considerations are "calculated" in the selectional solution does not involve calculation at all. Suppose that I have before me a ground-dwelling spider that responds to prey by taction. At some time in the distant past history of the species, there were spiders which responded predatorily by vision as well as those which responded by taction. The rough story of how it comes about that the present generation responds by taction (hence my spider Sally responds this way) is a story about the effects of an environment on populations. The weighting of consequences, the efficiency of a triggering mechanism in obtaining prey, the relative delicacy of the physical components used in the behaviours are conditions imposed by the physical structure of the environment (and the physical structure of the species involved). An environment with prey with keen vision would tilt the cost-benefit calculations away from sight-based predatory behaviour. An environment in which a predator of the spider generated vibrations very similar to those generated by the spider's prey, by making the cost of "false negative" states caused by taction very high, would tilt the cost-benefit calculations away from taction-based predatory behaviour. Of course it is odd to speak of any sort of calculation here, for properly speaking there is none. The

structure of the environment determines which of several competing behaviour-packages is (all things considered) more advantageous and selects a population in which that package is prolific by eliminating species members with the competing structures. The higher reproduction rate for spiders with predatory behaviour based on taction is (an essential part of) what accounts for Sally having this behaviour[6].

What is important here from a Dretskean perspective is that this rough depiction of how natural selection solves the Design Problem shows the possibility of states acquiring functions that is natural, in the sense of non-conventional. No intervening intentional agent is required in order for a certain type of state in Sally, caused by the sensing of vibrations, to have the function of indicating prey. Completion of the task then appears to require only that one say "and operant conditioning works analogously".

The problem is that operant conditioning appears not to operate analogously. Something must take the place of the environment as the weightor of behaviour-consequences. Moreover what takes the place of the environment must be non-conventional. Suppose that for some organism prey-catching is a learned behaviour. Wyle E., a member of this species, has displayed two different "trial movements" M_1 and M_2. M_1 is a movement that, given the requisite learning period would fall under the control of internal state C_1. M_2 is a movement that, given the requisite learning period would fall under the control of internal state C_2. Let us suppose that he has displayed several trials of each and that, as a matter of fact, M_2 has the best relative success rate. For example, M_2 is not as reliable as M_1, but this is compensated for by M_2's having a lower cost in terms of the risks associated with false negatives, false positives, and so on. For simplicity of comparison, one will need to make a pair of simple assumptions at this point that will not hold in

many actual world examples. The first is that one learned behaviour will be selected to the exclusion of the other. The second is that when operant conditioning makes several Design Solutions available to a learning subject, the optimal one tends to be selected. I set out these assumptions simply to try to maintain a relatively clear comparison between learning and natural selection.

Now the weight of consequences of adopting one prey-catching strategy is still going to be partly determined by the physical environment and the physical structure of Wyle E. Given the simple assumption that one prey-catching strategy will be adopted, how does the weight carried by false negatives versus false positives, the risk to personal health of one movement compared to the other, the number of successful catches (given that prey is present) manage to figure in to the adoption of M_1 rather than M_2? Once again, it is crucial in natural selection that when a corresponding genetic behaviour M_1 is selected over M_2 that what is being reproduced is a genetically transmitted trait, part of a replicating structure. An inherent version of M_2 is just, as it were, passed along when M_2-carrying species-members reproduce and if M_2 is advantageous this happens more often than does the reproduction of species-members lacking M_2. So the environment takes care of weighting of behaviour-consequences: the structure of the environment and its effects on relative reproductive frequency of M_2 versus non-M_2-carriers determines independently of intention or cognition the extent to which the "M_2 package" is of benefit.

This observation highlights the gap in the developmental story. The reproduced entities in development are not biological species members. They are instances of behaviour. An instance of M_2 does not replicate further M_2 instances. Behaviour-instances, unlike members of biological species, do not have reproductive systems. Nor are they parts of Wyle E.'s or any other individual

organism's replicating material. Assuming that cost-benefit weightings play some role in the way that Wyle E. is eventually conditioned (C_2 comes to cause M_2, C_1 is not recruited as a prey-indicator), the effect of the environment on the reproduction-rate of instances cannot be the explanation of why Wyle E.'s "population" of behaviour instances includes many instances of C_2 causing M_2 and no (or few) instances of C_1 causing M_1. If the higher cost of C_1 causing M_1 has been a factor in the selection of C_2 causing M_2, it can only have become a factor as a result of abilities of Wyle E.: Wyle E.'s cognitive assessment of the relative cost of obtaining prey by means of M_2 when F_2 is present as opposed to the cost of obtaining prey by means of M_1 when F_1 is present. Wyle E. has to *decide* between these strategies.

The foregoing line of thought might be taken to lead to a fairly short argument which shows that Dretske's theory is circular after all. Deciding is a mental process. It certainly will involve mental states. Dretske has said that all mental facts are representational facts. So the mental facts in a case of decision are representational facts. So the developmental solution, insofar as it yields adaptive behaviours, presupposes (mental) representational states. The developmental solution is supposed to point the way to an account of mental representation but it appears to presuppose it. Appealing to Dretske's Design Solution as an account of how active mental representation is possible is thus circular.

This conclusion is premature. The described case involved an organism confronted with two prospective prey-catching strategies. This is a relatively complex example of development of behaviour. Not every instance of a developmental solution must be free of intentional presuppositions in order for Dretske's proposal to avoid circularity. No-one, surely, could mean to maintain that operant conditioning *never* presupposes existing beliefs or other intentional states.

All that is required is that some plausible "ground-level" of developmental solutions be free of such presuppositions. The previous example might fail to be a ground-level developmental solution that could come about without prior mental states (or facts). In what follows I will make the case for circularity so as to avoid this problem by linking the circularity to reinforcement itself rather than applications of it.

Let us first consider an example of a less artificial sort than the definitive adoption of one of a pair of occasionally reinforced behaviour alternatives. Suppose that Wyle E. has learned to approach its prey from downwind. Wyle E. hears sounds characteristic of a rabbit (its usual prey). If it is already downwind of the source of the sound, it closes in directly. If it is not, it circles around the sound source until it is downwind. The routine might be controlled by pairs of sensory stimulation: intial awareness of a rabbit is auditory. The closing in on the prey is linked to olfactory stimulation. So if Wyle E. hears and smells a rabbit, it closes in. If it hears but does not smell a rabbit, then it continues to circle around the sound source to a position from which it can smell a rabbit.

The downwind approach movements have come under the control of some internal state. What does the internal state represent (if anything)? Let us suppose that content is determined along the lines suggested by Godfrey-Smith. Reinforcement of behaviour is no more dependent on reliability of correlation than is (natural) selection of behaviour. It is dependent on relative benefits. So the content is determined by whatever satisfies the relevant counter-factual: C has come to cause circling movements in the presence of F and C would not have come to cause these movements had F not been present. It is difficult to know exactly what should be filled in here. It is the presence of a rabbit and the opportunity for obtaining food that is reinforcing. On the other

hand, C is not specifically dependent on the presence of a rabbit. It is specifically dependent on sounds that are sometimes caused by rabbits. Or rather it is dependent on Wyle E.'s hearing of sounds of a sort that are sometimes caused by rabbits. The counter-factual should be reformulated slightly:

> C has come to cause circling movement in conditions in which F has on some past occasions has been present and C would not have come to cause these movements had F not been present on some past occasions.

So the representational content of C is "that a rabbit (or food or prey) is present".

Wyle E. must have undergone some trial process in virtue of which this piece of behaviour is acquired. The question which is of interest to me is the question of what is involved in that trial process. The trial process is the process by which the particular routine "move around without closing until the prey is smelled" becomes established as a repeated part of Wyle E.'s repertoire (call this the circling routine). Noting that the circling routine is reinforcing is far from a complete explanation of why the routine is "reproduced". For one thing, it may be that the routine is no more reinforcing than a straight approach to the prey[7]. Moreover, understanding how a developmental solution works should require an understanding of how reinforcement works. How does something's being of benefit or advantage to an organism manage to bring about entrenchment of behaviour that tends to result in that benefit? What does entrenchment of a behaviour by some reinforcing condition entail (if anything)?

Let us try to fill out the picture somewhat by imagining a couple of trial processes. Since what is wanted is an account of how intentional mental states can arise in a way that does not presuppose capacities that operate over such states, suggested trial processes should be of a sort that

would promise to be intentionality-free. Natural selection appears to select behaviours, roughly speaking, in a way that is intentionality-free by some combination of trial and error and elimination. Natural selection, once again, is not directly helpful to the present case: it operates over populations (rather than within individuals) and it relies on biological reproduction. My two suggestions for trial processes that are most likely to be intentionality-free are nevertheless inspired by natural selection.

The first process (borrowing its name from comments in Godfrey-Smith's article) I call "the Random Experimentation Process": A particular behaviour routine emerges from randomly generated alternative routines. The second I will call "the Mutation Process": A particular behaviour routine emerges over time from minor variations, produced by accident of circumstances, on a prior behaviour or movement. If there are more plausible examples of intentionality-free developmental processes, I cannot think what they would be. I shall illustrate what I mean by these processes with examples. Then I will show how representation creeps into the developmental solution in a way that is independent of the particular trial process by which a Design Problem is developmentally solved. The points I will make will be independent of the specifics of a trial process.

The Random Experimentation Process: Wyle E. settles on the circling routine as a result of a blind trial of various random approaches to its prey. Wyle E. is programmed to "reach" out its environment behaviourally, to engage in exploratory movements. Moving through its environment in various ways, Wyle E. occasionally stumbles upon a rabbit. It is prewired to attempt to eat any animated inhabitant of its environment of manageable size. Attempting this when the animated object in question is a rabbit is receptivity reducing. Wyle E.'s movements, now

rabbit-directed, converge on the "circle and close" routine which is under the joint control of internal states (or an internal state -- we need not assume that issues of state-individuation have been settled) caused by certain types of auditory and olfactory stimulation.

The Mutation Process: Wyle E.'s adoption of the routine is the result of minor variations on particular instances of prey pursuits over time. The variations themselves may be accidental: in chasing fleeing rabbits, Wyle E. is forced to move side to side to avoid obstacles. In closing on still stationary rabbits Wyle E. similarly is forced by the structure of the environment to move around rather than directly at its intended target. These "mutations" of directly closing on prey put Wyle E. in contact with more cases in which it can smell the rabbit. Over time, Wyle E.'s closing behaviour is triggered by an olfaction-caused internal state, its circling behaviour triggered by an audition-caused internal state. So when it hears, but does not smell a rabbit, it begins to circle even in the absence of obstacles.

There are two questions that should be answered about the described trial processes. The first is the question of whether or not such trial processes (if they are intentionality-free) are generalizable enough to form a plausible ground-level of developed behaviour. I do not know how to answer this question. The second is the question of whether, as the obvious candidates for intentionality-free developmental processes, they are indeed intentionality-free. I think the answer to the second question is no and for reasons that have to do with the requirements of reinforcement rather than the particulars of the described processes.

Reason (1)

It is an unfortunate fact that there is a substantial gap between what is in fact of benefit to complex cognitive organisms such as ourselves and what actually motivates us. The latter is what

we *take* to be of benefit, or simply what is pleasure-generating or pain-avoiding. Reinforcement operates on these perceptions of benefits or pleasures rather than what is of objective benefit. Access to what is taken to be beneficial or pleasurable comes to us through motivational states (desires). In Dretskean terms a desire is a receptivity that is reduced by some outcome R. R may or may not be a benefit in the objective sense of actually having some positive value for an organism. R is rather a benefit in the sense of being receptivity-reducing. Survival of course depends on receptivity-reduction being *some* guide to genuine (objective) benefits, but no-one could seriously imagine that it is an unerring guide. The counter-examples are too obvious and numerous. So for a condition to be of "benefit" to an organism in this context is for it to be receptivity-reducing. This is a point which hardly needs argument in relation to Dretske since his examples make it clear that this is how he is thinking of "benefit".

As I argued in Chapter four, Dretske's account of receptivities as non-intentional prior to actual satisfaction is quite unconvincing. Wyle E., one might suppose, has some receptivity which turns out to be for rabbits (or food or prey). Prior to that actual eating of a rabbit, Wyle E. may have no epistemic access to the satisfaction-conditions of that receptivity. Perhaps nothing has that access. A biologist observing Wyle E. as the first member of a new species may observe restless behaviour characteristic of organisms that want something, but she also may not know either, prior to seeing Wyle E.'s response to catching and eating a rabbit that its state was a motivational state *for* food. The fact that the biologist may not have epistemic access to this fact about Wyle E. is no reason to deny that Wyle E.'s receptivity is a food receptivity or that the receptivity in question is (intentionally) *for* something. Why, one wonders, should Wyle E.'s position with respect to its own receptivity-states be any different? Why should Wyle E. need to be in a position to be aware of

what a certain receptivity state is for before it can be for anything? If that is the requirement, there seems to be no answer to the question of why a given receptivity-state of Wyle E.'s turns out to be satisfied by some condition and was not satisfied by others.

Dretske wavers on the status of such basic receptivities as hunger and thirst, suggesting in his response to Bratman (Dretske 1990, p. 835) that these are (at least initially) like drives (what I termed "p-d drives"). I think he is right in so far as he is acknowledging such states as being inherited (rather than learned). However, this is an observation that sits very uncomfortably within the framework of his theory.

First, unlike other examples of inherent states (selectionally acquired indicator functions and n/n drives), p-d drives do have an effect on behaviour development. P-d drives have conditions of satisfaction (in the case of hunger, being satisfied by food, for example) in virtue of which they are effective in shaping behaviour. This makes it plausible to identify these as states which are *for* the outcome(s) which would satisfy them. Such receptivities are not inherent in the sense of being a product of the selectional solution. The selectional solution yields inherent behaviour-processes, not inherent internal states. It yields inherited "states-causing-movements", and when the state in question is a receptivity (an n/n drive) its conditions of satisfaction are ineffective in the development of behaviour. A blowfly will continue to have its probiscis extensions caused by a certain type of internal receptivity regardless of outcomes. This is Dretske's reasoning for excluding n/n drives, drives inherently part of a tropistic behaviour, from the class of genuinely intentional states.

Second, the framework of his theory does not allow Dretske to postulate inherent, contentful states that are unlinked to behaviour. Dretske certainly *denies*, where an internal state of an

organism is *selectionally* linked to behaviour, that any true representational properties (or other intentional properties if there are any) can be explanatory or causal. In the case of p-d drives, conditions such as thirst or hunger, there appears to be a category of states which have intentional properties that are inherent, and Dretske's reason for denying that these properties are efficacious is not available: they do have an effect on an organism's behaviour. Basic p-d receptivities then appear to provide an example of states with some kind of intentional content, at least being *for* something, without prior benefit of a Design Solution to give them that content.

One might well continue by asking what it is that makes p-d drives effective (in making some conditions reinforcing) whereas n/n drives are not similarly effective. A reasonable answer would be that p-d drives have a *phenomenology*. I mean this in the minimal sense that a state such as thirst involves awareness. That an fluid-deprived organism feels the difference before and after executing some exploratory routine that leads to the obtaining of water is part of what accounts for the increased probability of the routine being repeated.

I agree with Dretske that it would be unreasonable to define the content of a previously unsatisfied p-d drive in terms of some particular outcome from among those that would satisfy it. Nevertheless, I believe that such states are reasonably construed as intentional and representational, and that they they are so prior to having participated in a Design Solution. What they represent are states of the organism's own internal systems. Hunger or thirst are, for example, an organism's apprehension of states of the organism's digestive or metabolic systems. They are representations of internal conditions of the organism.

Reason (2)

Training by reinforcing requires some measure of repetition. A subject exposed to a negative

condition in thirty randomly chosen and quite different situations may learn to cower but it is unlikely to acquire any sort of *discriminative* behaviour. There needs to be some common or similar feature to a set of cases for an organism to learn to behave differentially with respect to them. For a hungry Sniffy to engage in bar-pressing only in some circumstances but not others, one needs to establish a conditioned stimulus. In the "Sniffy" example described earlier, this conditioned stimulus was a red light.

Simply illuminating a red light in cases when food is in Sniffy's hopper is not immediately helpful, of course. A red light illuminated in a nearby room will not help. The light has to be accessible to the subject, for it is the discriminations that Sniffy makes that results in the red light's illumination being a conditioned stimulus. Similarly the feature of the cases which separated instances of food being available for those in which food is unavailable must be something that Sniffy can discriminate. Installing a light in Sniffy's environment which is sometimes illuminated, sometimes not, stands no chance of conditioning Sniffy's behaviour if Sniffy either 1) cannot differentiate the light's being on or 2) *cannot reidentify cases of the light's being on*, in the sense of at some level recognizing these as of the same type. If for Sniffy cases in which the light is illuminated do not tend to be significantly different from those in which the light is not illuminated, or if each such case is a novel situation, Sniffy's bar-pressing movements will never fall under the control of an internal state dependent on the illumination of the light. Dretske clearly acknowledges point 1) (in a very strong way through his indication requirements) but overlooks point 2). Reidentification, awareness of features of the perceptual array and awareness of distinct occurrences as of the same type, is requirement for and is not a product of developmental Design Solutions. This also would seem to require some intentional property: "perceiving as".

The current suspicion that Dretske's model may rely on intentionality in order to explain it is further fed by a discussion of procedural belief which I raised at the end of my section on motivation and behaviour[8]. Recall Dretske's description of Sniffy's acquisition of a new behaviour structure: in virtue of a certain environmental arrangement, food being placed in a hopper when a light is illuminated, Sniffy acquires a procedural belief: pressing the bar when the light is illuminated will yield food. This is an expression of a dispositional state, a kind of rule.

In Dretske's description of the rat being trained to respond discriminatively to the illumination of the red light, the behaviour-structuring that occurs is that of instances of an internal state caused by the illumination of the light, together with hunger, coming to cause bar-pressing movements. It looks as if the semantic content which is supposed to be the product of this structuring is the procedural belief that pressing the bar when the light is illuminated will release food from the hopper. In the rat's case, we are to suppose that this is a "single-tracked" dispositional state, an implicit belief the content of which is unavailable for other cognitive activity.

Dretske clearly claims at one point that Sniffy's procedural belief, the intentional state which comes out of the behaviour-structuring comes about in virtue of Sniffy's *knowing* that the light is illuminated and having a desire for food. One might suggest that the reference to prior intentional states is incidental, a simple bit of carelessness. However, it becomes clear in Dretske's exposition that reference to these states is intended to be quite essential. He has to distinguish between simple physical dispositional states which are non-intentional, and those dispositional states of organisms which are intentional. It is precisely Sniffy's procedural belief's being a product of a structuring involving a perceptual belief (or piece of knowledge) and a desire that is supposed to make the difference:

> A rule or disposition that says, simply, "Produce movement M when conditions F and D obtain" is *not* an implicit belief. It is, at best, merely a regularity of some kind. My car "follows" the rule "Start the engine when the key is turned on, the battery is charged, and there is gas in the tank." But this is not a belief. In order to qualify as a *belief*, the disposition or rule must be defined over intentional elements (beliefs and desires) *qua* intentional elements. So, for example, the rule that says "Produce movement M when you believe (or represent things as) F and when you want R" gives rise to the implicit belief that M will yield R in conditions F. (Dretske 1988, p.118)

Moreover, acquisition of procedural beliefs appears to be quite basic to Dretske's developmental solution. He writes,

> Any animal that has been trained to do something in a certain set of conditions in order to secure a certain result has what we might call *procedural knowledge* -- a knowledge of *how* to achieve those results, a knowledge of what to do and when to do it. (Dretske 1988a, p. 116)

A little farther he continues as follows:

> This practical knowledge, this cognitive skill, is the *inevitable accompaniment* [my emphasis] of instrumental learning: you can't by appropriate rewards and punishments, shape an animal's responses to stimuli without also teaching it, at some level, *that* those responses lead, in these conditions, to those rewards and punishments. (Dretske 1988a, p. 117)

This is puzzling. On the one hand, I think one is led to believe that the behaviour-structuring story is the one which explains how any of our internal states get to be representational, or otherwise intentional states at all. Sniffy's coming to press the food-release bar discriminatively is what is supposed to account for B-tokens getting promoted from being F-indicators to F-representations. This promotion is what made such states beliefs that something is F. The present explanation would suggest that Sniffy's training results in the procedural belief, and what accounts for this is that it is defined over a prior perceptual belief and a desire.

Are these intentional states the simultaneous product of Sniffy's training? If so, there does not seem to be any way of fitting the perceptual belief that the light is illuminated into the

explanation of how the procedural belief comes about. It cannot be the case that Sniffy's having an internal state which constitutes believing that the light is illuminated both explains and comes about contemporaneously with the procedural belief. If the internal state B has already acquired the function of indicating that the light is on, then what accounts for its having acquired this function? The answer cannot be a selectional Design Solution, for the content of a selectionally acquired representational state, on Dretske's theory, would still be semantically inactive. It would be of a sort that, *qua* representation is irrelevant to that which an individual system does. In that case the "procedural belief" that Sniffy acquires would not be defined over intentional elements as *such*. Even if the answer is another appeal to the developmental solution, this appears to postpone the questions only. A prior behaviour-structuring which did promote B to an F-indicator will still be regarded as the formation of some sort of procedural belief, and the same questions arise again.

Minimally, a naturalistic theory about doxastic intentional states should be expected to account for these questions in a way which does not essentially appeal to other such states *as intentional states*. Dretske's discussion of procedural belief-acquisition appears essentially to require reference to perceptual belief. There are no obvious resources in his theory to account for the formation of these prior perceptual beliefs. The theory thus commits a circularity somewhat more general in form than that which Lynne Rudder Baker has tried to show: It is the circularity of accounting for basic procedural beliefs by appeal to prior perceptual beliefs, where there are no resources to explain what gives rise to the latter.

c. The End

The foregoing observations on how motivation and perceptual discrimination figure into the developmental Design Solution are intended to be basic reflections of a conceptual rather than empirical nature on what reinforcement seems to require. Receptivities are required. Basic p-d receptivities (for food, water, sex, for example) are to be understood as having their intentional "*for*-ness" prior to the occurrence of a satisfying condition. They are further plausibly construed as having some representational dimension: they represent states of the organism's own internal systems. Conditioning behaviour by exposing an organism to a set of (perceptual) situations of the same type will not have the desired effect unless the organism subsumes those conditioned cases under a type: there must be some identification of numerically disparate circumstances (which also may differ in some qualitative aspects) as being of the same kind. In order to get Sniffy to perceive circumstances in which the light is illuminated as "food affording" (as a case in which bar-pressing is worth the trouble), it must first be possible for Sniffy to perceive different illumination-instances as instances of a like kind.

This is not meant to be a reference to any very impressive capacity. I do not mean, for example, that rats must be reflectively conscious or aware of perceiving different things as being things of the same kind. All that I require is discrimination of aspects of visual arrays and sameness/difference perception across discreet perceptual situations.

I have followed up here on the question of whether Dretske's theory might commit a kind of general circularity. I have tried to make it at least initially plausible to think that the developmental solution presupposes what it is supposed to explain. "Perception as" and desire are reasonably construed as both mental and representational if any processes are. Besides, Dretske is committed

to the notion that "all mental facts are representational facts", so it is hard to see how he would avoid conceding their representationality. Moreover, I have tried to show that the developmental solution presupposes representationality in a way that is not linked to any particular example of the developmental solution. It does not matter especially which model is supposed to provide the "ground level" of representation, the Random Experimentation process, the Mutation process, or some other that I have failed to imagine. It does not matter because the source of the "creeping representationality" is reinforcement itself and this is presupposed by any model of how the developmental solution solves the Design Problem.

Dretske might try to dismiss the seriousness of the circularity I am alleging. One might reason that while my points about reinforcement are true, the role that prior representation plays in content-determining by development is a bit-player role. Possibly subsequent theorizing will reveal some modifications that will clear up the issue.

Also, recall that Dretske maintains a very strong notion of indication. In the face of criticisms by Slater that Dretske's indication is no principle of operant conditioning, Dretske responds by differentiating "mere operant conditioning" from cases of operant conditioning which yield genuine concepts. Taking this cue, one might reason in the following way:

> Operant conditioning, because dependent on reinforcement, does presuppose some mental bits and pieces. Operant conditioning by itself is not a content-determiner, however. Content is determined by H-indication, and content that is active is that which is so determined in cases of operant conditioning. Such a view thus still substantially respects Dretske's original theoretical presuppositions. Indication is the dominant factor in content-determination and it is a relational property

having external, environmental conditions at one end.

Now we are in a position to see how Dretske is led to an uncomfortable dilemma: either maintain or weaken the standards for indication (and H-indication). On the one hand, one can maintain a strong Dretskean notion of indication. However, this leads to very difficult consequences:

a) Through arguments by Slater (and Godfrey-Smith) there is a substantial gap between behaviour that is conditioned and conditioned behaviour which supports an internal state as an indicator-function.

b) States having been indicators, even where H-indication is satisfied is, I have argued not a causal fact. This means that representation does not even get a vicarious causal role in Dretske's theory.

c) That there is no representation without H-indication leads to the consequence that, with respect to much of our early, conditioned behaviour we are effectively "Swamp-persons" since most of such behaviour will have been conditioned in the absence of the appropriate indicator-history. What is more, learning about the genuine beliefs of others will no longer be a simple matter of (say) finding out what they say they believe. It will involve checking that the appropriate historical conditions under which that verbal behaviour was acquired has been satisfied.

On the other hand, one can give up on Dretske's strong concept of indication, justifying this, as Slater and Godfrey-Smith do, on the grounds that such a condition has little to do with operant conditioning anyway. Godfrey-Smith's reasoning, however, works not simply against Dretske's strong version of indication but against "reliabilist" accounts of content-determination generally. Given Dretske's approach to the notion of representation, content-determination is more plausibly

handled counter-factually: what must have been present in order that an organism adopts a behaviour. That is, C represents that F iff C has come to cause M and C would not have come to cause M had C not occurred in the presence of F. The shift in criterion, however, takes the weight of content-determination off of the environmental relation and raises the centrality of such considerations as an organism's receptivities, the discriminations it can make and how it can subsume different instances under kinds, and even assessment of consequences. Thus the mental "bits and pieces" which seem to be required by behaviour-structuring based on reinforcement are no longer bits and pieces. Dretske's functional interpretation of representation rests on prior processes which are themselves arguably intentional and representational.

Dretske's approach has a number of noteworthy features. As previously mentioned, it is an approach that is materialist in an ontologically reductive sense, and the content of intentional states is taken to be relational in nature. There are a couple of additional presumptions that are built in to the approach. The first is that subsumption of all sorts of intentional phenomena under one explanatory framework is desirable. The awkwardness with which Dretske's informational framework accomodates motivational states serves to underscore, once again, the point that the properties of motivational states and how they figure in the possibility and the shaping of systems with original intentionality are topics deserving of independent and thorough treatment. In fact, I am inclined to think that a better understanding of how genuine cognitive systems are engaged by and driven by receptivities to interact with a local environment is a fundamental part of the effort to understand the intentionality of epistemically relevant states.

The second presumption is that the nature of the sort of *original* intentionality which cognitive systems have can be adequately analyzed by thinking about what makes a given (type of)

state intentional. This seems to preclude the idea that there is something about the system *itself* that is intentional, a system which is capable of trading in an economy of directed, representational states and is capable of *naturalistically* forming or acquiring such states. It is then a natural move to make a distinction, for example, between the physical constitution of a system which has intentional/representational states and whatever it is that *makes* those states intentional or representational. An approach in the relationalist tradition of behaviourism and functionalism, a tradition to which Dretske's indicator theory belongs, then begins to look like the only reasonable tack to take on meaning-related properties.

Dretske's idea of accounting for the acquisition of "active" intentional properties through operant conditioning is interestingly reminiscent of Hume on induction. One may think of Hume as interested in the problem of how simple perceptual input manages to acquire significance for us beyond being a singular sensory "impression". Why do we generalize from what is given in a perceptual situation? His answer is that we acquire "habits" or "customs" as a result of repeated observation of conjunctions of types of events:

> Custom is that principle, by which this correspondence has been effected; so necessary to the subsistence of our species, and the regulation of our conduct, in every circumstance and occurrence of human life. Had not the presence of an object instantly excited the idea of those objects, commonly conjoined with it, all our knowledge must have been limited to the narrow sphere of our memory and senses; and we should never have been able to adjust means to ends, or employ our natural powers, either to the producing of good, or avoiding of evil. (Hume 1975, sec. 44)

In Hume's view, the basis for our generalizations is psychological and arises from, as it were, "repeated trials", repeated experiential exposure to regularly conjoined events. Dretske's picture of how the semantics of intentional states arise and have a role to play in behaviour is analogously based in trial learning.

A modern improvement on Hume's picture of how we come to make the "inductive move" occurs in Nelson Goodman:

> Hume thought of the mind as being set in motion making predictions by, and in accordance with, regularities in what it observed. This left him with the problem of differentiating between the regularities that do and those that do not thus set the mind in motion. We, on the contrary, regard the mind as in motion from the start, striking out with spontaneous predictions in dozens of directions, gradually rectifying and channelling its predictive processes. (Goodman 1983, p. 87)

Perhaps an analogous "re-orientation" is in order in coming to grips with the nature of intentionality. Why does Dretske's approach appear not to get to the heart of the issue? One suggestion is the it is in virtue of our being cognitive beings, the kinds of beings with the capacities and achievements requisite for taking an interested perspective on a local environment, that we are systems that "trade" in intentional or other representational states. In "Goodmanesque" terms, in relation to being a system of meaning-bearing states, the "mind is in motion from the start". Rather than representationality of or directedness toward an environment of *states* following from behaviour-structuring, the development of behaviour follows from the former. Cognitive beings such as ourselves are inherent "view-takers" whose behaviour develops from initial motivations and epistemically relevant states, which are in turn adjusted in light of the consequences of behaviour.

Of course these comments are highly speculative and do not at all constitute an alternative theory. The musings here do raise an interesting possibility, however. Suppose that one wants to maintain (what is arguably one of Dretske's most plausible commitments) the commitment to ontological materialism. In terms of the analogy drawn here with Goodman, the mind which is in motion from the start is *the brain*. So if there are capacities and and achievements in virtue of

which cognitive beings are intentional (and in virtue of which we take an interest in and interact with an environment), these are capacities and achievements of brains. To the extent that such capacities are meaning-involving, the having of such original meaning properties may require rethinking of the common contemporary assumption that, as neurophysiological systems, cognitive systems are *merely* physical, at best syntactic, systems, and are certainly non-semantic ones.

NOTES

1. This is a point over which Dretske may be changing. In his more recent work, Naturalizing the Mind, he is endeavouring to give a unified treatment of mental phenomena, including both the phenomenal as well as the intentional. Consideration of phenomenal states (eg. perceptual experience), how they get their phenomenal content and what role it plays leads one naturally to suppose both that their content is explanatorily relevant and that it is derived selectionally, not developmentally. What one believes seems partly grounded in one's perceptual experiences. This would lead to the view that selectionally acquired (phenomenal) contents at least are "to or for" the cognitive system in which the occur, which conflicts with Dretske's earlier, official view that selectionally acquired contents are not thus efficacious. In conversation (Oct. 1996), Dretske did indicate that he was aware of the conflict and was reconsidering the view that the selectional solution does not yield causally active content.

2. Mutation is caricatured here. In actual cases series of very small mutations may take place over generations and which small mutations which happen to survive will have some impact on which of some range of possible solutions actually come to be instantiated. Some "ideal" solution may never come about because the mutation-steps that would have led to it were not realized, or some were but they did not reproduce. Nor is it clear to me that every mutational step has to be itself advantageous in order to proliferate.

3. Dretske is of course aware of this. It is not clear whether he regards it as a serious difference.

4. He argues this in relation to selectional as well as developmental solutions.

5. Notably chapter five.

6. Godfrey-Smith would most likely object to the rough characterization of natural selection and what it explains that I have given here. A mutation to some complex function does not occur in a single step but rather through a series of very minor mutations which, over successive changes, make the advantageous mutation "mutationally accessible", to use Godfrey-Smith's term. This complication is of no importance to the points that I will make.

7. "No more reinforcing" in the sense that it is not more reliably correlated with instances of rabbit-catchings. I take it that Godfrey-Smith's well-taken point is that a behaviour can be of some advantage or benefit to an organism without it being either a) well correlated with the condition that reinforces it, or b) better correlated with the reinforcing condition (as compared to alternative behaviours).

8. I do not of course mean to suggest that Dretske is unaware of distinction between simple and conceptual perception. Nor am I endorsing a view on which all perceiving is "perceiving as". The present point is a concern about what may be required in order for operant learning to work and what I mean to suggest is that Dretske may be over-estimating the extent to which such a process can be, as natural selection is claimed to be, a "blind" process from which functions can arise.

9. Chapter four, section (c).

WORKS CITED

Baker, Lynne Rudder. 1991. Dretske on the Explanatory Role of Belief. Philosophical Studies 63: 99-111.

Baker, Lynne Rudder. 1995. *Explaining Attitudes*. Cambridge: Cambridge University Press.

Bratman, Michael E.. 1990. Dretske's Desires. Philosophy and Phenomenological Research 50: 795-800.

Cummins, Robert. 1991. The Role of Mental Meaning in Psychological Explanation. In *Dretske and his Critics*, ed. Brian P. MacLaughlin, 102-117. Oxford: Blackwell.

Davidson, Donald. 1987. Knowing One's Own Mind. Proceedings and Addresses of the American Philosophical Association 60: 441-58.

Dretske, Fred. 1981. *Knowledge and the Flow of Information*. Cambridge, Massachusetts: MIT Press. Paperback Edition.

Dretske, Fred. 1986. Misrepresentation. In *Belief: Form and Content*, ed. Radu J. Bogdan, 17-36. Oxford: Clarendon Press. Paperback edition.

Dretske, Fred. 1988a. *Explaining Behavior*. Cambridge, Massachusetts: MIT Press.

Dretske, Fred. 1990. Reply to Reviewers. Philosophy and Phenomenological Research 50: 819-839.

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

Dretske, Fred. 1991b. Dretske's Replies. In *Dretske and His Critics*, ed. Brian P. MacLauglin, 180-221. Oxford: Blackwell.

Dretske, Fred. 1994a. Mind and Brain. In *The Mind-Body Problem: A Guide to the Current Debate*, ed. Warner, Richard and Szubka,Tadeusz, 131-136. Oxford: Blackwell.

Dretske, Fred. 1994b. Reply to Slater and Garcia-Carpintero. Mind and Language 9: 203-208.

Dretske, Fred. 1995. *Naturalizing the Mind*. Cambridge, Massachusetts: MIT Press.

Dretske, Fred. 1996a. Absent Qualia. Mind and Language 11: 78-85.

Dretske, Fred. 1996b. How Reasons Explain Behavior: Reply to Melnyk and Noordhof. Mind and Language 11: 223-229.

Field, Hartry. 1990. "Narrow Aspects of Intentionality and the Information-Theoretic Approach to Content. In *Information, Semantics, and Epistemology*, ed. Villanueva, Enrique, 102-116. Oxford: Blackwell.

Godfrey-Smith, Peter. 1992. Indication and Adaptation. Synthese 92: 283-312.

Goodman, Nelson. 1983. *Fact, Fiction, and Forecast* (4th edition). Cambridge, Massachusetts: Harvard University Press.

Horgan, Terence. 1991. Actions, Reasons, and the Explanatory Role of Content. In *Dretske and His Critics*, ed. Brian P. MacLauglin, 73-101. Oxford: Blackwell.

Hume, David. 1975. Enquiries Concerning Human Understanding and Concerning the Principles of Morals, ed. P.H. Nidditch. Oxford: Clarendon Press.

Kim, Jaegwon. 1991. Dretske on How Reasons Explain Behavior. In *Dretske and His Critics*, ed. Brian P. MacLauglin, 53-72. Oxford: Blackwell.

Lloyd, Dan. 1989. *Simple Minds*. Cambridge, Massachusetts: MIT Press.

Noordhof, Paul. 1996. Accidental Associations, Local Potency, and a Dilemma for Dretske. Mind and Language 11: 216-222.

Parfit, Derek. 1971. Personal Identity. Philosophical Review 80: 3-27.

Ross, Don, and Zawidzki, Tad Jr.. 1994. Information and Teleosemantics. The Southern Journal of Philosophy 32: 393-419.

Ryle, Gilbert. 1949. *The Concept of Mind*. London: Hutchinson.

Searle, John. 1983. *Intentionality*. Cambridge: Cambridge University Press.

Slater, Carol. 1994. Discrimination Without Indication: Why Dretske Can't Lean on Learning. Mind and Language 9: 163-180.

Stalnaker, Robert. 1984. *Inquiry*. Cambridge, Massachusetts: MIT Press.

Stampe, Dennis. 1977. Toward a Causal Theory of Linguistic Representation. In *Midwest Studies in Philosophy II*, ed. P.A. French, T.E. Uehling Jr., H.K. Wettstein, 81-102. Minneapolis: University of Minnesota Press.

Stich, Stephen P.. 1983. From *Folk Psychology to Cognitive Science: The Case Against Belief*. Cambridge, Massachusetts: MIT Press. Paperback edition.

Strawson, Galen. 1994. *Mental Reality*. Cambridge, Massachusetts: MIT Press.

Additional Works Consulted

Adams, Frederick. Causal Contents. In *Dretske and his Critics*, ed. Brian P. MacLaughlin, 131-156. Oxford: Blackwell, 1991.

Aikens, Kathleen. Of Sensory Systems and the "Aboutness" of Mental States. Journal of Philosophy 93 (1996): 337-372.

Antony, Louise. Equal Rights for Swamp-persons. Mind and Language 411 (1996): 70-75.

Bach, Kent. De Re Belief and Methodological Solipsism. In *Thought and Content: Essays on Intentionality*, ed. Andrew J. Woodfield, 121-151. Oxford: Clarendon Press, 1982.

Baker, Lynn Rudder. *Saving Belief*. Princeton: Princeton University Press, 1987.

Bogdan, Radu J.. The Importance of Belief. In *Belief: Form, Content and Function*, ed. Bogdan, Radu J., 1-16. Oxford: Clarendon Press, 1986.

Boghossian, Paul. The Status of Content. Philosophical Review 99 (1990): 157-184.

Burge, Tyler. Individualism and the Mental. In *Midwest Studies in Philosophy IV*, ed. P.A. French, T.E. Uehling and H.K. Wettstein, 73-121. Mineapolis: University of Minnesota Press, 1979.

Burge, Tyler. Other Bodies. In *Thought and Content: Essays on Intentionality*, ed. Andrew J. Woodfield, 97-120. Oxford: Clarendon Press, 1982.

Burge, Tyler. Individualism and Psychology. Philosophical Review 95 (1986): 3-45.

Churchland, Paul. *Matter and Consciousness* Cambridge, Massachusetts: MIT Press,1984.

Cummins, Robert. *Meaning and Mental Representation*. Cambridge, Massachusetts: MIT Press, 1989.

Dennett, Daniel. Beyond Belief. In *Thought and Content: Essays on Intentionality*, ed. Andrew J. Woodfield, 1-95. Oxford: Clarendon Press, 1982.

Dennett, Daniel. Ways of Establishing Harmony. In *Dretske and His Critics*, ed. Brian P.

MacLaughlin, 118-130. Oxford: Blackwell, 1991.

Dennett, Daniel. "Cow-sharks, Magnets, and Swampman". Mind and Language 11 (1996): 76-77.

Dretske, Fred. Putting Information to Work. In Information, *Language, and Cognition*, ed. P.P. Hanson, 112-124. Vancouver: University of British Columbia Press, 1988.

Dretske, Fred. The Explanatory Role of Content. In *Contents of Thought*, ed. Grimm, R. and Merrill, D.D., 31-61. Tucson: University of Arizona Press, 1988.

Dretske, Fred. Precis of "Explaining Behavior: Reasons in a World of Causes". Philosophy and Phenomenological Research 50 (1990): 783-786.

Dretske, Fred. The Nature of Thought. Philosophical Studies 70 (1993): 185-199.

Dretske, Fred. Mental Events as Structuring Causes of Behavior. In *Mental Causation*, ed. Heil, John and Mele, Alfred, 121-136. Oxford: Oxford University Press, 1993.

Egan, Francis. Must Psychology Be Individualistic? Philosphical Review 100 (1991): 179-203.

Egan, Francis. Individualism, Computation, and Perceptual Content. Mind 101 (1992): 443-459.

Egan, Francis. Computation and Content. Philosophical Review 104 (1993): 181-203.

Field, Hartry. Narrow Aspects of Intentionality and the Information-Theoretic Approach to Content. In *Information, Semantics, and Epistemology*, ed. Villanueva, Enrique, 102-116. Oxford: Blackwell, 1990.

Field, Hartry. Deflationist Views of Meaning and Content. Mind 103 (1994): 249-285.

Fodor, Jerry. *Psychosemantics*. Cambridge, Massachusetts: MIT Press, 1987.

Fodor, Jerry. A Modal Argument for Narrow Content. The Journal of Philosophy 88 (1991): 5-26.

Fodor, Jerry. *The Elm and the Expert*. Cambridge, Massachusetts: MIT Press, 1995.

Francescotti, Robert M.. Qualitative Beliefs, Wide Content, and Wide Behavior. Nous 28 (1994): 396-404.

Garcia-Carpintero, Manuel. Dretske on the Causal Efficacy of Meaning. Mind and Language 9 (1994): 181-202.

Gauker, Christopher. Mental Content and the Division of Epistemic Labour. Australasian Journal of Philosophy 69 (1991): 302-318.

Godfrey-Smith, Peter. Misinformation. Canadian Journal of Philosophy 19 (1989): 533-550.

Hanson, Philip P.. Are Contexts Semantic Determinants? Canadian Journal of Philosophy (suppl.) 6 (1980): 161-183

Levine, Joe. SwampJoe: Mind or Simulation? Mind and Language 11 (1996): 86-91

Ludwig, Kirk. Duplicating Thoughts. Mind and Language 11 (1996): 92-102.

McGinn, Colin. The Structure of Content. In *Thought and Content: Essays on Intentionality*, ed. Andrew J. Woodfield, 207-258. Oxford: Clarendon Press, 1982.

McGinn, Colin. Conceptual Causation: Some Elementary Reflections. Mind 100 (1991): 573-586.

McLaughlin, Brian P.. Belief Individuation and Dretske on Naturalizing Content. In *Dretske and his Critics*, ed. Brian P. MacLaughlin, 157-179. Oxford: Blackwell, 1991.

Melnyk, Andrew. The Prospects for Dretske's Account of the Explanatory Role of Belief. Mind and Language 11 (1996): 203-215.

Miller, Richard B.. A Purely Causal Solution to One of the Qua Problems. Australasian Journal of Philosophy 70 (1992): 425-434.

Millikan, Ruth Garrett. Seismograph Readings for "Explaining Behavior". Philosophy and Phenomenological Research 50 (1990): 807-812.

Millikan, Ruth Garrett. On Swampkinds. Mind and Language 11 (1996): 103-117.

Morton, Adam. Folk Psychology is not a Predictive Device. Mind 105 (1996): 119-137.

Neander, Karen. Swampman Meets Swampcow. Mind and Language 11 (1996): 118-129.

Neander, Karen. Dretske's Innate Modesty. Australasian Journal of Philosophy 74 (1996): 258-274.

Papineau, David. Doubtful Intuitions. Mind and Language 11 (1996): 130-132.

Peacocke, Christopher. The Metaphysics of Concepts. Mind 100 (1991): 525-545.

Putnam, Hilary. The Meaning of 'Meaning'. In *Language, Mind, and Knowledge: Minnesota*

Studies on the Philosophy of Science VII, ed. Keith Gunderson. Minneapolis: University of Minnesota Press, 1975.

Putnam, Hilary. *Representation and Reality*. Cambridge, Massachusetts: MIT Press, 1989.

Schiffer, Steven. Belief Ascription. The Journal of Philosophy 89 (1992): 499-521.

Stampe, Dennis. Desires as Reasons -- Discussion Notes on Fred Dretske's "Explaining Behavior: Reasons in a World of Causes". Philosophy and Phenomenological Research 50 (1990): 787-793.

Sterelny, Kim. *A Representational Theory of Mind*. London: Blackwell, 1990.

Stich, Stephen P.. On the Ascription of Content. In *Thought and Content: Essays on Intentionality*, ed. Andrew J. Woodfield, 153-206. Oxford: Clarendon Press, 1982.

Stich, Stephen P.. Building Belief: Some Queries about Representation, Indication, and Function. Philosophy and Phenomenological Research 50 (1990): 801-806.

Tuomela, Raimo. Are Reason-Explanations Explanation by Means of Structuring Causes? Philosophy and Phenomenological Research 50 (1990): 813-818.

Woodfield, Andrew. On Specifying the Contents of Thoughts. In *Thought and Content: Essays on Intentionality*, ed. Andrew J. Woodfield, 259-297. Oxford: Clarendon Press, 1982.

Woodfield, Andrew. Conceptions. Mind 100 (1991): 547-572.