2009

Naturalism, Norms, and Intentionality: The Substitutionary Aspects of Mental Representations

Philip Kuchar

Naturalism, Norms, and Intentionality: The Substitutionary Aspects of Mental Representations

(Spine Title: Naturalism, Norms, and Mental Content)

(Thesis Format: Monograph)

by

Philip Kuchar

Graduate Program in Philosophy

A thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

Abstract

Naturalistic philosophical theories of semantic content, such as the influential ones proposed by Fodor, Dretske, and Millikan, on which I focus, are typically judged on whether they account for the differences between a semantic relation and a more naturally fundamental relation, such as a causal or an informational relation. One of the apparent differences is that a semantic relation, between a symbol and something else, has better and worse ways of being instantiated, which is to say that the semantic relation seems to be normatively determined. But these theorists have also tended to identify the semantic relation with one of those naturally more fundamental relations which isn't normatively determined, adding a theory of content determinacy to a metaphysical account of the relation.

I argue that because these theorists take this approach to explaining semantic content, their theories tend to have internal contradictions. These theorists need to explain the apparent normative determination in a way that is consistent with their claim that a semantic relation is a naturally fundamental relation that is not at all normative. Thus, Millikan, for example, posits a purposive function as the determinant, but a function which she claims is only descriptively, or objectively, normative. I argue, though, that this function turns out to be prescriptively normative, and that her metaphysical claims about the nature of a semantic relation conflict with her account of the purposive function.

I propose an alternative naturalistic strategy, one that takes naturalistic methodology rather than ontology as the starting point in an explanation of semantic content, and one that can therefore afford to accept the role of prescriptive norms in determining this content. A philosophical naturalist takes for granted not just what scientific theories say, but the methods scientists use, including the use of explanatory models. These models have substitutionary aspects, which I argue are crucial to the intentionality of mental symbols in general and which aren't addressed by the other three theorists. I provide, then, a naturalistic, noncircular account of how a mental symbol's standing in for something else is determined by prescriptive norms.

Keywords

naturalism, normativity, intentionality, mental representation, Fodor, Dretske, Millikan

Contents

List of Tables

Abbreviations and Symbols

CP *ceteris paribus*

Abbreviations for Fodor's Theories:

ADT Basic Asymmetric Dependence Theory of Content

AHC The Actual History Condition added to ADT

LMT Theory of the Locking Mechanism for Concept Acquisition

Symbols in Dretske's Theory:

S an organism, especially one trained to be a semantic system

C an internal condition of *S* that causes *M*

M a bodily movement of *S*

F property indicated by *C*

R external condition that reinforces *S*'s behaviour

B belief

D desire

N neural type of which *C* is a token

Symbols in Millikan's Theory:

R	reproductively established family
m	token member of R
F	purposive function, proper to m
C	m's character, or set of properties, enabling m to perform F

Chapter 1

Introduction: Naturalism and Intentionality

What is it for one thing to be *about* something else? This is an old philosophical question. Speaking, writing, and thinking are ways of using symbols that bear a certain relation to other things just by being about them. But there seems no scientific explanation of this semantic relation, or of what philosophers often call "intentionality," because there are normative, correct and incorrect instantiations of symbols. Norms are taken to govern what *should* happen, and it's hard to see where these norms fit into the scientific picture of the natural world as full of events that simply *do* happen as a matter of brute fact. There may only *seem* to be a normative dimension of nature, in that its appearance may be like that of a stick under water that appears bent but that isn't really so. Thus, a semantic relation might be contrasted with a more clearly objective, fundamentally physical relation, such as someone's eating of food. Eating is realized ultimately by material processes that can all be scientifically explained. But a symbol's being about something else doesn't seem to be similarly realized. This conflict between the commonsense view of the use of symbols, and the scientific picture of the world is one mark of a philosophical problem.[1]

Take, for example, a written word's reference to an object. Use of written

[1] Sellars (1962), for example, makes this point about philosophy, by speaking of the conflict between the "manifest image" and the "scientific image" of ourselves, where the former is how we appear naively to ourselves, and the latter is how we discover ourselves to be, using scientific methods.

language depends on the implicit assumption that certain marks on paper, such as "dog", are directed towards certain other things, in this case towards dogs. However, this relation of directedness doesn't seem realized ultimately by a physical process, which is to say that there seems to be no concrete *means* by which the marks on the page have this ability of being about something. Of course, the writing of the word is just such a process, as is the seeing of the word, but the marks on paper don't themselves point to dogs. Still, the writer treats the word as a symbol that has meaning, that bears the relation of being about dogs. Perhaps this linguistic relation derives from a deeper, mental one, such as the thought about dogs. By itself, "dog" might have no meaning, but in so far as the word causes people to think of dogs, the word might have determinate content. This raises the similar question of whether a thought's relation to dogs is fundamentally physical or whether the relation is, once again, just some sort of illusion. When a person thinks of dogs, there is some activity in the person's brain, but even were there a distinct neural state whenever the thought is had, the neurons wouldn't reach out and connect physically to dogs. Suppose the thought causes the thinker to pet a dog. The act of petting would be objective, but there are many reasons to think the thought's having content isn't just the ability of the neurons to cause movement of some type. For one thing, the thought is about dogs, not just the petting of dogs. Moreover, some mental symbols don't have behavioural consequences; others are about distant objects about which nothing can be done, or about different things despite having the same behavioural response to each; still others are about things that no longer exist or that never will exist, such as fictions, to which no objective connection can be made. And yet these symbols are all used as though they had content.

This philosophical problem, of reconciling the commonsense way of understanding symbols with the scientific way of understanding nature, should be divided into at least two subproblems. Indeed, the theories I consider in the following chapters assume there are these two different questions to answer. First, there is the question, "What is it for one thing to be about something, that is, what sort of relation is at issue here?" For example, one answer would be that the relation is a causal, mechanistic one. Second, there is the question, "What determines, or sets the limits of, that relation, so that each symbol gets its own content?" For example, the reason a certain thought is about dogs and not foxes might be that instances of the thought are caused only after an *infallible* detector goes to work, so that only dogs cause someone to have the thought. Assuming this is not so, and things other than dogs can cause someone to think of dogs, the determinant of the causal relation would have to be more complicated. Thus, were the detector fallible, after all, the detector might have the *function* of distinguishing between dogs and foxes, and there might be ideal conditions under which the detector fulfils its purpose. In this case, what would set the limits of the relation between a symbol and its content, despite the variety of situations in which the relation is found, is the detector's function.

There have been at least three ways of dealing philosophically with the overall problem of explaining content. First, there are *non-naturalistic dualists* who posit an immaterial substance that is supposed to account both for the semantic relation and for the naturalist's inability to explain it. In this case, a symbol's being about something would be real rather than somehow illusory, despite the lack of any concrete way of realizing the relation. A symbol's ability to be about something would be comparable to a

ghost's alleged ability to manifest itself in the natural world. Second, there are *naturalistic internalists*, who argue that, because of the limits of scientific methods, there is no naturalistic explanation to be had of reference or of the semantic aspect of symbols; the part of the use of symbols that can be explained naturalistically is just the set of internal, syntactic relations between them. Third, there are *naturalistic externalists*, who argue that there can be a naturalistic account of the semantic relation, because this relation is identifiable with some natural, scientifically recognizable relation. I'll say something here about each of these three approaches.

I reject non-naturalistic dualism, because I think there is a naturalistic explanation of a symbol's having content, or of what philosophers have called "intentionality." By "naturalistic", I mean, if not a scientific explanation, then a philosophical one that is continuous with scientific explanations. Internalists such as Chomsky (1995) disagree with this broad definition of "naturalistic," and argue on methodological grounds that there can be no naturalistic account of semantic relations. I can't respond at length here to Chomsky's case against externalism, but I do want to say something about his view of naturalistic methods, since I too appeal to these methods in Chapter 5. According to Chomsky, naturalism is just a set of methods that have proven highly successful in the sciences, and one of these methods is to liberate inquiry from the dogmas of commonsense, regardless of the subject matter. Thus, even when studying human beings, theorists need not be bound by intuitions or by popular opinions; after all, in many cases scientific truths have been counterintuitive. However, what Chomsky calls methodological dualists, such as philosophers, claim that there is one method for studying human beings, including our languages and our thought processes, and another for

studying nature. In particular, commonsense is supposed to have a special role in our self-understanding even if it has no such role in physics, cosmology, or in any other inquiry into things other than human beings.

Chomsky limits naturalistic inquiry not only to science, it seems, but to *physical analysis*. Thus, he says, the quest for naturalistic inquiry "seeks to account for some aspects of the world on the basis of usually hidden structures and explanatory principles" (28). A genuine science finds data in what can be carefully observed, and then posits underlying, often unobservable entities and processes to account for those data. And so, regardless of whether words or thoughts appear to bear the relation of being about certain things, a scientific explanation of symbols won't likely stop at that appearance. On the contrary, a linguist, for example, posits the underlying structure of the language faculty. Science is limited to studying what can be measured, and thus the language faculty is defined by syntactic, or formal, symbol-to-symbol relations, not by immaterial symbol-to-world relations. And so, "general issues of intentionality, including those of language use, cannot reasonably be assumed to fall within naturalistic inquiry" (27). From Chomsky's narrow view of naturalistic methods, the mind should be studied as a physical system. Thus, he grants that there is a triadic relation between speakers, words, and other things in the world, assuming the semantic relation between words and what they are about may depend on a semantic relation between thoughts and what they are about. But he says this relation holds "in more or less the sense in which a relation holds of people, hands, and rocks, in that I can use my hand to pick up a rock" (44).

If Chomsky is right, there's very little a philosopher can do to add to a naturalistic account of language and the mind. But I think Chomsky's scientistic notion of naturalistic

methods is wrongheaded. In ancient animistic interpretations of nature, intuition and commonsense were indeed overextended, on a sort of hypothesis that minds are found everywhere, even in stones, rivers, and clouds. The scientist's demonstration that there is a hidden microworld of nonliving material elements undermined this animistic, panpsychist interpretation, but not realism about mental properties, including semantic ones. Just because minds aren't everywhere doesn't mean they're not somewhere. Chomsky is arguing that the scientist's method of positing unobservable entities, to account for the way things appear to be on the surface, should replace intuition and commonsense in psychology, because this replacement happened in all of the sciences whose subject matter was formerly thought to be knowable by intuition and commonsense. But he overlooks the possibility that intuition and commonsense are legitimate methods of self-knowledge and of knowledge of other minds, methods that were overused on the naïve and false assumption that minds are everywhere. This is important, because philosophers take intuition and commonsense more seriously as sources of data than do scientists, and if there is some knowledge to be had using these methods in the special case of psychology, then philosophers *can* add to a naturalistic account of mental properties, again including semantic ones. Just because intuition and commonsense don't work in the study of things that aren't really minds, doesn't mean they're useless also in the study of things that are indeed minds and that seem to have semantic properties.

I'll clarify my point by returning to Chomsky's comparison of the relation between speakers, words, and other things in the world to the physical relation between people, hands, and rocks. According to intuition and commonsense, this comparison is

weak, because the former relation is normative whereas the latter is not. The semantic relation between the speaker's thought and what the thought is about seems unlike the relation between a hand and the rock that's picked up by the hand, in that the former relation holds even when what the thought is about doesn't exist, whereas the latter relation doesn't hold when the hand or the rock doesn't exist. The semantic relation seems normative in the same way that an obligation is normative, in that someone can be ethically related to the performance of an action, by having the obligation to perform it, even though the action is never performed. Indeed, the assumption of animism that was overturned by modern science is another example of the way semantic relations appear very different from physical ones. People thought minds were everywhere, and most of these minds turned out not to exist; nevertheless, the animist's thoughts managed to connect *semantically* to these nonexistent minds. A hand can't connect *physically* to a nonexistent rock.

Thus, there's at least this one apparent difference between semantic and physical relations. Rocks and rivers once appeared to intuition to be alive, and that intuition was shown to be in error, but it's a mistake to infer that intuition must likewise be mistaken when applied just to ourselves and to other minds.[2] Now, Chomsky's narrow point may

---

[2] There's a similar argument implicit in Churchland (1981). Churchland argues that there's no strong scientific basis for positing such things as beliefs, desires and other such mental entities, because folk psychology (FP) fails as a scientific theory. He points out that there's resistance to treating the commonsense picture of the mind as purely an empirical theory, because most folk theories, whether of diseases, the weather, or other natural phenomena, are now regarded by scientists as false. Thus, before Churchland can apply scientific standards to FP, he has to argue that the commonsense assumptions about the mind constitute an empirical theory.

There is, however, a relevant difference between FP and the other folk theories, which is also implicit in Churchland's case against folk psychology. As Churchland says, FP used to be applied to most of the natural world. But this amounts to saying that the other folk theories, which have been replaced by more sophisticated scientific theories, were derived from the overextended version of FP, that is, from the animistic worldview. And once again, there are obvious reasons why the application of FP to ourselves isn't likely to be just as groundless as the application of those folk theories, derived from animism, to most natural phenomena. There are, after all, obvious differences between things, such as rocks or clouds, that

be that there are still semantic relations even though there's no *scientific* account of them. This point may or may not be correct, depending on what scientific methods are taken to be. But in going on to say there's no *naturalistic* account of semantic relations, Chomsky is discounting the philosophical practice of reflecting on, as opposed to ignoring or explaining away, intuitions about how things appear. In particular, he's discounting a rational reconstruction of the intuition that minds use symbols that can relate to nonexistent things and that have normatively correct and incorrect instantiations, or uses. Again, such a reconstruction of intuitions may be groundless when the intuitions are about things that aren't minds, such as rocks or clouds. Still, a mind may have nonscientific, intuitive insight into itself and other minds. And assuming rational arguments that are consistent with scientific theories contribute to a broader naturalistic project, philosophers can add to a naturalistic account of mind and language, by trying to reconcile how symbols appear to commonsense, with the way scientists say certain things objectively are. With respect to a naturalistic account of semantic relations, this is largely a matter of making sense of how semantic relations can be natural while apparently having norms for determinants.

As for naturalistic externalism, on which I focus in this dissertation, I agree that intentionality is a real relation between symbols and what they are about, and that there is a naturalistic philosophical explanation of this relation. In the next three chapters, I examine three influential externalistic theories of content, those given by Fodor, Dretske, and Millikan. But I reject each of these explanations, not for the reasons given by others

---

are passively explained in sophisticated or else in commonsensical ways, and the things that are actively doing the explaining in the first place, what FP calls "minds." FP may not be the best theory of minds, but the argument, that FP is likely false because most folk theories are false, is weak. The subject matters of FP and of the other folk theories are very different, so the restricted application of FP isn't likely to be false just because the other folk theories are false.

in the literature over the last few decades, but because of the shared strategy these three theorists employ, of turning to fundamental elements of naturalistic ontology when saying what intentionality is, and of then adding an account of how semantic relations are determined. The problem with this strategy, as I say in Chapter 5, is that the answers given to these two questions tend to contradict each other, as I intend to show in the next three chapters. Also in Chapter 5, I propose a different strategy, one that promises to avoid the common failing of the other externalistic theories, by appealing to naturalistic methods instead of ontology— specifically, to the scientist's use of explanatory models. On this alternative approach, the way symbols are used as stand-ins for other things is seen as crucial to intentionality, and normative determinants of semantic relations aren't ruled out at the outset. In Chapter 6, I follow through with a sketch of how semantic relations are normatively determined, given their substitutionary aspects.

Chapter 2

Semantic Robustness and Asymmetric Dependence: Fodor's Theory of Content

2.1 Introduction

Fodor offers his own theory of content as an improvement on the sort of teleological theories, put forward by Dretske and by Millikan, that I consider in the next two chapters. There's a reason, though, I consider Fodor's theory first, and explaining why will double as an introduction to my critique of Fodor's theory. One of the questions about content that each of the three theorists tries to answer is of how a symbol's content is determined, which is to say how each symbol gets its own content. Like any other relation, the relation between a symbol and what the symbol is about must be distinguished from other relations by having some determinants that set the relation's limits, ensuring that there's some pattern in the way things bear the relation to each other. In the literature, this question of how semantic relations are determined is usually put in terms of the need to account for *misrepresentation.*[1] A thought experiment is posed: a creature has a symbol with content *X*, and the creature perceives *Y* which resembles *X* and thus which causes the creature to token, or to instantiate, the symbol for *X*. What makes the symbol such that it's semantically related to *X* even when a token of the symbol can

[1] See, for example, many of the papers in Stich and Warfield (1994).

be caused by *Y*? Whether it's a frog seeing a bee bee instead of a fly and flicking its tongue, or someone seeing a fox instead of a dog and mistaking the fox for a dog, this problem of misrepresentation is often taken to be central to naturalistic theories of content.

I think there are two reasons for this, although in the literature only one is typically given. The given reason is that naturalistic theories of content are often causal theories of some sort, and causal theories have difficulty accounting for misrepresentation, since whatever causes a symbol, at least in some fashion, would seem to be what the symbol is about, on such theories. The reason naturalistic theories are often causal ones is because of a strategy these theorists use, which is to identify intentionality, the aboutness relation, with some relation that is already well-understood, such as a relation that has a place in basic naturalistic ontology. The theorists then add an account of how this relation is specially determined to make for a semantic version of the relation. Thus, a semantic relation might be identified fundamentally with a causal or a nomic relation, since naturalism takes as granted causation and natural laws.

The problem is that semantic relations and causal or nomic relations don't seem determined in the same way, as the thought experiments about misrepresentation show. This points to the second reason for taking the problem of misrepresentation to be so important. The determinants of semantic relations seem to be *prescriptive norms*, which alone could allow for any mistake in a case of misrepresentation. A norm is a standard that is supposed to apply to something even when the thing doesn't follow the standard. In the same way, despite being sometimes caused by objects not in the symbol's extension, a symbol token seems nevertheless to have its own, independently established

content. To come back to Fodor, he appreciates both problems for naturalistic theories, the problem of accounting for misrepresentation in causal terms, and the problem of the independence of semantic and of causal or nomic relations. The other two theories I consider, given by Dretske and Millikan, deal with the second problem by offering explicit naturalistic accounts of the norms that seem to govern semantic relations. Thus they speak of purposive functions which are assumed to be only somehow descriptively (objectively) rather than prescriptively (subjectively) normative. Fodor's asymmetric dependence theory of content speaks not of norms, but of semantic robustness. As I say later, in section 5.2.1 (n.8), I think the best interpretation of semantic robustness introduces the normative aspect of semantic relations. But Fodor doesn't interpret his theory this way, so I postpone an explicit discussion of the normative determinants of semantic relations until the next two chapters.

However, Fodor's talk of semantic robustness ironically comes closest to saying that the determinants are prescriptive norms, and so I want to consider his theory first. "Semantic robustness" is a euphemism for the prescriptively normative aspect of symbols. Meanwhile, descriptively normative purposive functions are posited, by Dretske and Millikan, to account for the normative determinants of semantic relations without accounting for the prescriptive aspect of norms. In both cases, there's an avoidance of the issue of prescriptive norms. The difference is that, by saying that a naturalistic theory of content needs to explain semantic robustness, Fodor appreciates the *problem* of the normative determinants without offering a solution, such as a reductive or a nonreductive account of prescriptive norms. With their accounts of descriptively normative functions, Dretske and Millikan offer *solutions* without acknowledging the problem, since they

don't claim to be offering reductive accounts of prescriptive norms, and I argue later that the so-called descriptively normative functions turn out to be prescriptively normative. I argue also, in Chapter 5, that in each of the three cases, the claim that intentionality is a fundamental relation in naturalistic ontology conflicts with the account given of the determinants of semantic relations. This is because the latter account, one way or the other, introduces prescriptive norms which don't themselves have a clear place in basic naturalistic ontology. In Fodor's case, he takes his theory of content determinacy to account for semantic robustness, which is actually a quasi-normative property. I argue in this chapter that his theory of content determinacy contradicts his account of how the special nomic relations, which he says are the elements of semantic relations, are implemented by mechanisms, such as the mechanism featured in his theory of the acquisition of basic symbols.

I begin, then, with a detailed interpretation of semantic robustness (sections 2.2 to 2.5). I argue that there are two interpretations of what Fodor means by saying that semantic relations are "robust," and that the ambiguity can be seen in the philosophical reviews of Fodor's theory of content (2.3). The two interpretations, however, which I call the empirical and the metaphysical ones, present a dilemma for Fodor. On the empirical interpretation, semantic robustness poses no problem for teleological theories for content, whereas Fodor says semantic robustness is the main problem for those theories, calling for an alternative such as Fodor's own, asymmetric dependence theory (2.4). On the metaphysical interpretation, semantic robustness is indeed hard to explain in teleological terms, but for this very reason it will turn out that the claim that semantic content is metaphysically robust conflicts with any account of the mechanistic implementation of

symbols, including Fodor's own theory of how symbols are acquired by a locking mechanism (2.5). So while the metaphysical interpretation is preferable in at least one way, Fodor can't help himself to it. To support this latter point, I go on to show how Fodor's theory of content is designed to account for semantic robustness (2.6). Then I summarize Fodor's theory of the locking mechanism for acquiring symbols (2.7), and I argue that his theories of content and of the locking mechanism have contradictory implications: the former theory assumes semantic relations are metaphysically robust, while the latter theory implies that these relations aren't so robust (2.8). Finally, I consider an objection to some of my arguments in this chapter (2.9).

2.2 The Disjunction Problem and Semantic Robustness

Fodor (1990) proposes an informational theory of content that is supposed to account for how an object not in a symbol's extension can still cause the symbol token, which Fodor says is the mark of semantic robustness. Fodor refers to the problem of accounting for semantic robustness as "the disjunction problem." An informational theorist tries to explain the sort of symbol that has semantic content, which means at least the symbol that can be misused, with a cause that can be misidentified. A simple informational theory of content takes semantic content to be the information carried by a signal of a source. This informational relation is generally regarded as being nomically determined. Roughly speaking, information is what can be learned from a property of a signal about a property of its source in so far as the former indicates the latter. By its very presence, a signal that carries information eliminates possibilities about what the signal's

source could be, and so a property of the source can be learned from the signal. Information in this sense depends on nature's being an ordered place, and thus on natural law. The reason one thing rather than another is probably the source of some signal is that a property of the signal is nomically related to a property of its source.[2]

But if a symbol's content is identified with the type of object that nomically suffices for the symbol's tokening, the symbol can't be tokened in an erroneous way. Either there is only one way of nomically tokening the symbol type or there is more than one such way. If there is only one way, whatever causes the symbol's instantiation is what the symbol is about, and so there can be no reliable, law-like way, say, for foxes to cause the tokening of the symbol, "dog." The symbol is then used always correctly, because the symbol's intentional object instantiates the type that nomically suffices to instantiate the symbol type whenever the symbol is used. If, instead, there is more than one way nomically to instantiate the symbol, the informational theorist has to identify the symbol's content with one or the other source of the signal. If dogs or foxes under certain conditions are nomically related to the tokening of "dog," the symbol has disjunctive content, and so once again the symbol can't be erroneously instantiated. The disjunction problem illustrates the point that there can be no misinformation, in the technical sense of "information", according to which information is due to a nomic connection. So if semantic content is fundamentally information, there can be no misrepresentation, no

---

[2] Dretske (1981) argues for this link between information and natural law, but the link can be disputed. Information might be thought to require reliable correlation rather than nomic connection. The question then would be whether reliable correlation provides evidence justifying a judgment about one of the correlated things, so that a property of the one can be known from knowing a property of the other. Just because two things are usually found together doesn't mean there's a reason for the correlation, since the correlation can be accidental. A reason for judging one thing on the basis of what is known about something else would be knowledge of a nomic connection between the two, since a natural law could then be cited in a deduction of the judgment. In any case, I'll assume, with Fodor, Dretske, and Millikan, that the information that is supposed to be relevant to semantic relations depends on nomic connection.

false symbol tokens. Assuming symbols can misrepresent, this simple informational theory fails.

Teleological theorists attempt to solve the disjunction problem by pointing out that, assuming the laws governing the tokening of symbols are *ceteris paribus* (CP) rather than strict, or exceptionless, the instantiation of the relevant sort of nomic relation depends on conditions that aren't necessarily met. When the conditions are met, a symbol works like a signal and the semantic relation is no more than an informational one. But when the conditions are not met, the symbol's tokening could be nomically related to some type of object that isn't in the symbol's extension, and so the symbol could misidentify its cause. What makes this a teleological theory is that, in this case, the CP law stating the conditions needed for a symbol's tokening to be caused by the semantically relevant type, states also what function a certain mechanism is *supposed* to perform. Thus, the conditions needed for the symbol to be caused in this way are called Normal, or optimal, meaning that the conditions are needed for the mechanism to perform its function of implementing the denoted type's causing of the symbol's instantiation. The conditions under which a symbol can misrepresent its cause are abNormal, meaning that under these conditions the mechanism can more easily fail to perform its function. I'll focus on contrasting Fodor's theory of content with a teleological theory, as opposed to a pure informational one.[3]

[3] A naturalistic teleological theory of content posits an objective function subject to so-called descriptive rather than prescriptive norms, but I'll argue in Chapters 4 and 5 that there is no such function, because any norm that could generate a purposive function, as opposed to a disposition or a tendency, would be prescriptive. For this reason, when turning in the present chapter to a teleological theorist's possible replies to Fodor's theory of content, I'll often talk, more neutrally, about a mechanism's instantiation of a special causal relation or its working as needed by a CP law, rather than about a mechanism's fulfillment of its purposive function.

Fodor (1990) raises a number of objections against teleological theories, but he says that they share their main defect with informational theories. He points to this defect in his diagnosis of the source of the disjunction problem. While errors raise the disjunction problem, he says, the problem "isn't really, deep down, a problem about error." Instead, the problem is "the difference between *meaning* and *information*." The key difference between meaning, or semantic content, and information, is that "information follows etiology and meaning doesn't." By "etiology," Fodor means the causal history of a symbol's tokening, the nomic path to instantiating the symbol. Again, if there is more than one such path, the symbol carries more than one kind of information; each token symbol carries information only about its actual cause since there is no misinformation. By contrast—and Fodor emphasizes this point—"*the meaning of a symbol is one of the things that all of its tokens have in common, however they may happen to be caused.*" He goes on to say that "what's *really* wrong with teleological theories of content" is that their talk of Normal conditions "underestimates what one might call the *robustness* of meaning: In actual fact, 'cow' tokens get caused in all sorts of ways, and they all mean *cow* for all of that."[4] Not only can a symbol be caused by something semantically irrelevant to the symbol, as in cases of misidentifications of

[4] With regard to notation, I'll use quotation marks to indicate a reference to a symbol, and I'll switch to the use of capital letters to designate a mental representation such as a concept, when the context calls, in section 2.7, for the distinction between, say, concepts and words in natural language expressions. Fodor assumes that the primary bearers of content are syntactically primitive mental representations, which are the symbols whose content he tries to explain. So although quotation marks are often used to indicate reference specifically to words rather than to concepts, here they should be taken to indicate reference to symbols in general, to the primary symbols and to symbols whose content derives from the primary ones. Italics are often used in the literature to designate that to which a symbol refers. In the case of informational theories of content, such as Fodor's, the referent is the property of being an instance of the type that causes the instantiation of a symbol type, because the content is given by a nomic relation between properties. Fodor assumes, then, a certain view of natural laws. In any case, I'll speak of a symbol's reference to the objects in its extension, and of these objects as having the property that makes them instances of the type that is identified by the symbol. The symbol may identify these objects by picking out some of their other properties. For example, "dog" refers to dogs, and dogs are the objects that have the property *dog* and that may be picked out by their superficial, detectable properties.

what's perceived, but a symbol can be used correctly and still not be caused by the type of object the symbol denotes, such as when "dog" is caused by a doghouse or by something else associated with dogs. A symbol's meaning is "insensitive to variability in the causes of its tokening," and so a causal theory of content needs a way of "picking out *semantically relevant* causal relations from all the other kinds of causal relations that the tokens of a symbol can enter into." Teleological theories, however, are guilty of "implicitly denying robustness" by "idealizing to contexts of etiological homogeneity." On Fodor's view, false tokenings of a symbol "can happen whenever they like," and so a theory of content should be "compatible with any amount of heterogeneity in the causal history" of symbol tokens, that is, with the assumption that meaning is "arbitrarily robust" (90-92).[5]

I think there are two interpretations of Fodor's claims about semantic robustness, the empirical and the metaphysical, and I discuss each in turn. I'll argue that the metaphysical interpretation makes the most sense of Fodor's claim that robustness is the main stumbling block for teleological and causal theories of content, and thus of his statement of the goal for his own theory. However, on the metaphysical interpretation it's hard to see how the instantiation of symbols could be mechanically implemented, and so this interpretation conflicts with Fodor's account of a mechanism that carries out the acquisition of symbols. I go into the empirical interpretation in detail to do as much as I can to determine whether this interpretation is a viable alternative to the metaphysical one, since the latter, while posing the biggest problem for teleological and causal theories

[5] It should be noted that Fodor isn't just stipulating that semantic content is robust; instead, he's saying that there are empirical reasons to think semantic, as opposed to informational, relations are robust, and that any theory of semantic content should explain this robustness.

of content, also renders Fodor's overall representational theory of mind internally contradictory.

2.3 Two Interpretations of Semantic Robustness

To get an idea of the two interpretations of Fodor's talk of semantic robustness, consider how some papers on Fodor's theory of content can run together two claims. What happens, I think, is that a certain aspect of robustness is emphasized which is subject to different interpretations. This aspect is the *variety* of possible causes of a symbol type's instantiation. Fodor (1990) says that robust content is insensitive "to the heterogeneity of the (actual and possible) causes of its tokens" (90). By saying that symbol tokens can "get caused in all sorts of ways," Fodor is saying that symbols can be used properly and still not be caused by what they denote, so that the problem with some naturalistic theories isn't just the problem of explaining error. With this in mind, a symbol's *capacity* to keep its content, despite being caused by something other than its denoted object, might be discussed in terms of the wide variety of objects that can cause the symbol tokens. But this point supports two different interpretations, each of which can be found in the philosophical literature on Fodor's theory of content. The point about a symbol token's having a variety of causes might suggest what I call the empirical interpretation, since the wider the variety, the less the likelihood that conditions could be met under which there could be just one cause of the symbol token. Alternatively, the point about the variety might suggest what I call the metaphysical interpretation, that semantic relations are independent of causal ones.

Baker (1991), for example, speaks of the problem of semantic robustness in terms of tokens of a symbol type that "may have countless different kinds of causes...A thought of a cat may be produced, for instance, not only by an instantiation of a cat, but also by an instantiation of a shoe that you mistake for a cat" (17). Here, the capacity in question seems to be for a symbol type to have a wide variety of causes, but this point doesn't imply that the semantic relation between *cat* and "cat" is independent of the nomic connection between cats and "cat" tokens. Saying that there is such a variety is consistent with saying merely that the conditions for instantiating a special nomic relation can be imperfectly met. This, in turn, is consistent with a teleological theory of content, since this theory posits an abNormal situation in which there is just this sort of wide variety of ways in which a mechanism can fail, as it were, to implement a special causal relation, or as the teleological theorist says, can fail to fulfill its function. The special nomic relation here is a teleological relation between a mechanism, a set of special conditions, and the fulfillment of the mechanism's purpose. In any case, the metaphysical point about robustness goes beyond the empirical point, in saying not that a symbol token sometimes or often has a variety of causes, but that, to be a symbol, that which is caused *must always potentially* have such a variety of causes. There's no such potential in a Normal situation, and so a teleological theory seems inconsistent with the *metaphysical* interpretation of robustness.

Loewer and Rey (1991) seem to assume the empirical interpretation when they say that what meaning adds to information, namely robustness, is found in the fact that "*most* tokenings of sentences (whether in English or Mentalese) are produced in the *absence* of the conditions that they nevertheless mean." They contrast sentences with

symbols, by saying that symbol tokens are robust in so far as they mean "things that aren't on occasion their actual cause" (xxv). So most sentences are "wild," meaning that they are caused by things other than what the sentences correspond to, whereas symbols are wild only on occasion. In either case, the claim seems to be the empirical one about the extent to which a type of semantic relation actually amounts to a type of causal relation. Thus, they give the condition of robustness as follows: "tokenings of '*S*' are robust: i.e. are sometimes caused by instances of a property *G* other than *F*" (xxvii). Here, the point is that symbols *are* sometimes caused in a certain way, not that they *must* always have the potential to be caused in this way to count as symbols.

Seager (1993) says that the point about robustness is "the interesting point that a meaningful item *can* maintain a specific meaning although such items occur, or *are* produced, in a huge range of circumstances only some of which reveal the actual meaning of the item." He adds that the formal way of putting this point is to say that "Some non-*X*s cause *Y*s" (265, my emphases).[6] The formal point as well as the condition given above by Loewer and Rey seem to derive from Fodor (1990), where Fodor explains that if *A* causes *B*, and *B* causes *C*, *C* can't be a symbol with the robust content of *B*, since a condition of a symbol's having robust content is that there *are* tokens of the symbol type that aren't caused by their denoted objects. But those *C*s that are caused by *A*s are also caused by *B*s, since the three are related by a causal chain. It might seem that Fodor is saying here just that the tokening of a symbol type must sometimes actually be caused in a wild fashion, and thus that he's making the empirical point about robustness. But the point about causal chains is surely also that they are like Normal situations in making a wild tokening impossible. This latter point is the metaphysical one that a condition of

[6] By "*Y*s," Seager means the symbol tokens that refer to *X*s.

semantic robustness is that there *always* be the *possibility* of a wild tokening. To return to Seager's other claim, he seems to be working with both interpretations: his claim about what symbols *can* do is consistent with the metaphysical interpretation, although his claim that symbols *are* produced in a certain way suggests the empirical one. The metaphysical claim would be that, *however* symbols are actually produced, they *can* retain a semantic relation to some set of objects. This is different from affirming that symbols are actually produced in a certain way.

A questionable statement about semantic robustness is found in Livingston (1993). In speaking of the need for a theory of content to account for the "quality of robustness," he says that a "cow" token "had better mean *cow*, even if it is sometimes caused by horses. Teleological theories have exactly the wrong character if this is the goal, precisely because they attempt to specify Normal conditions in which 'cow' tokens have only the one kind of cause that fixes their meaning" (n.p.). Again, the point about a symbol being "sometimes" caused by some object the symbol doesn't denote suggests empirical robustness. Livingston takes this point to be inconsistent with a teleological theory of content that sets out Normal conditions under which there can be no such wild tokening of the symbol type. But were the point about robustness just the empirical one that symbol tokens are *sometimes*, or even *mostly*, caused by objects they don't denote, there would be no such inconsistency since this wild causing of symbol tokens would be free to happen under abNormal conditions. What's inconsistent with a teleological theory of content is the metaphysical interpretation of robustness, not the empirical one. On the metaphysical interpretation, a symbol can't exist under Normal conditions, since a symbol wouldn't have the *capacity* to be tokened in a wild fashion, and therefore

wouldn't be semantically robust, under those conditions. A teleological theory of content says, however, that a symbol can be found in a Normal situation, since content is determined by what happens in that situation.

Antony and Levine (1991) suggest both the empirical and the metaphysical interpretations when they speak of "the general phenomenon of *robustness*—the fact that representations, by their nature, can have an unlimited variety of causes, and still manage to mean something nonetheless." Again, the point about the variety of possible causes is consistent with the empirical view of robustness, but they link this potential variety to the *nature* of representations. They seem to take the potential for symbols to be caused by a wide variety of objects to derive from some potential of symbols themselves. Thus, they go on to say, "The robustness of genuine representations thus entails not only the possibility that they may on occasion, misrepresent, but that they display a certain 'detachedness' from their contents, that is, that they can occur without their contents being around to occasion them" (8). This seems to distinguish between the two interpretations. Here, the point isn't about what may actually happen on occasion, but about some property of detachedness that symbols possess.

2.4 Empirical Robustness

The empirical interpretation, then, is that conditions do not tend actually to be met under which the *only* way of causing symbol tokens is for the denoted objects to cause these tokens. What makes this an empirical claim is that there is no denial here that were certain conditions actually met, this might be relevant to explaining content; instead, the

point is that the meeting of such conditions is held out as in fact highly improbable. In other words, the point is granted that a Normal situation in which symbol tokens are caused is *possible*, but the claim is that as things *actually* stand, the Normal situation isn't realized. Usually, symbol tokens "get caused in all sorts of ways," because of the variety of prevailing conditions. No matter what the conditions that tend to be met in the actual world, such as the states of the environment or of the mechanism implementing the symbols, there is no nomic necessity that a type of symbol will be instantiated only as a result of the instantiation of the type of object the symbol denotes.

As it stands, this interpretation might seem to contradict Fodor's claim that symbol tokens are nevertheless, under some conditions, caused so that they carry information about their denoted type. To show why there might be a conflict here, I need to say more about the distinction between general, strict, and presumably physical nomic relations, on the one hand, and special, CP ones, on the other. Assuming, for example, that the laws of general relativity are perfectly general, in the sense of being fundamental, there are no special conditions that have to be met for them to be satisfied; on the contrary, what it means to say that these laws are general is that no matter what the specific conditions, such as the size of planets, their distance to each other, their composition, and so on, the laws apply to them.[7] Assuming there are also special natural laws, there are, by definition, special conditions that have to be met for these laws to apply. These conditions can be thought of as those that would obtain were the nomically

[7] This is a simplification, since the laws of relativity might not apply to quantum phenomena. Moreover, there may be no perfectly general laws, not even in physics, as Cartwright (1983), for example, argues. But I'll assume here that there is a needed distinction between perfectly general and special laws.

determined system isolated from other systems; when the system is not isolated, which is usually the case, there is interference from independent systems.[8]

Now, as I'll show when I present Fodor's theory of content, Fodor assumes that symbol tokens are caused by something they denote in a way expressed by a CP law. This is just the assumption that under certain conditions, a symbol token is at least a signal of something else. He assumes that semantics is part of psychology and that psychology is a special science, with laws that aren't reducible to physical laws. So if, say, "Dogs cause 'dog' tokens, *ceteris paribus*" is a law, there must be a special set of conditions for this law's application. Moreover, there must be some way to specify when all things are equal, as it were, or when all things are as they need to be to satisfy the psychological assumption that a symbol is used fundamentally to identify, and thus to be semantically related to, the distal type of object that causes its tokening.[9] Perhaps the conditions for a match between a semantic relation—a symbol's being about dogs—and a certain causal relation—the symbol token's being caused by a dog—are rarely or never fully met, in which case there is always some possibility of a mismatch under actual conditions: the symbol token that means *dog* may be caused by a fox. But the CP law implies that *were* these conditions met, only dogs would cause "dog" tokens. Moreover,

---

[8] As Pietroski and Rey (1995) point out, "the emergence of any theoretically interesting science requires considerable abstraction and idealization. The actual world is too complex to study all at once, so one proceeds by ignoring some aspects of the world in order to understand others. We idealize away from friction, electric charge, and nuclear forces, for example, when we seek to understand the effect of gravity on the motion of bodies. However, such abstraction guarantees a loss of descriptive adequacy in any generalization we lay down, since actual bodies are always affected by, e.g. friction, at least a little." They argue, then, that "cp-laws are a vehicle of such abstractions." That is, "a cp law holds only in a 'closed system', i.e. a system considered in abstraction from other, independently existing factors" (89).

[9] Again, as Pietroski and Rey (1995) argue, the "systematization" that a CP law assumes takes place in a closed system "is non-vacuous only to the extent that deviations from the regularities that are constitutive of it can be explained by those factors," that is, by the independently existing factors (89). If there is no specification of the conditions needed to close off a system, or of the independent factors that can interfere with the system when it is not closed off, there is little understanding of the system itself, and thus the CP law is empty.

assuming the CP law has some empirical evidence in its favour, the special nomic relation must actually be instantiated at some point. A natural law's being *ceteris paribus* means that there are special conditions of the law's application, and the law's being natural means that the law is justified by scientific procedures, and so there must be some empirical basis for asserting the CP law.[10] Thus, Fodor's own assumption that symbol tokens enter into special nomic relations with their denoted type seems to conflict with the empirical interpretation of semantic robustness. According to this interpretation, there is no *empirical* reason to assume there's a set of conditions under which a symbol token's being about a type correlates with the symbol token's being caused by this type.

However, strictly speaking, there's no contradiction here, since the empirical point about robustness is that there tend not to be conditions under which the *only* way of causing the tokening of a symbol type is for the denoted objects to cause this type. The empirical claim, then, is that, at best, there are likely conditions under which a type of symbol token is caused by either a semantically relevant type of object or by a semantically irrelevant one. If psychological laws are not strict laws, there must be special conditions for the application of these laws, and there must be some empirical basis for confidence that the special nomic relation obtains under the stated conditions.

---

[10] Fodor's response to Kripke's criticism of dispositional accounts of arithmetical rule-following suggests he might say that no such empirical basis is needed, since the CP law can refer to an idealized possible world. Kripke says a dispositional account of following the rule for addition fails, because either dispositions are finite whereas the rule has infinite applications, or the disposition corresponds to a CP law but there is no way to know what would happen were the relevant counterfactual statement true. For example, there would be no way of knowing what would happen were someone to have infinite memory to behave as predicted by the CP law and to follow the rule for addition even in a case involving very large numbers. Fodor says in response that even established CP laws, such as the ideal gas laws, refer to idealized scenarios. (See Fodor 1990 and Kripke 1982.) But a natural CP law must still have an empirical basis, in that there must be observed situations that approximate the unobserved, ideal one. Alternatively, the CP law can be derived from some law which is itself supported by empirical evidence. Otherwise, the CP generalization isn't a natural law, put forward by scientific methods, but an *a priori*, philosophical speculation. All of this is complicated by the question, to which I return in Chapter 5, of whether a scientific theory consists of laws or models, or both. In any case, I'll assume that if there are CP laws, there must be empirical evidence in favour of asserting them.

But given empirical semantic robustness, these special conditions must be such that either a denoted or a non-denoted object causes the symbol token under these conditions. For example, if a certain intensity of illumination in the ambient environment is needed for dogs to cause "dog" tokens in a type of creature, foxes may just as well cause these tokens given just that degree of illumination. So take all of the conditions of the special nomic relation's instantiation that have a realistic chance of being met; even were all of these conditions actually met, the symbol might be caused in some way other than that stated by the semantically relevant CP law. In other words, even in the most likely way of meeting the special conditions needed for the law's application, there will be only mixed evidence for the law. At least, this is what the claim of empirical robustness implies.

There are two ways of reading this implication, but I'll argue that neither is consistent with the empirical interpretation and with the claims that (1) a symbol type carries semantically relevant information, as dictated by a CP law, and that (2) empirical robustness is a deep problem for informational and teleological theories of content. The point could be that, empirically, a CP law is rarely, if ever, perfectly satisfied by real-world conditions so that there is no possibility of an exception to the law under those conditions. Alternatively, the point could be that, given the imperfect way actual conditions ever satisfy a CP law, there is always the possibility not just of an exception to the CP law, but of an overlap between this CP law and another one. These interpretations will lead to others, and I'll address them in order and in some detail to determine the merit of the empirical interpretation of semantic robustness.

On the former interpretation, the special conditions are only ideal in that their *full* realization in the actual world is highly improbable, and the reason the conditions aren't

fully met is because of an anomalous, circumstantial relation between some particulars. Thus, given the imperfect way in which actual conditions could ever satisfy the law that, *ceteris paribus*, dogs cause "dog" tokens, a fox as opposed to a dog can still cause the symbol token. Empirically, there is always the possibility of an exception to a CP law, even under part of the actual world's best approximation of the special conditions. In this case, when a symbol's being about a type of object doesn't match the symbol's being caused by this type, even under those ideal conditions that are most likely realized, this is because the symbol token can be anomalously rather than just nomically produced.

There is a problem, though, with this reading. Even if actual conditions are never perfect for a CP law's application, an approximation of these conditions must surely make some ways of tokening a symbol type less likely than others. A fox's causing of a "dog" token under conditions that approximate the special conditions for a dog's doing so would indeed be an *exception*. Therefore, a teleological theorist, for example, could claim that under Normal conditions, the nomic relation between the denoted type and the symbol type is the *only* nomic relation that is *probably* instantiated. This claim would be consistent with the empirical interpretation of semantic robustness. Indeed, Fodor's own way of stating the point about robustness, in terms of there being many ways of *causing* symbol tokens, seems to commit him to something like this teleological claim. Assuming there is, at best, a CP law that dogs cause "dog" tokens, as opposed to a law that, under the same set of conditions, dogs or foxes cause these tokens, the set of conditions must increase the probability that dogs cause the symbol tokens and decrease the probability that anything else causes them. Otherwise, there would be no empirical basis for the CP law at issue. Moreover, a random or an anomalous way of producing a symbol token is

precluded from being the semantically relevant way, since this way doesn't instantiate a nomic relation.

So while, as a matter of empirical fact, there might always be the possibility that a fox causes a "dog" token under conditions that are suited to a dog's doing so, this possibility must be less likely than the possibility of a dog causing the "dog" token under those conditions. Fodor himself must agree with this claim. But this claim is compatible with both a teleological theory of content and with the empirical interpretation of semantic robustness. The teleological theorist can identify the semantically relevant causal relation with the most probable causal relation under a realistic approximation of the optimal conditions, given a certain CP law. Again, were there no such causal relation that stands out under the *actual* conditions that are most suitable to instantiating a certain special nomic relation, there would be no empirical basis for the CP law in the first place. Thus, there would be no basis for the empirical interpretation of semantic robustness, according to which any type of symbol token is actually *caused* in many different ways. On the assumptions that a "dog" token is usually caused by a dog under optimal conditions, and that, in the anomalous case, the "dog" token might be caused by a fox under those same conditions, there is a principled way of distinguishing between the semantically relevant and the semantically irrelevant causal relations, while allowing for empirical semantic robustness. Thus, on this interpretation, the point about empirical robustness would hardly be a deep problem for a teleological theory of content.

Alternatively, the point about empirical robustness might be that under special conditions, the semantic relation is realized by multiple *nomic* relations. There are three ways this can happen, depending on whether the sets of conditions of the nomic relations'

instantiations have nothing in common, are identical, or are only similar to each other. I'll address each possibility in turn. Were the sets to have no conditions at all in common, the symbol's content wouldn't be empirically robust on the present interpretation. The point about empirical robustness is that instead of there being a single set of conditions under which the only way for a symbol type to be tokened is for the symbol tokens to be caused by the denoted objects, which would suit a teleological theory of content, there are actually always multiple ways of causing the symbol tokens. But were the symbols caused in different ways given only entirely different sets of conditions, the empirical evidence wouldn't generate even the appearance of a single set of conditions under which the symbol tokens are caused in multiple ways, conflicting with what a teleological theory predicts. And so there would be no deep problem here for a teleological theory of content. At most, there would be an epistemic problem of knowing which set of conditions and which nomic relation make for the semantic relation. But this wouldn't be a problem of empirical semantic robustness.

I'll go over this point again. The teleological theorist says that a semantic relation is determined by the only way of causing the tokening of a symbol type under a certain set of conditions. If there are two completely different ways of causing the symbol tokens, under two completely different sets of conditions, there is still one nomic relation per set of conditions, which is consistent with the teleological theory. The teleological theorist would need a reason to choose which nomic relation is semantically relevant. But the point is that the claim that there are multiple sets of special conditions for causing a type of symbol token doesn't conflict with the central teleological claim that semantic content is determined by what happens under *some* set of special conditions. Again, the

task is then to say which set of special conditions is semantically relevant, or Normal in the sense of being needed for a mechanism to implement a certain causal relation. This might be accomplished by giving an evolutionary explanation of the system that uses the symbol type. The reason this is an epistemic problem is that there may be insufficient access to the historical evidence that justifies talk of the Normality, or so-called functionality, of certain conditions under which the symbol was, and still can be, caused.

Now, suppose the two sets of conditions are *identical*—not just the special conditions that tend actually to be met, but all of the possible special conditions under which a symbol token is caused. For example, suppose there's a set of conditions under which either dogs or foxes might cause the tokening of a certain symbol type. In this case, the CP law would be that dogs or foxes cause certain symbol tokens under the same set of conditions. The challenge to a teleological theory of content, posed by empirical robustness, would be that this sort of theory is forced to regard a symbol's content as disjunctive even though the content may not be so. The response, though, would surely be that were indeed *all* of the special conditions under which dogs or foxes cause certain symbol tokens identical, this would show that the symbol user can't distinguish between dogs and foxes, so that the symbols in question aren't "dog" tokens. Fodor would need a reason to suppose that a creature could have the concept of dogs even though foxes are exactly as likely as dogs to cause the creature to token that concept under the very conditions that are nomically sufficient for dogs to do so.

Assuming, then, the empirical robustness of the content of "dog" doesn't require that the conditions under which dogs cause "dog" tokens are identical to the conditions under which foxes do so, Fodor can say that the two sets of conditions are only *similar* to

each other. The point about robustness, then, would be that usually dogs alone don't cause "dog" tokens, because the relevant conditions that tend to be met are similar to those under which, say, foxes cause these symbol tokens. This similarity between the two sets of conditions can take two forms, depending on whether one set of conditions is a subset of the other or whether the two sets overlap. Suppose one set is a *subset* of the other. For example, suppose the conditions that suffice for the causing of "dog" tokens by dogs are all among the conditions that suffice for the causing of these symbols by foxes. It might then be easier for dogs than for foxes to cause them, since additional conditions would have to be met for foxes to cause them. Moreover, foxes might cause these tokens only because dogs do so; that is, the one set of conditions might depend on the other set. In any case, the two sets of conditions would be asymmetrically related. But this would give the teleological theorist a way of distinguishing the semantically relevant special nomic relation from the irrelevant one. Thus, a teleological theory of content would be consistent with the claim that content is empirically robust, on this interpretation of empirical robustness.[11]

Now, suppose the two sets of conditions *overlap*. This would mean that the sets have positive rather than just negative differences: each set would have a member not included in the other set. The point about empirical robustness, then, would be that the conditions needed for one of the special nomic relations to be instantiated, as opposed to the other one, are rarely met so that the conditions that do tend to be met suffice for the

[11] As I'll go over in section 2.6, Fodor's own theory of content determinacy is that if there is more than one way of causing symbol tokens of a certain type, all but one of the ways, separately, are asymmetrically dependent on the one way. But his theory of content doesn't say the asymmetric dependency holds between sets of *conditions* for realizing the ways of causing the symbol tokens. It's this latter point that amounts to a teleological theory of content determinacy, that is, to the positing of a distinction between Normal and abNormal situations that does the work in determining semantic relations.

tokening of "dog" tokens by either dogs or by foxes. Still, for there to be two separate nomic relations here, there would have to be reason to speak of the other conditions which, when they are met, are needed only for the causing of "dog" tokens by dogs or only for the causing of these tokens by foxes. Once again, assuming there are two separate natural CP laws here, the laws must be justified by some empirical evidence, and so there would have to be an empirical basis for saying there's a set of conditions that suffices only for the causing of "dog" tokens by dogs. In that case, those conditions would differ from the conditions under which something not denoted by "dog" causes this type of symbol's instantiation, and there would be an empirical basis for speaking of the difference between the two sets of conditions.

But there are two ways in which this interpretation of empirical robustness might still pose a challenge to a teleological theory of content. First, were the conditions *shared by* the sets the only conditions in the two sets that tend actually to be met, the teleological theorist would have no empirical reason to explain the semantic relation in terms of only one of the nomic relations. This is because those of the special conditions that tend to be met also would be those under which another nomic relation is instantiated. But this proves too much, since there also would be no empirical reason to speak of two different nomic relations in the first place; instead, once again, the evidence would support the generalization that under certain observed conditions, dogs or foxes causes "dog" tokens. By hypothesis, however, there are two separate nomic relations whose instantiations depend on different sets of conditions.

The second challenge posed by the claim that there are multiple ways of causing a type of symbol token, whose special conditions overlap, is as follows. Assuming the

conditions *not* shared by the two sets are readily met, there would be an empirical basis for speaking of two nomic paths to the tokening, say, of "dog," but the teleological theorist would still need a reason for taking one rather than the other nomic relation to be the semantic relation. There would be two sets of conditions for instantiating "dog" tokens by two separate nomic paths, as it were. The question then would be why, if a semantic relation is, roughly, just some special nomic relation, the type of symbol has dogs rather than foxes in its extension. Unlike the case in which one of the sets is a subset of the other, in the present case in which the sets only overlap, there's no obvious reason why only one of the two nomic relations is semantically relevant. As in the case in which the two sets of special conditions are completely different, the teleological theorist has here an epistemic problem rather than a problem just of empirical robustness. Of course, the teleological theorist could say that although there would be different ways of causing "dog" tokens and no obvious asymmetry between these ways, only one of the ways would be Normal because only one would follow from a teleological, functional explanation of the existence of the later symbol tokens. For example, an evolutionary explanation might posit one rather than the other nomic relation, and so the constraints on this explanation would provide a reason for taking only one of the nomic relations to be semantically relevant.

But the following, more comprehensive reply might be in order. (1) If the conditions the two sets don't share aren't in conflict with each other, and if these different conditions are equally well met in the actual world, there might be no basis for speaking of two sets of conditions and of two separate nomic relations. Instead, the symbol might be caused by either dogs or foxes under some single, broad set of conditions, because the

symbol has disjunctive content. This raises the same issues discussed above about the empirical basis of justification for natural CP laws.

(2) Were there reason to speak of two sets of conditions because some of the conditions aren't as often met as are others, and thus, say, dogs more frequently cause the symbol tokens than do foxes, there would also be some asymmetry between the sets of conditions which might be explanatorily relevant. For example, a creature's internal mechanism might be adapted to detecting one type of animal rather than another, because one rather than the other type is a fixture in the creature's environment, and the creature's ancestors increased their fitness by having a way to detect the often-encountered type of animal. Thus, once again, the teleological theorist might have a principled way of explaining the semantic relation in terms of only one of the nomic relations.

(3) Suppose some of the conditions are in conflict with each other, so that certainly there are two sets of conditions under which "dog" tokens are caused by dogs and by foxes, respectively. Suppose also that foxes are just as likely as dogs to cause the symbol tokens; that is, the one set of conditions is just as likely to be met as is the other set, although not at the same time when some of the conditions in conflict with each other are met. For example, suppose dogs cause "dog" tokens when dogs appear to walk in a certain way, whereas foxes cause the tokens when foxes appear to walk in a different way, and no animal can appear to walk in both ways at the same time. But either kind of animal, with its own gait, is just as likely to cause "dog" tokens, given the meeting of other conditions shared by the two sets.

This scenario poses more of a problem for a teleological theory of content. Suppose there is no decisive information available about the evolutionary histories of the

two nomic relations that might reveal an asymmetry between the conditions of their instantiation. In this case, whether one or the other of the supposed nomic relations is instantiated would appear to be random, leaving the teleological theorist with no reason to take one rather than the other to be the semantic relation. Were the conditions of the nomic relations' instantiation asymmetrically related, as supposed above, there would be no such problem. On (3), however, the sets of conditions might be symmetrically related. There would simply be two independent ways of causing some symbol tokens, and yet, by hypothesis, one of the nomic relations wouldn't be the semantic relation. The teleological theorist seems to lack the resources, in this case, to explain the symbol's content. Without an appeal to some actual etiological asymmetry, I don't see how a teleological theorist can answer this epistemic question.

However, with regard to the issue of empirical robustness itself, the teleological theorist can respond in the above ways. If the conditions that the two sets don't have in common are typically met along with the shared conditions, there won't be the appearance of a single set of conditions, and thus the symbol tokens won't be empirically robust in a way that challenges a teleological theory of content. There will be various ways of causing a symbol token but different sets of conditions under which the token is caused. Some of these sets of conditions can be irrelevant to the special causal relation, or so-called function, that determines content, on a teleological account. If, instead, the conditions not shared by the two sets are not typically met along with the shared conditions, there will indeed appear to be a single set of conditions, and yet for this very reason, as discussed above, there won't be an empirical basis for positing the other conditions and thus for positing two different nomic relations. In this case, the symbol

tokens won't be robust in a way that matters to Fodor, because there won't be multiple nomic relations under a single set of empirically met conditions. For empirical robustness to challenge a teleological theory of content, there has to be the appearance of a single set of conditions and yet different ways of causing a symbol type under those conditions. After all, the teleological theorist says content is determined by the only nomic relation instantiated when certain conditions are met, namely those needed for the performance of a mechanism's function. But these criteria for empirical robustness are in conflict. When there appears to be a single set of conditions, this is an empirical reason to posit only a single, in this case disjunctive, special nomic relation, and this in turn is reason to say the symbol has disjunctive content.

So much for my analysis of the empirical interpretation of Fodor's claims about semantic robustness. I don't claim to have shown here that the empirical considerations are no threat at all to a teleological theory of content. What I do think I've shown is that empirical robustness doesn't pose a *deep* problem for this sort of theory, since on various interpretations of the point about empirical robustness, the teleological theorist has ready responses. Whatever semantic robustness comes to, though, Fodor says robustness goes to the heart of what's wrong with a teleological theory. Given this reason for speaking about semantic robustness, a stronger interpretation of the robustness claim is needed and so I turn to the metaphysical interpretation.

2.5 Metaphysical Robustness

Much of what Fodor says about robustness calls for a stronger interpretation. Recall Fodor's words, quoted above (see section 2.2). He says that "information follows etiology and meaning doesn't." The tokens of a symbol type have their content "*however they may happen to be caused.*" False tokenings of a symbol, for example, "can happen whenever they like," and symbols are "arbitrarily robust" in that there is "any amount of heterogeneity" in their causal history. Here, the point seems to be, not just that there aren't actually special conditions under which a semantic relation amounts to a causal one; rather, the point is that even were there a set of conditions under which there's only one nomic path to the tokening of some type, semantic content couldn't be explained in terms of such a nomic relation, because what is caused in such a case wouldn't be a symbol, or something with semantic content as opposed to information. There are indefinitely many kinds of causes of a symbol token, making the symbol's semantic content arbitrarily robust, because this content isn't determined by what causes the symbol token.

The stronger, metaphysical interpretation, then, is that symbol tokens keep their semantic content no matter how they are caused; their content isn't determined by any causal relation, no matter how optimal the conditions for instantiating this relation. Were there conditions under which tokens of some type are caused necessarily in only one way, these wouldn't be *symbol* tokens, with semantic content, because they would lack the *capacity* to be caused in either semantically relevant or irrelevant ways under those conditions. Symbols with metaphysical semantic robustness can *always*, under any set of

conditions, be caused in more than one way. Thus, a "dog" token, for example, must always be able to misrepresent its cause, even when conditions are actually such that a dog is the most likely cause of the symbol token. Dogs rather than anything else may actually cause "dog" tokens under certain conditions, but the symbol's content is so robust that it stands on its own, independent of how the symbol token is caused. The robustness gives the symbol the capacity to keep its own content regardless of the probability of a certain semantically relevant or irrelevant causal relation's instantiation under special conditions; a metaphysically robust semantic relation is simply *independent* of what probably happens when conditions come together for a mechanism to implement a certain nomic relation.

I call this second interpretation of robustness "metaphysical", because on this interpretation a higher-order theory of content is needed. Fodor himself offers just such a theory, which is yet another reason to regard the empirical interpretation of robustness as flawed. As Fodor (1990) says, "Robustness captures the point that some ways of using symbols are *ontologically* parasitic on others" (128, my emphasis). In saying this, Fodor is committed to saying that content isn't determined by natural laws, including physical and CP laws. Were a symbol's content determined by a physical law, the symbol tokens wouldn't be free to be caused in more than one way in so far as the physical law applies to them. On the contrary, the way the symbol tokens are caused would correspond just to what the law says. Now, a CP law allows for at least two ways of causing something, since this sort of law is such that the special conditions needed for the nomic relation's instantiation may or may not be met. But when the conditions *are* met, a CP law implies that there is only one way of causing certain tokens, namely the way that corresponds to

what the law says would happen under the special conditions.[12] In general, then, whatever is determined by a natural law happens in a way that corresponds just to the law. Assuming physical laws are fundamental, Fodor's point about semantic robustness implies, on this deeper interpretation, that symbols require a *meta*physical explanation, that is, an explanation given in terms that are deeper than, or prior to, the terms given by natural (general or special) laws.

If semantic content is robust in Fodor's sense, and the best interpretation of what Fodor means by "semantic robustness" is what I'm calling the metaphysical one, this is surprising because, as I'll show later in this chapter, metaphysical robustness poses a problem for any naturalistic theory of mind, including Fodor's, that takes a reference relation to be implemented by a mechanism, that is, by some complex causal system. No mechanically implemented symbol token has a natural capacity to be caused in multiple ways, at the higher, semantic level of explanation, when conditions are optimal for a mechanism to work, satisfying some CP law. So as soon as such a mechanism is posited, the metaphysical robustness claim is denied. This is so even in the case of the mechanism Fodor himself posits to account for how symbols are acquired since, as I'll argue in sections 2.7 to 2.9, this mechanism, or complex causal relation, would, in effect, determine the semantic relation. Thus, a teleological theory of content, which denies the metaphysical robustness claim, can be read off of Fodor's account of the mechanism for acquiring symbols. To try to save Fodor's account from resting on a notion of semantic content that doesn't fit with his claim that symbols are mechanically implemented, I

[12] Assuming the special nomic relation is multiply realized, there may be more than one mechanism that realizes the nomic relation in a Normal situation, or else a different Normal situation for each mechanism. Either way, there is only one way of causing a certain effect in a Normal way of producing the effect, at the higher level of explanation.

analyzed the alternative, empirical interpretation in some detail, to see whether this alternative is viable. But Fodor's claim, that semantic robustness poses a deep problem for teleological theories of content, implies that the robustness is metaphysical, not just empirical. Empirical robustness poses no such problem, since the teleological theorist's distinction, between conditions that are and that aren't conducive to a mechanism's ability to implement a special causal relation, easily accounts for whether the causal relation that realizes a semantic relation is or isn't actually instantiated. Indeed, on Fodor's own assumption that a symbol type is caused, *ceteris paribus*, by its denoted type, there are special conditions that have to be met for this causal relation's instantiation. The teleological theorist can take these conditions as the Normal ones that determine a symbol type's content.

Fodor has to say, then, that what is left out of a teleological theory is an explanation of the independence of semantic and of special causal relations. Metaphysical semantic robustness is a real problem for a teleological theory of content, and for any theory that takes content to be nomically determined, because the point about this robustness is precisely that content is not so determined. Something other than a mechanism, a set of special conditions, or a causal relation must be the determinant of semantic content, if this content is metaphysically robust. Even were the conditions that are suitable to instantiating a nomic relation met, a symbol with metaphysically robust content wouldn't be explainable in terms of these conditions, because such a symbol would have the capacity to keep its content regardless of whether these conditions are met. But this capacity would be like a capacity of someone to walk on a lake's surface despite the meeting of conditions under which the person is naturally compelled to sink.

The only thing that has the capacity to violate natural laws is a creature, where the laws are prescriptive, not descriptive or scientifically discovered. Symbols with metaphysically robust content, then, are determined, not by a mechanism, but by prescriptive norms when they are used by creatures subject to these norms. Although the neural particulars that realize mental representations are subject to mechanistic interactions, there must be a level of explanation according to which symbols are subject to prescriptive norms rather than to scientific generalizations about mechanisms. While a neural particular can't be caused in multiple ways when conditions are met for a neurological mechanism to satisfy some CP law, a creature that uses the symbol and that can fail to follow a prescriptive norm isn't so restricted. The problem for Fodor is that he thinks content is determined at the level of symbol types, not at the level of the symbol-user, and that he posits a mechanism for acquiring symbols that, in effect, determines content. No such symbols can have metaphysically robust content. At least, this is what I'll argue later in this chapter.

2.6 Fodor's Theory of Content Determinacy

I turn now to a brief presentation of Fodor's theory of content. The theory, as given in Fodor (1990), is meant to apply directly to syntactically simple, lexical mental representations, such as the concepts DOG or HORSE. These are taken to have primitive, original content, while the content of a complex expression, such as the thought of large dogs, is assumed to derive from the contents of the simple parts of the expression and from the rules used to combine them. Moreover, given the representational theory of

mind, the content of linguistic expressions is assumed to derive from that of mental representations.[13] Intuitively, a mental representation is a mental particular that stands in for something in the environment even in the absence of this other thing. Fodor takes the job of a naturalistic theory of content to be, minimally, the giving of sufficient natural conditions of this semantic relation of representation.

He begins with the assumption that a symbol is fundamentally a signal that carries information. This is usually taken to mean that a symbol is nomically related to the set of objects denoted by the symbol. Therefore, as Fodor says, "*X*s cause '*X*'s" is a natural law (121). There is an external relation, then, between a symbol in the mind and the denoted object, which is a causal, nomic relation.[14] This means, for Fodor, not just that there is a correlation between, say, particular dogs and some token symbols, but that there is a necessary connection between the two types. Subjunctive conditionals about what would happen, given the law's antecedent, can be inferred from the law, which is to say that the law is a generalization covering not just the actual world but possible states of the world. Fodor assumes, though, that this connection between types, or certain properties, is a special nomic relation that depends on conditions that may not actually be met. The law at issue, then, is a CP law.[15]

---

[13] Again, I'll speak loosely, though, of symbols, using quotation marks instead of capital letters, and switch to capital letters only when the context warrants the distinction between words and concepts.

[14] Technically, as Fodor (1990) says, the sort of nomic relation that is relevant to his theory of content holds between "the property in virtue of which *X*s [or *Y*s] cause '*X*'s and the property of being a cause of '*X*'s" (102, my italics). For example, the property *dog* is the distal property in virtue of which some set of objects causes "dog" tokens, and *dog* is nomically related to the property of being a cause of those tokens. In other words, dogs are generally the sort of things that cause "dog" tokens.

[15] Fodor (1990) qualifies his statements a number of times by saying that the laws relevant to his theory are CP. For example, he says that the asymmetric dependence condition, which distinguishes his theory of content and which I'll come to in a moment, "requires that, ceteris paribus," possible worlds in which the relevant laws hold be related in a certain way (108).

But semantic content is robust and therefore can't be just information. So Fodor considers the possibility that tokens of a symbol type can be caused by objects not denoted by the symbol. Indeed, symbols are often caused by such semantically irrelevant objects, which is to say that the conditions that would have to be specified in the above CP law aren't usually met. Thus, "non-*X*s cause '*X*'s" is also a CP law. This leaves the question of why *X*s rather than non-*X*s should be the content of the symbol type, or of why only one of the nomic relations is semantically relevant. Fodor's answer is that there is a higher-level, asymmetric relation between the nomic relations that favours one of the nomic relations. Specifically, he says, "For all *Y* not=*X*, if *Y*s qua *Y*s actually cause '*X*'s, then *Y*s causing '*X*'s is asymmetrically dependent on *X*s causing '*X*'s" (121). For example, if "dog" is caused by either dogs or by foxes, there would be no nomic connection between foxes and the symbol type were there no such connection between dogs and the symbol type, but the converse statement is false: there would be a nomic connection between dogs and the symbol type even were there no nomic connection between foxes and the symbol type. This asymmetric dependence condition of semantic content can be put in terms of possible worlds. If a symbol type denotes the property *dog* rather than *fox*, even though the symbol tokens are sometimes caused by foxes, the possible world in which dogs but not foxes cause the symbol tokens must be closer to our world than any possible world in which foxes but not dogs cause these tokens. But for the fact that dogs cause these symbols, foxes wouldn't do so, but dogs could cause them even were foxes incapable of doing so. The nomic relation between foxes and "dog" tokens is, as it were, parasitic on that between dogs and "dog" tokens.[16]

[16] Fodor (1987) says his theory of content rests on a Platonic intuition: "It's an old observation—as old as Plato, I suppose—that falsehoods are *ontologically dependent* on truths in a way that truths are not

This asymmetry gives Fodor a reason to call one of the nomic relations the semantically relevant one for some symbol type. Any symbol type's semantic relation to a denoted type is an *independent* nomic relation relative to any dependent, parasitic nomic relation. Whereas informational content is determined by a nomic relation, semantic content is determined by an asymmetric relation between nomic relations. As I said, Fodor takes nomic relations to hold between types or properties, not between tokens. Indeed, he assumes that the nomically related properties needn't be instantiated, and thus that they are ontologically distinct from particulars.[17] Moreover, he assumes that properties and their relations are ontologically prior to particulars and their relations.[18] Thus, his theory of content allows for a symbol type to have content even though the creatures that use the symbol never encounter the denoted type of object. As long as foxes, say, *would* causes a certain symbol token in a type of creature only because dogs would, and not the other way around, the symbol denotes *dog*, even if there were actually only foxes and no dogs. Were this creature to use this symbol to refer to foxes, which are animals the creature does come into contact with, this would be a case of misrepresentation. The symbol would carry the informational content *fox*, but the semantic content *dog*. What matters in determining semantic content isn't how tokens are actually related given the prevailing conditions for satisfying CP laws, but how the nomic relations themselves are related. So, according to this theory of content, a creature that

---

ontologically dependent on falsehoods. The mechanisms that deliver falsehoods are somehow *parasitic on* the ones that deliver truths" (107).

[17] In speaking about the content of "unicorn", Fodor (1990) says "I take it that there can be nomic relations among properties that aren't instantiated; so it can be true that the property of being a unicorn is nomically linked with the property of being a cause of 'unicorn's *even if there aren't any unicorns*" (100-101).

[18] Fodor (1990) says that "*Ontologically* speaking, I'm inclined to believe that it's bedrock that the world contains properties and their nomic relations, i.e., that truths about nomic relations among properties are deeper than—and hence are not to be analyzed in terms of—counterfactual truths about individuals" (93).

never actually perceives a dog, but does encounter foxes, can still have a mental particular that refers to dogs rather than to foxes.

On Fodor's view, then, semantic relations aren't established in the domain governed, as it were, by physical or by CP laws, that is, in the natural domain, properly speaking; rather, these relations are established metaphysically prior to anything that happens in the nomically determined domain of particulars. The point isn't just that semantic relations have a top-down rather than a bottom-up, or evolutionary, explanation, since on his view all nomically determined events depend on relations between properties which are ontologically distinct from particulars. In explaining semantic content, Fodor posits not just relations between properties, but higher-level relations between nomic relations. A semantic relation requires not just a nomic relation, but the independence of such a relation from certain other possible nomic relations. A nomic relation's independence from another one isn't established by any nomic relation or by any nomically determined, that is, natural process or actual evolution of particulars. Of course, were the asymmetric dependency itself somehow determined by a natural law, such as by a physical law, semantic content wouldn't be metaphysically robust. This is because the semantic relation wouldn't be independent of a certain nomic relation.

To recap, then, Fodor claims that the robustness of semantic content is a deep problem for naturalistic, and specifically for teleological theories of content. There seem to be two interpretations of the claim about robustness, but only one of these interpretations, the metaphysical one, seems to support Fodor's claim that semantic robustness can't be explained by a teleological theory. Metaphysically, a symbol's content is robust if a symbol keeps its content no matter what actually happens to

particulars in a nomically determined domain. This feature of symbols calls for a metaphysical theory of content, such as Fodor's which assumes that the asymmetric dependency between nomic relations is fixed prior to anything determined by nomic relations themselves. On Fodor's view, there is *always* the *metaphysical* possibility for a symbol token to be caused by something not denoted by the symbol, which means that the symbol's content isn't determined by any one nomic relation, such as one instantiated under Normal conditions, even assuming there were such conditions. Instead, a symbol's content is determined by the way certain nomic relations are structured, and this abstract structure can be in place regardless of how nomic relations are instantiated, and thus regardless of how a symbol token might actually be caused. This Platonic aspect of Fodor's asymmetric dependency theory does the same work that the purposive function does in teleological theories, such as Dretske's and Millikan's, except that Fodor doesn't say that talk of the asymmetric dependence of some relations between ontologically distinct and prior properties has any normative implications.

Fodor (1990) also considers a version of his theory of content, in response to what he takes to be a questionable verificationistic implication of the theory. Again, his basic theory has two parts. First, there is the informational, nomological condition that if "*X*" means *X*, "*X*s cause '*X*' tokens" is a CP law. Second, there is the condition that if "*Y*s cause '*X*'s" is also a CP law, the nomic relation corresponding to this law is asymmetrically dependent on the nomic relation corresponding to the law given in the first condition. Information depends on nomic relations, and a nomic relation is given by a law that applies to possible as well as to actual states of the world. If there is a nomic as well as an accidental way to produce a symbol token, only the nomic way is semantically

relevant. And symbol types are distinguished, on informational grounds, only if there are multiple laws at issue and therefore different predictions about possible ways in which the symbol types are instantiated. If "dog" and "fox" have different contents for a type of creature, the creature must have the capacity to distinguish between dogs and foxes, because there must be a possible world in which one rather than the other nomic relation is instantiated. Conversely, if a creature can't distinguish between dogs and foxes, even though there are differences between these two types, the creature's symbol for dogs refers also to foxes. This means that content depends on the symbol user's capacity to distinguish the denoted type from some other type, not directly on real similarities or differences between denoted tokens.

Fodor is suspicious of the verificationistic claim that having a symbol type that means *X* depends on having the capacity to distinguish between *X*s and *Y*s. If there is a real difference between *X*s and *Y*s, and between *X*s and the disjunction of types, (*X*s or *Y*s), Fodor (1990) asks, "why shouldn't we be able to talk (/think), in ways that respect" those differences? That is, why should content depend on the mere capacity to *recognize* real similarities or differences in the world, and not simply on these real similarities and differences? Therefore, he adds a third condition to his theory: if "*X*" means *X*, "*X*" tokens must actually be caused by *X*s. When this condition is met, contents can be distinguished not only by a difference between nomic relations, but by a difference between actual conditions in which the symbol types are used. The objects denoted by a symbol must be among the actual causes of the symbol's tokening; that is, if "*X*" means *X*s, "Some '*X*'s are actually caused by *X*s" (121). Call this third condition the actual

history condition (AHC).[19] On this version of the theory, if a creature has a symbol meaning *dog*, dogs must have actually caused some of the symbol tokens. AHC provides a more objective way of distinguishing between symbols, since if foxes don't actually cause "dog" tokens, because conditions aren't actually met, perhaps, for instantiating the nomic relation between foxes and "dog" tokens, foxes can be excluded from the extension of "dog" without having to appeal to whether the symbol-user could distinguish between dogs and foxes in some possible world. When multiple types of objects, such as dogs or foxes, *actually* cause the tokening of a symbol type, or when each nomic relation is instantiated, the version of the theory that includes AHC says that either the symbol has disjunctive content or the symbol user has the capacity to distinguish between the two types. When the symbol user can distinguish between these two types, the asymmetric dependence condition applies, privileging one of the ways of actually causing the symbol tokens.

Fodor points out that AHC "violates the assumptions of pure informational theories," since this version of the theory assumes that semantic relations may be fundamentally contingent, historical relations rather than nomic ones (121). It's clear, however, that AHC is also inconsistent with the metaphysical interpretation of semantic robustness. Recall that on this interpretation, a symbol's content is fixed regardless of how the symbol tokens are actually caused and regardless of whether any special set of conditions is met for instantiating the symbol type (see section 2.5). On the empirical interpretation, the point about robustness is that while a semantic relation may amount to

[19] Fodor (1990) points out that, given AHC (presumably), semantic content isn't "arbitrarily robust" (132, n.5). I'll argue in this section that AHC is also incompatible with content's being metaphysically robust. In fact, talk of *arbitrary* robustness seems a way of getting at the distinction between empirical and metaphysical robustness.

a causal relation instantiated under some special conditions, these conditions tend not to be met in the actual world, which is why a symbol is actually instantiated in a variety of ways. On the metaphysical interpretation, though, the point is that even were such conditions met, that which would be tokened under these conditions wouldn't be a symbol, with semantic content, because a symbol token has the capacity always, and thus no matter which conditions are actually or possibly met, to denote something other than what causes the symbol's instantiation. What allows for this independence, in Fodor's theory, is the metaphysical priority of the higher-level asymmetric dependency.[20] A semantic relation is independent of any causal relation, including any that tends to be instantiated when some conditions are actually met. But AHC says that a symbol must denote something that actually causes the symbol's instantiation. Actual ways of causing symbol tokens are of no consequence to the determinant of content, given metaphysical robustness, whereas AHC takes one such actual way, for each symbol type, to be a condition of the symbol's having determinate content.[21]

I assume Fodor needs the metaphysical interpretation of semantic robustness to motivate his alternative to causal and to teleological theories, and since AHC is inconsistent with this interpretation, I set aside AHC. Next, I want to determine whether the mechanistic part of Fodor's broader theory of mind is likewise inconsistent with the

---

[20] To anticipate, I would point out that, likewise, assuming prescriptive norms aren't reducible to objective, descriptive states of affairs, and purposive functions are prescriptively normative, a teleological theory could account for this metaphysical robustness. As I'll argue in the next two chapters, though, the hidden prescriptive normativity of the functions posited by Dretske and by Millikan only makes trouble for their theories of content.

[21] Fodor (1990) raises another problem for his theory of content when AHC is added to the theory: Fodor is forced to say that a symbol such as "unicorn", which isn't actually caused by its denoted type, is syntactically complex rather than simple, so that this sort of symbol doesn't fall within the purview of Fodor's theory (124).

simpler version of his theory of content, that is, with the asymmetric dependence theory which I've argued needs to assume that content is metaphysically robust.

2.7 The Locking Mechanism of Concept Acquisition

I've argued that Fodor's main criticism of informational and teleological theories of content is that these theories don't explain a symbol's immunity, as it were, to the variety of ways of causing the symbol tokens. Fodor contends that the deeper problem revealed by the fact that symbols are used in misidentifications of their cause is the problem of semantic robustness. But there is a question of how to interpret this deeper problem. I've tried to show that the interpretation most compatible with Fodor's statements and with his simpler theory of content is that content is metaphysically rather than just empirically robust. The semantic content of symbol tokens, as such, is determined not by a nomic relation, but by the metaphysical fact of an asymmetric dependence between nomic relations.

In this section, I summarize his theory of concept acquisition, and in the next I argue that this theory contradicts his theory of content.[22] On the one hand, he says content is metaphysically determined by the asymmetric dependency. On the other, he says symbols enter into special nomic relations that must be mechanically implemented, where

[22] Technically, if I'm right about the contradiction, the objection is to his broader theory of mind which encompasses his theories of content and of concept acquisition, not just to his theory of content, since his account of concept acquisition isn't implied by his account of content. However, the contradiction is derivable solely from implications of his asymmetric dependence theory. This latter theory implies that symbols enter into special causal relations which, on Fodor's view, must be implemented by mechanisms. My objection is that such symbols can't have metaphysically robust content, contrary to Fodor's theory of content, because the properties of symbols that work in naturally selected mechanisms are teleologically determined. Fodor's theory of the mechanism for acquiring concepts just provides for a striking way of deriving the contradiction.

a mechanism is a complex system that works only when sufficiently isolated from other systems. He posits a mechanism to explain how mental symbols are acquired, but in so doing he makes semantic relations dependent on causal ones, contrary to the assumption of metaphysical semantic robustness.

With regard to concept acquisition, Fodor (1998) defends the earlier view, in Fodor (1975), that many syntactically simple, lexical concepts can't be learned by abduction, since this type of learning would require forming a hypothesis that already employs the concepts to be learned. This suggests that these concepts must be innate. In the later work, though, he argues that these concepts themselves aren't innate; instead, what is innate is the capacity to lock to kinds denoted by them, from limited experience of certain particulars. When a simple concept is acquired, an internal structure is given a semantic relation to some type, by configuring the internal structure so that the concept comes to be tokened for the first time in, and thus made available to, a creature. Fodor (1998) calls that which configures the internal structure "the mechanism of concept acquisition," and thus he thinks of concept acquisition as a nomically determined process (128).[23]

The simple lexical concepts in question, that is, the ones that are in some sense innate, are the concepts that are acquired by means of a creature's innate mechanisms for recognizing things as instances of a certain type. There are, broadly speaking, two kinds of concepts that can be acquired in this way. First, there are concepts of subjective, mind-

[23] He says, for example, that explaining why denoted objects cause the acquisition of the concept, or explaining what is called the doorknob/DOORKNOB effect, to which I'll turn in a moment, "requires postulating some (contingent, psychological) mechanism that reliably leads from having *F*-experiences to acquiring the concept of *being F*" (133). Mere correlation or historical connection between the objects to be denoted by an acquired concept and the creature that acquires the concept doesn't explain why a *concept* is acquired, so there must be some mechanism that connects the denoted objects and the creature when the concept is acquired.

dependent kinds, such as DOORKNOB.[24] When we perceive objects with certain phenomenological properties, they strike us as being doorknobs, and there is nothing more to being a doorknob than for these objects to strike us in a certain way. Being a doorknob is nothing more than having certain features that certain creatures can understand. Second, there are concepts of objective, mind-independent kinds that are acquired by innate mechanisms rather than by learned, scientific theories. The concept of water, for example, is of a natural kind that is more than just the interaction between certain stuff and some creature's way of understanding the stuff. Children, animals, and people who lived prior to the practice of modern science acquire the concept of the natural kind *water*, but not of water *as* this kind, because they lack the scientific theory of water's real properties, which are those that underlie, as it were, its superficial, phenomenological ones. In either case, there is an innate mechanism for acquiring a concept by locking to a mind-dependent or to a mind-independent type, by means of a causal relation between the type, given by its detectable properties, and creatures with the locking mechanism. The detectable properties are the ones that trigger the locking mechanism, causing a type of creature to conceive of some set of objects in a way that differs from conceiving of them as having simply the detectable properties. Whatever these detectable properties are, whether they have to do with the object's shape, motion, or some other property, the property to which the creature locks isn't among them.[25]

---

[24] This assumes that DOORKNOB is syntactically simple which, technically speaking, it's not. Fodor (1998) raises this point about the inappropriateness of this example in the literature, but assumes for the sake of argument that DOORKNOB is primitive (122, n.3).

[25] In this way, the set of objects that causes a creature to lock to a type denoted by a mental representation of these objects can be independently specified. Thus, reference to the detectable properties can be used in an explanation of why, for example, perceiving a doorknob, or an object with the property of being a doorknob, causes the acquiring of DOORKNOB. Were there no such independent specification, Fodor's point might be simply the platitudinous one that DOORKNOB is acquired from experience of certain objects in virtue of their doorknobhood. As Fodor (1998) says, this point would be simply a restatement of

Moreover, with regard to either DOORKNOB or WATER, the innateness of the locking mechanism accounts for why, for example, doorknobs cause the acquisition of DOORKNOB or why water causes the acquisition of WATER.[26] In the case of a concept with mind-dependent content, the match is hardly mysterious, because there is a necessary connection between the causal and the semantic relations. There is nothing more to being the *semantically relevant* objects, in this case, than to have certain features that *cause* a type of creature to token a certain concept, and thus that cause the object to appear to the creature in a certain way. For example, assuming the property of being a doorknob depends on some interaction between the object and the perceiver, our acquiring of DOORKNOB reflects our being causally struck in a certain way by the object. The purest case of an object's causing of a symbol token, the content of which reflects solely how the object appears *innately* to the perceiver, is the object's *earliest* appearance to the perceiver, since in this case the perceiver has no memories or related symbols to add to the object's impact and thus to the symbol's mind-dependent content. When someone perceives a doorknob after having formed memories of interacting with doorknobs, the later perception reflects not just the object's striking the perceiver, but the perceiver's previous experience. So to the extent that the mind-dependent content of a certain symbol reflects, as Fodor argues, the way an object with certain features strikes a type of creature, the content should be determined by the creature's first encounter with

---

a fact that requires an explanation, namely the fact that encounters with doorknobs cause the acquisition of DOORKNOB.

[26] Fodor (1998) calls this the doorknob/DOORKNOB problem. He asks why it is "so often experiences of doorknobs, and so rarely experience of whipped cream or giraffes, that leads one to lock to *doorknobhood*" (127).

the object, which is when the creature acquires the symbol by locking to the property the object has in virtue of its way of appearing to the creature.[27]

In the case of a concept with mind-independent content, there is no such necessary connection between the causal relation of acquiring a concept and the semantic relation between the concept and its denoted objects. Instead, there is a correlation between the detectable properties in virtue of which certain objects trigger a locking mechanism for acquiring a concept, and the kind-constituting property denoted by the concept. For example, the stuff that first appears to creatures as water tends to be chemically composed of H_2O molecules, and animals that first appear to be dogs and that thus cause a perceiver's acquisition of DOG usually are dogs rather than foxes. These statistical facts might be explained in evolutionary terms. The creatures with locking mechanisms for recognizing certain distal types tend to operate in the same environment in which early instances of the mechanisms succeeded in recognizing these types, or in correlating certain detectable properties with certain kind-constituting ones, increasing the creature's fitness, and so the later instances of the mechanisms tend also to succeed. Also, for social or more directly genetic reasons, extra care tends to be taken by procreators to ensure that their offspring encounter the semantically correct types as they acquire concepts by a locking mechanism. Still, mistakes are possible in the acquisition of concepts of mind-independent types, which is why DOG can be acquired as a result of perceiving a fox that shares enough detectable properties with dogs to cause the locking

[27] The claim that DOORKNOB and other such concepts have subjective or mind-dependent content is surely questionable. However, Fodor's point isn't that the mind-dependent kinds aren't real, but that these kinds are relational, that they are constituted by a type of creature's engagement with a set of objects. This type of moderate realism about doorknobs is at least less questionable than outright antirealism about them, assuming an antirealist position could be maintained, according to which a mind-dependent type has no reality at all even while there may be the illusion or the appearance of instances of this type.

mechanism to token the wrong concept. In this case, the causing of the concept's being acquired doesn't match the concept's being about a certain type. But under conditions that are suitable to telling dogs apart from foxes on the basis of the perceptual evidence, a fox's causing the acquisition of DOG wouldn't generally happen, and so under these conditions, dogs are the animals that tend to have the detectable properties that cause the early tokening of DOG.

As to the nature of the locking mechanism, Fodor (1998) says that "If our minds are, in effect, functions from stereotypes to concepts, that is a fact *about us*. Indeed, it is a *very deep* fact about us" (140). I think that "messy" can replace "deep" in the last quoted sentence. Instead of a single mechanism for acquiring nonscientific concepts, there's likely a large set of mechanisms, including neurological, evolutionary, psychological, and social mechanisms. Some of these mechanisms, such as social ones, may not be innate. However, granted that there is likely this complication, I'll continue to simplify and to speak of just one locking mechanism for acquiring nonscientific concepts. Whether the locking mechanism is innate isn't crucial to my argument; rather, what matters is that the mechanism's behaviour is nomically determined, as opposed to being determined by a higher-level relation between nomic relations.

2.8 The Inconsistency of ADT and LMT

I've argued that Fodor's theory of content is meant to account for what he calls the robustness of semantic relations, but that there are two interpretations of this robustness, the empirical and the metaphysical. The empirical one can't be correct,

because the claim that semantic relations are empirically robust is consistent with a teleological theory of these relations, whereas Fodor says the main problem with teleological theories is that they can't explain semantic robustness. A teleological theory does indeed seem unable to account for metaphysically robust content. However, I think Fodor's theory of content is just as unable, since he assumes that symbol tokens are nomically related to the semantically relevant type, as implemented by a mechanism that works under certain conditions; moreover, Fodor posits just such a mechanism to explain how mental symbols are acquired. It's time now to show that there is a contradiction between his theory of content, which assumes semantic relations are metaphysically robust, and his assumption that symbols are mechanically implemented.

There is some reason to believe, on the contrary, that talk of the locking mechanism, in particular, is just another way of talking about asymmetric dependence. Fodor (1998) doesn't repudiate his theory of content, given in Fodor (1990). Granted, the later work on concepts doesn't discuss that earlier theory. However, as Viger (2001) points out, Fodor's theory of the locking mechanism is adapted from Loewer and Rey (1991), and they themselves assume Fodor's theory of content, defining "locking on" in terms of asymmetric dependence. Indeed, Loewer and Rey begin by saying "let's define a predicate, '*x* is locked onto *y*,' to capture this asymmetric causal structure" (xxvii). Thus, they take talk of locking on to be another way of talking about asymmetric dependence. Fodor (1998) himself seems to equate these ways of talking when he says, "Locking reduces to nomic connectedness. (I hope). See Fodor (1990[a])" (145, n.18), citing his work on his theory of content. For the theory of the locking mechanism (LMT) to be a restatement of the semantic theory of asymmetric dependence (ADT), the property a type

of creature locks to would have to be determined by a metaphysical asymmetric dependency between nomic relations. For example, a creature might lock to *dog* not because of what the locking mechanism has been naturally selected to do when special conditions are met, such as when the creature perceives a stereotypical dog, but, roughly speaking, because the creature would sooner lock only to this property, given that experience, than to lock only to any other property, such as *fox*, given the same experience.

But LMT can't be equivalent to ADT. For one thing, Fodor says the locking mechanism has to be triggered by actual perception of stereotypical particulars, which means that when a creature encounters a particular that lacks the stereotypical triggering properties, the creature doesn't acquire a certain concept by the locking mechanism. This point about the need to perceive certain particulars seems to conflict with the assumption that content is metaphysically robust just as much as does the actual history condition of Fodor's theory of content (AHC). According to AHC, a symbol's content depends on some of the symbol tokens' being actually caused by semantically relevant objects, and this conflicts with the metaphysical claim that the determination of content is independent of how a symbol type might ever actually be tokened. On the assumption that ADT is more closely connected to the thesis of metaphysical robustness than to AHC, LMT's claim that concepts are acquired by actually perceiving certain particulars shows, at least, that LMT isn't just a restatement of ADT.

Also, although LMT isn't supposed to be a theory of content, LMT seems to lay out conditions for determining content, and these conditions differ from those given by ADT. On ADT, content is metaphysically robust, whereas on LMT the content of

symbols acquired by a locking mechanism seems to be causally determined. A teleological theorist could say that a symbol with mind-dependent content has its content because of the way certain objects strike the creature who has the potential to acquire the symbol. What would be teleological about this is that this causal relation would be implemented by a naturally selected mechanism, and this sort of mechanism is often thought to have a purposive function. Now, a symbol with mind-independent content would have its content because of the strength of a correlation, when certain conditions are met, between a set of detectable properties that trigger the locking mechanism, and the objective property to which a creature locks, which is also the property shared by the objects denoted by the symbol. Again, this causal relation would be implemented by a naturally selected mechanism. In this way, a teleological theorist can use LMT to explain content, because LMT addresses a problem that is similar to the problem of explaining content in a naturalistic way, which is the problem of why the semantically relevant objects cause the acquisition of concepts. Informational and teleological theorists assume the problem of naturalizing semantic relations is to show that these relations arise out of, or are nothing but, causal or other lower-level natural relations. The doorknob/DOORKNOB problem that Fodor thinks LMT solves is, in effect, the problem of how the causal relation involved in acquiring a concept could be the semantic relation between the concept and its denoted objects.

LMT addresses the problem of concept acquisition by positing a locking mechanism that works under special conditions. Indeed, LMT is committed to saying there are, in effect, non-metaphysical, historical asymmetric dependencies as opposed to metaphysical, synchronic ones, and a teleological theorist can use the historical relations

to explain content. This is because conditions under which the semantically relevant objects cause the acquisition of a concept, and thus under which a causal relation amounts to a semantic one, are those under which these objects *first* cause the symbol tokens. In the case of subjective content, the earliest appearance of an object is the purest case of a symbol's having mind-dependent content because of the way an object by itself strikes the creature. In the case of objective content, there are evolutionary reasons why objects that will be denoted by a symbol tend to cause a creature to acquire the symbol. There may be other, abNormal sets of conditions under which a creature that has already acquired a symbol will be caused to use the symbol by its denoted objects. But there would be no conditions under which semantically irrelevant objects cause the symbol tokens, or under which the symbol actually misrepresents its distal cause, were the symbol not acquired by members of the species who first benefited from having the locking mechanism. Once again, though, a causal or a teleological theory of content doesn't account for metaphysical robustness, or for the independence of semantic relations and of special, context-dependent causal relations.[28]

For these reasons, Fodor needs to distinguish between ADT and LMT; otherwise, in exchanging ADT for LMT, Fodor would no longer have a theory of content that succeeds where other theories fail, in explaining metaphysical semantic robustness. I think Fodor's well-known distinction between generalizations about mechanisms that implement special nomic relations, on the one hand, and generalizations about the nomic

[28] There is a type-token distinction that is relevant here, between the acquisition of a concept by a species and by a member of the species. As I suggested earlier in this section, there are evolutionary reasons why the conditions for the earliest instantiation of a symbol type for either a species or for a member of the species might be crucial to determining the symbol's content.

relations themselves, on the other, applies here.[29] For Fodor, this distinction is based on his view of special science laws, given in Fodor (1974). He argues that special sciences are autonomous, meaning that their laws are not reducible to lower-level laws, because their laws quantify over relations between multiply realizable properties. A multiply realizable property is one possessed by particulars that are regarded as instances only of heterogeneous kinds when these particulars are characterized without the special science's theoretical framework. Fodor (1989) goes further and says that what distinguishes special from basic natural laws is, "Nonbasic laws want implementing mechanisms; basic laws don't" (155). A physical nomic relation is primitive, since there is no further question of how, as a result, say, of some change of microstructure, the physical structure came to be. A special nomic relation, though, is always dependent on some lower-level process, and because the lower-level process works in its own way, the higher-level process may be disrupted, resulting in an exception to the special science law and the need for this law's *ceteris paribus* conditions.

So, then, Fodor can say that LMT and ADT explain different aspects of symbols. The vocabulary needed to explain how a concept is acquired is different from that needed to explain how a concept's content is determined. In particular, talk of mechanisms and of causal and historical relations applies to concept acquisition, while talk of metaphysical asymmetric dependencies applies to content determination. The laws of how certain locking mechanisms work are different from the metaphysical

---

[29] See, for example, Fodor (1989): "the vocabulary that's appropriate to articulate a special-science law is systematically different from the vocabulary that's appropriate to articulate its implementing mechanism(s)" (146). Fodor and Pylyshyn (1988) use this distinction in defense of the classical view that cognitive processes are digital computations over strings of symbols, not just neural processes or simplified versions of these processes as the connectionist contends. Specifically, the systematic nature of cognitive processes can be explained only in classical, not in connectionist terms. At best, say Fodor and Pylyshyn, the connectionist's talk of changes in connection strengths between nodes in a network provides a model of the *mechanism* that implements the cognitive software.

generalizations about dependencies between certain nomic relations. Thus, LMT is a theory of how some implementing mechanisms work, not a theory of content, and the metaphysical semantic generalizations don't reduce to mechanistic ones. Actual symbols with semantic properties are dependent on the locking mechanism, in that an instantiated symbol has to be acquired in the first place, but explaining how a symbol is acquired isn't the same as explaining the symbol's semantic relation to some set of objects.

Note, however, that while Fodor can distinguish between these two levels of explanation, at least in principle, he can't say that their difference lies in the requirement of a mechanism to *establish* the asymmetric dependency. On the contrary, for symbols to be metaphysically robust, the asymmetric dependency must be established metaphysically prior to the actual meeting of conditions for a mechanism to satisfy a CP law. Were a mechanism or any causal relation or Normal situation to determine the asymmetric dependence of any semantically irrelevant causal relation on the semantically relevant one, once again Fodor's theory of content would reduce to a teleological one. A mechanism can implement a special nomic relation, but can't account for what are supposed to be *meta*physical facts about relations between special causal relations. Moreover, any mechanism that could implement the asymmetric dependency between special causal relations would be explainable simply in terms of the higher-level causal relation that, say, the mechanism produces symbols that are caused by *X* only because they are caused by *Y*, and not the other way around. Instead of being a metaphysical fact, the asymmetric dependency would be just a fact explained by the special science that explains a pattern produced by the mechanism. Thus, a teleological explanation of the mechanism's work would account directly for semantic relations, as far as Fodor's point

about asymmetric dependence is concerned, albeit without accounting for content's metaphysical robustness.

In any case, even if LMT isn't just a restatement of ADT, and the two theories operate at different levels of explanation, the two can contradict each other and they indeed seem to do so. Fodor needs a way to affirm both LMT, as a nonsemantic theory of concept acquisition, and ADT as a theory of where content derives from. On ADT, content is metaphysically robust, but on LMT content is not metaphysically robust, because the locking mechanism amounts to a special way of causing symbol tokens by the semantically relevant objects. LMT says there are conditions under which a type of symbol is unable, or at least unlikely, to misrepresent the cause of its tokening, because there tends to be only one type of cause under those conditions, namely the denoted type. While ADT makes a semantic relation independent of the conditions needed for a mechanism to implement a special causal relation, LMT lays out conditions that have to be met for the acquisition of a symbol type, and these happen to be conditions under which the symbol tokens are caused by the denoted objects.

To see how exactly the conflict arises, consider, as an example, how the locking mechanism is supposed to implement concepts that have subjective, mind-dependent content. In the case of DOORKNOB, for example, the concept denotes, in effect, the type determined by how some particulars strike the perceiver when the perceiver acquires the concept by locking to a property. This means that, in general, the way in which some particulars strike a perceiver—when the perceiver acquires this sort of concept—*already determines* this concept's mind-dependent content. To say that the content here is mind-dependent, or subjective, is to say that the content depends on the *process* by which a

type of creature, with a working locking mechanism, locks to some property, by perceiving an object with certain detectable properties. This seems to imply that there can be no mistake made in the case of recognizing a mind-dependent type. If a doorknob is just whatever *seems* to be one, then whatever causes the tokening of DOORKNOB, or whatever has detectable properties the perceiving of which causes someone to understand them in a certain way, must be a doorknob. If a non-doorknob couldn't be mistaken for a doorknob, there would be no possible world in which non-doorknobs cause the tokening of DOORKNOB while doorknobs fail to do so. Moreover, the mind-dependent semantic relation would depend on the locking mechanism, since the content would be defined in terms of it. In this case, the content of DOORKNOB would be neither metaphysically nor empirically robust, and so Fodor's account of how subjective concepts are acquired would conflict with an assumption of ADT.

Suppose, though, there were no such necessary connection here, between the causal relation that makes for the acquisition of a concept with mind-dependent content, and the relevant semantic relation. So DOORKNOB could somehow have mind-dependent content even though non-doorknobs could be mistaken for doorknobs. In this case, Fodor's mechanistic level of explanation would still undermine his metaphysical one. If what it is for an object to "strike" a creature in a certain way is dependent on how the object strikes the creature for the first time, and thus on how the locking mechanism is triggered, resulting in the acquiring of a concept, any possible world in which, say, non-doorknobs cause DOORKNOB tokens would be one in which doorknobs cause these symbol tokens *first*. The symbol's subjective content would be already determined by the initial way in which the symbol token would be caused in a creature, since the acquiring

of the concept by the locking mechanism would be the process that generates the mind-dependent property of being the type of object that is denoted by this concept. In this case, there would be no asymmetric dependence of the nomic relation, between non-doorknobs and DOORKNOB tokens, on that between doorknobs and these symbol tokens, because there would be no possible world in which *only* non-doorknobs cause the symbol tokens. Whatever triggers the locking mechanism and causes the concept to be acquired in the first place would be an instance of the semantically relevant mind-dependent type, and so something semantically irrelevant could cause the symbol token only if semantically relevant objects do, and indeed only if they would have already done so in that very world. But a symbol has metaphysically robust content only if there's a possible world in which the symbol keeps its metaphysically determined content even though the symbol token is never caused by anything denoted by the symbol, so that the symbol might always be used in a misidentification. Were a symbol with mind-dependent content always to be caused by something it denotes, whenever the symbol token is first instantiated in a creature, LMT would contradict ADT's assumption that content is semantically robust, that a semantic relation could be what it is regardless of how the symbol token is caused under any conditions.

But suppose the above assumption is false, and what it is to be an instance of a mind-dependent type isn't dependent on the process by which the concept is acquired. Suppose, for example, that DOORKNOB can be acquired by perceiving a hologram of a doorknob, something that has only some of the detectable properties that trigger the locking mechanism. A holographic doorknob might have relevant visual but not tactile features; a hand that tries to use the object by grasping it would pass through the

hologram, and the object then wouldn't strike us as a doorknob after all. But perhaps the holographic doorknob has enough of the relevant detectable properties to trigger our locking mechanism for DOORKNOB. Still, this would not be a possible world in which only holographic doorknobs cause the symbol tokens, since similar objects that have also the tactile properties would likewise cause these symbol tokens. So once again, the content of DOORKNOB would lack metaphysical robustness: the semantic relation here would depend on how the symbol tokens would be caused were certain conditions met.

That is, it's not as though the semantic relation would be determined merely by the asymmetric dependence of one nomic relation on another one, without reference to a certain fact about any world in which there is the type of symbol in question. DOORKNOB would have metaphysically robust content only if the symbol could keep its content regardless of how the symbol tokens are caused under certain conditions. But if any world in which there is this symbol type would be one in which the symbol tokens are caused by the denoted objects under certain conditions, such as when the object's tactile properties are detected, there would be no way to show that DOORKNOB has metaphysically robust content, or that the semantic relation really is independent of any causal one. This is because there would be no counterexample to the view that the content lacks this robustness, since there would be no possible world in which the symbol retains its content despite the fact that the symbol tokens aren't caused by their denoted objects in that world. Assuming there are always conditions under which DOORKNOB tokens are caused by doorknobs, even though, under different conditions, these symbol tokens are caused by holographic doorknobs, the teleological theorist can point to what happens when one of the sets of conditions is met as that which determines the content. For

example, the content of DOORKNOB might be whatever causes the symbol token when a creature detects *all*, or as many as possible, of the object's detectable features that trigger the locking mechanism.

Finally, suppose there is, after all, a possible world in which DOORKNOB is caused by non-doorknobs and not by doorknobs. The question, then, is how the symbol could be said still to have *mind-dependent* content, or how the symbol could be acquired by a process of detecting certain surface properties and of locking to another property, where the denoted type is nothing more than the type locked to by this process. Assume that, for some reason, holographic doorknobs, but not real ones, cause the symbol tokens, and yet the concept at issue is DOORKNOB and not the concept of holographic doorknobs. In this case, the content would have to reflect the *only* way objects with certain detectable features could appear to a type of creature even though the objects that do so appear wouldn't be included in the extension. Again, were there other ways in which objects with those (and with other) detectable features could appear, real doorknobs might cause the symbol tokens under certain conditions, which is not what is here being supposed. Instead, were the semantically relevant objects incapable of causing the symbol tokens, the assumptions are that these symbol tokens could be semantically related to doorknobs; that doorknobs could be nothing but the objects whose detectable properties cause the objects to appear in a certain way to a type of creature; that doorknobs possess at least some of the detectable properties of holographic doorknobs; and yet that only non-doorknobs cause the symbol tokens. These assumptions are inconsistent, so with regard to a symbol that has subjective content, the symbol tokens must be potentially caused by their denoted objects whenever the symbol tokens are

caused by semantically irrelevant ones with similar detectable properties. There is no way to show that the mind-dependent content is metaphysically robust, by showing how the semantic relation could be instantiated without any potential instantiation of the semantically relevant causal relation. So given LMT's account of how these symbols are acquired, the symbols lack metaphysically robust content, and ADT, which assumes otherwise, is false.

The upshot is that it seems mind-dependent content isn't metaphysically robust, and that LMT's account of the acquisition of symbols with such content contradicts ADT. Moreover, the non-robust, mind-dependent content seems explainable in terms of the locking mechanism and thus in teleological terms that are inconsistent with ADT. Of course, there should be no surprise that ADT conflicts with a teleological theory of content, since Fodor offers ADT on the assumption that content is metaphysically robust and thus that a teleological theory is false. But what is surprising is that mind-dependent content can't be metaphysically robust, given LMT. There could be no such content without the perceptual process by which these concepts would be acquired by a locking mechanism. Therefore, if these concepts are acquired in this way, there is no need for ADT to explain their content, and so it is just as well that LMT supplies a teleological explanation of content that is consistent with LMT's own explanation of the acquisition of these concepts.

I won't go into detail here with regard to LMT's account of objective, mind-independent content. The contradiction is more striking in the case of mind-dependent content since, on one interpretation at least, there's a necessary connection in that case between the semantic relation and the work of the locking mechanism. There's only some

looser connection between the mind-independent semantic relation and the locking mechanism's work, since in this case the surface, triggering properties are only correlated with some objective, type-distinguishing property possessed by the denoted objects. But the contradiction still arises. Under conditions that are suitable to distinguishing between two sets of surface properties, say those of dogs and foxes, the locking mechanism locks to the objective property correlated with one of the sets of surface properties, depending on which animal is perceived. So what the locking mechanism does under these conditions, in effect, determines the semantic relation. The reason there's no necessary connection here is that the *actual* conditions under which a concept with objective content is acquired may not suffice for the locking mechanism to distinguish between the two sets of surface properties, and so perceiving a fox can cause the acquisition of DOG.

Suppose, however, two objectively different distal types have exactly the same surface properties, as in the case of H_2O and XYZ, where XYZ is a liquid that has the same surface properties as H_2O. I can think of three ways the locking mechanism might still determine the content of the concept acquired by perceiving a sample of either liquid. In one scenario, the locking mechanism might have an indexical component, so that the objective, type-distinguishing property to which the mechanism locks is just an underlying one that is causally related to the surface properties. So perceiving samples of liquid that have the same surface properties might still cause the locking mechanism to lock to different underlying properties, even though the locking mechanism alone won't reveal the underlying properties to the concept's possessor. Alternatively, the premise that the two liquids have different underlying, but identical surface, properties might be questioned. Whenever two things are instances of objectively different types, with

different underlying properties, there may be conditions under which they display different surface properties. By hypothesis, XYZ has a different chemical composition than H_2O, and this difference might cause, say, a distinctive boiling time of a sample of XYZ. So the locking mechanism might determine mind-independent content, under conditions that suffice for the object's underlying properties to surface, as it were, in the perceivable properties that trigger the mechanism.[30] Finally, if the intuition is that at the time of perceiving a sample of either liquid, the acquired concept doesn't refer to a type defined by the sample's underlying properties, the mechanism might lock to the disjunctive type, *H_2O or XYZ*, in which case the concept would have disjunctive content. The mechanism would still determine the content, but in this case the surface rather than the underlying properties would be crucial. Later on, when the liquids' different underlying properties are discovered, a new concept, with narrower content, might be acquired by other means, replacing the one acquired by the locking mechanism. The content of this later concept would likely be determined, in part, by other symbols used in the process of discovering the different underlying properties, so accounting for this content isn't the job of the sort of naturalistic theory in question here.[31]

In each case, the teleological theorist has a way of showing that mind-independent content is determined by what the locking mechanism does when conditions enable the

---

[30] It does seem unlikely to me that objectively different types could have exactly the same surface properties, under conditions that are ideal for the perceiving of either type. Even though foxes and certain dogs may look the same in the dark, to creatures without night vision, a teleological theorist will say that the Normal situation in which content is determined includes conditions that are suitable, say, to a naturally selected trait's ability to perform its purposive function. If a mechanism for acquiring a concept isn't supposed to work in the dark, it doesn't matter to this sort of theory of content that certain distal types can be mistaken for each other in such an abNormal situation. What matters is whether they're mistaken for each other under conditions that are optimal to the mechanism's ability to fulfill its function, or to satisfy some CP law, given natural selection or some other source of so-called objective norms. Most, if not all, objectively different distal types would seem to have distinguishing surface properties under Normal conditions.

[31] This is just what happened in the case of the concept of jade, which split into the concepts of jadeite and nephrite when the underlying objective differences were discovered.

mechanism to lock to some type that correlates with a set of surface properties. This is so whether the distal type is a narrow and undetectable one, such as *XYZ*, a similarly narrow but detectable one, or a broader one such as H_2O *or XYZ*. But metaphysically robust semantic relations aren't determined by any such mechanism, since they're independent of any causal relation instantiated under any set of conditions. Thus, ADT contradicts LMT.

In summary, I've argued that whether the content of syntactically simple concepts, acquired by a locking mechanism, is subjective or objective, the content seems not to be metaphysically robust. Although LMT is supposed to be a theory simply of the mechanism that causes a symbol to be acquired, I've tried to show that LMT implies that a semantic relation's determinacy depends on a special causal relation. This causal relation is needed for the symbol type's acquisition. Even if a symbol is sometimes acquired from encounters with objects not denoted by the symbol, the reason one symbol type is acquired rather than another is that under certain conditions, only encounters with the denoted objects cause the acquisition of the symbol. LMT implies that content is not metaphysically robust, since concepts acquired by the locking mechanism don't have their content regardless of a mechanism's work. A teleological theory of content is derivable from LMT, and so LMT is incompatible with ADT's assumption that a teleological theory is false. If, however, ADT is assumed, some other explanation of how symbols are acquired is needed. This would have to be an explanation that doesn't assume a symbol is implemented by a mechanism that works when conditions are met for a special causal relation's instantiation.

Assuming, though, any naturalistic account posits such a mechanism, at least on a construal of how a naturalist should explain content, which I reject in Chapter 5, metaphysically robust semantic relations seem hard to naturalize. If semantic relations are metaphysically robust, a symbol type mustn't be implemented by a mechanism in that sense. In Chapter 6, I argue, indeed, that if the content of a mental symbol is determined by a norm, or by how the symbol ought to be used, the symbol can't depend on a mechanism's work, in so far as the symbol is something that enters into a semantic relation. And saying that content is normative amounts to saying that content is at least metaphysically robust, since in either case the semantic relation is independent of the work of any mechanism.

2.9 Is Metaphysical Semantic Robustness just Multiple Realizability?

Finally, I want to consider whether ADT and LMT can be shown to be consistent with each other, on the assumption that a semantic relation's metaphysical robustness can be equated with the relation's *multiple realizability*.[32] To take an example of what is often regarded as a multiply realized property, *pain* is found in nomic patterns that aren't explained in terms of the mechanisms that realize the property of being in pain, because the property might be realized by disparate mechanisms. Thus, there is human pain, but also the pain potentially of intelligent extraterrestrials, with different physiologies, and of inorganic, artificially intelligent machines created by humans. The important point is that

[32] The idea that a property can be multiply realized is a functionalist idea from the philosophy of mind. Resting on a distinction between causal role and an occupant of the role, the point is that a mental property might be identified with the causal role played by different mechanisms. I'll consider this sort of functionalist point, in section 6.3, as a possible response to my account of the normativity of mental content.

while there would be no pain without *some* mechanism realizing the property, pain in general isn't explainable in terms of any specific mechanism. For example, extraterrestrial pain isn't explainable in terms of human brains, since the extraterrestrial might have a different kind of control center. Similarly, semantic relations might be realized by different mechanisms, including the locking mechanism. So while the instantiation of a semantic relation naturally depends on some mechanism that realizes the relation, the relation isn't explainable in mechanistic terms, contrary to a teleological theory of content. Therefore, LMT doesn't imply that semantic relations depend on causal relations in a sense that threatens ADT's assumption that content is metaphysically robust.

A teleological theorist, however, can reply that the conclusion doesn't follow. Once the point is granted, that a *kind* of semantic relation depends on a *kind* of mechanism, in the same way that human pain depends on the human brain, the implication is that content isn't metaphysically robust. Human pain is determined by a type of neural mechanism, alien pain by a different physiology, and AI pain by yet another set of mechanisms.[33] An account of the Normal conditions that allow the human brain to realize the property of being in human pain accounts also for the limits of when and how this pain is felt. Perhaps the subjective aspect of pain isn't explainable in mechanistic terms, but the functionalist's point about multiple realizability grants that a kind of pain would be determined by the workings of a kind of mechanism.[34] This is

---

[33] This is Kim's point about local reduction, which is opposed to a functionalistic account of multiple realization. Kim (1989), for example, speaks of "species-specific biconditional laws" connecting laws about the mental states of a species to laws about the species-specific neurological realizers of these states.

[34] In one of the surprisingly few explicit discussions of what it is to "realize" a property in the sense relevant to the thesis of multiple realization, Polger (2007) assumes the relevant sense is of a "non-destructive, non-causal, synchronic dependence relation," with the function, or role, depending on the properties of the role's occupant. Following Gillett (2003), Polger assumes that whatever has a function, or

enough to contradict ADT's assumption that content is metaphysically robust, that a semantic relation isn't determined by any causal or mechanistic relation, or by any Normal situation.

Suppose, however, some limits of semantic relations are like the more general features of pain shared by human, extraterrestrial, and AI pain. For example, the way different symbols have different contents might be compared to the nomic relation between pain and defensive behaviour. The latter relation might not be explainable in terms of any specific mechanism, assuming there are different ways of causing defensive behaviour as a result of being in pain. Likewise, one way to determine content might be a locking mechanism, but there might be other ways as well, and so the property of having determinate content wouldn't be explainable in terms of any one of these mechanisms.

Still, in the case of the nomic relation between pain and defensive behaviour, all of the mechanisms would seem to have enough in common that there would be a *broader* mechanistic account of the nomic relation. In particular, this nomic relation would have to be implemented by *brains* of some type. Only were the mechanisms so disparate that no scientific account were possible of the realizers of pain's causing of defensive

---

a multiply realized property, must have whatever is needed to *individuate* the function. As Polger summarizes Gillett's view, "a system *s* instantiates a certain property *G* when it or its parts have the causal powers that individuate *G*," where "instantiates" is synonymous with "realizes" (238). Gillett thinks a function is individuated by certain causal powers of whatever occupies the role at some mereological level. Polger disagrees, pointing out that the mathematical functions relevant to artificial intelligence aren't causally individuated. Polger's view is that "To occupy a role is to have the relations that are *distinctive* of the role," and these relations may or may not be causal (251, my emphasis).

The only point I want to take from this is that whatever is supposed to realize a property has that which individuates or distinguishes the property, which is the same as having that which determines the property, setting its limits. To take an analogy, the role of Hamlet is distinguished or limited by different actors that play the same role. Although the role isn't *identical* with the work of any one actor, but is only *realized* by each performance, an actor sets the limits of the role by performing it, or by *instantiating* it. If a nomic relation is instantiated by a complex causal system (what I'm calling a mechanism), the system determines the nomic relation, even if different systems may determine the relation differently in other instantiations, or realizations, of the nomic relation. This point will be important in my later discussion of multiple realization, in section 6.3.

behaviour, would the claim that the determinants of content are multiply realizable refute my argument against ADT and LMT. Only if, say, a brain could feel pain and cause defensive behaviour, but so could a rock, or a nebula, or something completely different from a brain, could the nomic relation be independent of even a broadly conceived mechanism, so that there would be no teleological account of the nomic relation.[35]

In the first place, it's doubtful that systems of causal relations establishing a higher-level nomic pattern could be this disparate. Any property realized by human brains, by AI machines, and by rocks probably wouldn't figure in a special nomic relation (although physical laws would apply to all). But even were there some highly abstract property shared by very different systems, any theorist who could recognize such a property should be able to conceive of a property shared by the realizing mechanisms that determines, or sets the limits of, the higher-level property. Were there in fact no such lower-level property, and no broad mechanistic account of the highly general nomic relation between, say, pain and defensive behaviour, this would be a reason to think there is no such nomic relation, that the terms used in the CP law are empty or subjective.

---

[35] The idea that functionalism in the philosophy of mind supports the thesis of multiple realization derives from Turing's mechanistic account of computation. A Turing machine is definable in terms of a mathematical function that makes no reference to the machine's physical makeup. So if cognitive processes are computational, and a Turing machine can carry out any computation, cognitive processes are likewise thought to be definable without reference to the mechanisms that implement the processes. Given the thesis of functional equivalence, that systems with the same functional descriptions have the same mental states, systems can have the same mental states with different hardware implementing these states.

But more recently, connectionists, dynamical systems theorists and other critics of functionalism have argued that cognitive processes are not so independent of implementation. Eliasmith (2002), for example, follows Kolmogorov in arguing that in the real world in which computers have finite symbol-strings, time, and other resources with which to operate, the complexity of the hardware limits the algorithms that the hardware can carry out. The same algorithm can be run on different implementations only with the help of an emulator program, which increases computational complexity, which in turn significantly affects performance. "Since an increase in computational complexity necessitates an increase in the time and power needed to perform a computation, the class of actual computable functions *within a given period of time and with a fixed amount of computational resources* will vary for different physical computers" (5). The claim that there are no constraints on the mechanisms that can implement a certain property abstracts away from these real-world considerations, and isn't strictly true even in the case of properties that can be computationally defined.

Were there instead such a lower-level, physiological property, this wouldn't mean that talk of pain is reducible to talk of the physiological property. Pain might have epiphenomenal properties or other features that don't figure in scientific laws; moreover, talking about pain in terms of the realizing property might be impractical for certain purposes. In any case, even if the locking mechanism weren't a necessary determinant of semantic relations, this mechanism would seem a sufficient one. So were human primitive concepts acquired by a locking mechanism, and were this mechanism (and the Normal conditions of its working according to a lower-level CP law) to account for why different concepts refer to different distal types, the content wouldn't be metaphysically robust. Thus, LMT would imply a teleological theory of content that contradicts ADT. For these reasons, I don't think the equating of metaphysical robustness with multiple realizability helps the proponent of both ADT and LMT.

2.10 Conclusion

Fodor's theory of content can be seen as an attempt to deal with a worst-case scenario for the naturalist interested in explaining semantic relations. The idea is that if content is metaphysically robust, and therefore not nomically determined, there is still a naturalistic explanation of content. The explanation is given in terms of subjunctive conditionals, which are needed anyway to make sense of the nomic relations found in basic naturalistic ontology. The asymmetric dependency is metaphysically prior to the work of mechanisms or to the evolution of particulars in the natural world. Fodor explains semantic relations as ways in which causal relations are themselves organized,

but the special causal relations posited in explanations of neurally implemented primitive symbols have their own mechanistic, evolutionary and thus teleological explanations. By turning to basic naturalistic ontology as his starting point in naturalizing content, while also affirming the metaphysical robustness of content, or the independence of semantic relations from relations found in this ontology, Fodor is left with the problem of showing how symbols with that sort of content could be implemented by a naturally selected mechanism. While ADT doesn't posit normative determinants of content, metaphysical robustness plays the same role in ADT as does a certain normative determinant in Dretske's and in Millikan's theories, as should become clear in the next few chapters.

The deep problem for this approach to explaining content is to explain the disorder in, or wide variety of, the ways symbol tokens are actually caused, while also explaining semantic relations as dependent on the deepest source of order in nature, that is, on anything from basic naturalistic ontology, such as a nomic relation. For Fodor, the disorder is due to the metaphysical priority of a certain asymmetric dependency, and thus to the possibility of a mismatch between the dependency and an actual way of causing a symbol token; there is a lack of pattern in the way symbol tokens are actually caused, despite their having determinate content, because a semantic relation is pre-established and isn't affected by the circumstances under which symbols are actually used. For Dretske and Millikan, the determinants of a semantic relation withstand disorder in how symbols are actually caused or used, because the determinants are purposive functions and thus are in some way normative. Unlike a descriptive law, a prescriptive norm needn't actually be followed by the symbol-user to effectively determine a semantic relation. In either case, though, there's a conflict between positing such determinants of

content, and identifying intentionality fundamentally with a causal relation or with anything already posited by naturalists to explain, at a metaphysical level, the scientifically-discovered natural order. I'll return to this diagnosis of the internal conflicts in sections 5.2 and 5.3. Next, though, I turn to Dretske's theory.

Chapter 3

Receptivity, Information, and Learning: Dretske's Theory of Content

3.1 Introduction

As I pointed out in the last chapter, Fodor's theory of content is supposed to succeed where a teleological one fails at solving the problem of the metaphysical difference between intentionality and information. Information is nomically determined, and therefore a token signal's informational content is just the nomically sufficient condition for tokening the signal. A symbol with semantic content, however, can be caused in all sorts of ways, including semantically irrelevant ones, which means there isn't simply a nomic relation that determines the symbol's semantic content. Instead of being nomically determined, semantic content is metaphysically determined by an asymmetric dependency between the semantically correct way of nomically producing the symbol token and any of the semantically incorrect or less relevant ways of doing so. But this raises the question of how a symbol's semantic content could be determined in this way, from above, as it were, without being affected by how the symbol is implemented by a mechanism. This is especially puzzling, given the semantic implications of Fodor's own theory of how symbols are acquired by a locking mechanism.

Dretske's theory of content is also fundamentally an informational theory, but Dretske doesn't claim that semantic content is metaphysically robust. He argues, instead, that a symbol's semantic content is the informational content the symbol is supposed to carry, given how the symbol is implemented by a mechanism which the creature itself sets up when the creature learns that this function is needed to get what the creature wants. Whereas a symbol token with metaphysically robust content need never be caused by the semantically relevant object, a symbol token with teleologically determined content at some point must be so caused. This is because the mechanism has a function which is the result of a process of development. In this case, what develops is how an individual creature is internally configured as its inclinations are reinforced and the creature learns how to control its movements by internal representations of external conditions. When a creature's internal conditions control its movements because of their relations to external conditions, the creature is said to *behave* based on its own *reasons*, on its beliefs and desires. The creature is then said to be a semantic rather than a syntactic "engine," which means that the external, semantic relations can be causally responsible for the creature's behaviour.

On Dretske's view, the symbol's actual indication of what will become the semantically relevant external object is part of the process by which the symbol acquires the purpose of being nomically related to that type of object. Once the purpose of the relevant mechanism is established, however, the environmental conditions are free to vary so that the symbol may be nomically related to types that are semantically irrelevant to the symbol. This is because the symbol's purposive function remains. Just as a metaphysical asymmetric dependency is the sort of thing that can supposedly control a

symbol's semantic content without being affected by any of the nomic relations that determines how the symbol type is tokened under various conditions, a certain mechanism's function is supposed to persist despite the same sort of nomic heterogeneity. Whereas Fodor distinguishes between nomic relations and asymmetric dependencies, the teleological theorist distinguishes between Normal and abNormal situations, and argues that what a signal does in a Normal situation serves as a sort of standard for what the signal is supposed to do in an abNormal one. On Dretske's view, the Normal situation precedes the abNormal one, making his teleological theory an etiological one, according to which a symbol's content is determined by conditions at the symbol's origin.

That, at least, is a summary of Dretske's theory. In this chapter, I present Dretske's theory in more detail (sections 3.2 and 3.3). Next, in section 3.4, I consider a common objection to etiological theories of content, the Replacement Argument, and I argue that this objection can be overcome. Then I consider two of what I take to be much more serious objections. In sections 3.5 and 3.6, I argue that the learning process Dretske says turns some internal conditions into beliefs and desires doesn't produce a semantic engine after all. The problem I raise here is of whether the content of each token internal state is causally relevant to the creature's behaviour. The last objection I raise, in sections 3.7 and 3.8, is that Dretske's theory is, at best, incomplete, because the learning process depends, roughly, on the creature's interest in organizing itself, by way of its receptivity to the reward that reinforces some configuration of the creature's internal structures. Dretske should be committed to saying that this initial receptivity already has content, but this content isn't explained by his theory.

3.2 Nomic Relations and Structuring Causes of Behaviour

Drestske's theory of content should be seen, in part, as a response to the Twin Earth arguments in Putnam (1975). These arguments convinced many philosophers that a symbol's reference to something doesn't supervene just on the speaker who uses a symbol; reference isn't intrinsic to the speaker, since type-identical speakers can have symbols that are narrowly both about a liquid that has similar superficial properties, but that still refer to kinds with different underlying properties, such as to H_2O and XYZ, giving the symbols different truth conditions. This externalist view of content led Fodor (1980) to embrace methodological individualism as a strategy in cognitive science. Even were semantic relations somehow causally relevant to behaviour, differentiating between them would be difficult given their indexicality, their dependence on environmental contexts that may differ from one token speaker to the next. But precisely because reference is extrinsic to a speaker, reference relations shouldn't be relevant to cognitive science, with its prevailing computational perspective on the mind. If a mind is a computer program, the syntactic relations are the proximate causes of the body's behaviour.[1] Stich (1983), likewise, took externalism to show that semantic properties are causally irrelevant to an organism's neural states that cause its behaviour. If physical processes in the head control an organism's behaviour, and semantic relations don't

[1] What Fodor (1980) says about methodological solipsism is that naturalistic psychology is hopeless, since a naturalistic theory posits nomic relations, which are relations between properties, but a mental representation relates potentially to *any* real property. Thus, naturalistic psychology has to wait for all of the other sciences to say what these properties are, before psychologists can specify their own theoretical entities, the mental representations. As Stich (1983) points out, this overlooks Fodor's own way of showing that psychology is an autonomous science. Mental representations may be about H_2O, but not H_2O *as such*, and so the psychological law could be specified even were there no scientific explanation of water, given a folk characterization of the represented property.

causally affect these processes, content is irrelevant to the pursuit of a causal explanation of behaviour. By contrast, Dretske wants to argue that even though semantic relations are extrinsic to an organism, these relations can be causally relevant to the organism's behaviour.

According to Dretske (1988), the mind is a semantic rather than just a syntactic "engine", or what I'll call a semantic system.[2] A semantic system is such that its semantic relations are really part of nature in the sense that the system's mental states have causal power in virtue of their semantic properties. A syntactic system may be a mechanism, in the sense of a complex causal system that implements a nomic relation under special conditions, but the system's causal power derives only from the intrinsic properties of the mechanism's parts, such as their form or shape. However, these intrinsic properties interrelate in such a way that the system is interpretable as a semantic one; a syntactic system doesn't run on content, as it were, but behaves *as if* it were a semantic system. On Dretske's view, a semantic relation can itself have a causal impact on the world even without the work of an interpreter. In particular, content can cause an organism's behaviour, in that the content of some internal state may be the reason why the organism behaves as it does, and this reason can be explained in terms of how the organism came to be configured, or how the internal state came to have its internal role. In this way, beliefs and desires, as content-bearing mental states, can be reasons for purposive behaviour. But for content to have a causal role, a semantic relation has to be identified with some natural relation, as opposed to being theoretically eliminated as an illusion or reduced to something that depends on interpretation.

---

[2] Calling a creature an engine is a figurative or an archaic way of speaking, so I'll use "system" instead of "engine." Still, the distinction between the two kinds of systems is fundamental to Dretske's project.

Dretske (1981) argues that semantic relations are fundamentally informational. *X*'s being about *Y* is a matter of *X* carrying information about *Y*, and *X* does this if *Y* is made more probable, given *X*. This isn't exactly to say that *X* tokens are *caused* by *Y*; instead, the point is that the possibilities of what might obtain, which conflict with *Y*, are eliminated or at least narrowed, given *X*, so that *X* signals, indicates, and provides strong evidence, in effect, that *Y* is the case. That *Y* obtains can be learned from *X*. Indeed, Dretske (1981) is concerned mainly with an informational epistemology. Knowledge can be thought of as the having of a belief that *s* is *F*, when the belief is caused by the information that *s* is *F*; the belief carries a signal sent by the state of some object *s*. But what makes this belief knowledge is that the belief's truth is guaranteed by the ultimate source of information, which is the nomic relation. Although signals must be sent across channels of some sort, and received by detectors which can malfunction, the state of a signal's source can be learned from the signal, or from the state of a receiver, because the latter makes the former probable, and this in turn is so because an informational relation depends on a nomic one. As Dretske (1981) says, "The ultimate source of the intentionality inherent in the transmission and receipt of information is, of course, the *nomic regularities* on which the transmission of information depends. The transmission requires, not simply a set of de facto correlations, but a network of nomic dependencies between the condition at the source and the properties of the signal" (76-7). Indeed, Dretske traces the intensionality of reports of intentionality to the intensionality of natural laws: just as it can't be inferred that someone believes that *s* is *G*, given that the person believes that *s* is *F*, and that *F* and *G* are extensionally equivalent, so too it can't be

inferred that a signal indicates that *s* is *G*, given that the signal indicates that *s* is *F*, and that *F* and *G* are extensionally equivalent.[3]

With regard to the intensionality of natural laws, Dretske (1977) argues that this is due to their necessity, which is a property that requires an "ontological ascent," a Platonic rather than a Humean view of laws. On the Platonic view, a law of nature states that there is a contingent relation between *properties*, and this relation *necessitates* that the covered particulars be related in the corresponding way. On the Humean view, though, a law is about a relation that holds *universally* just between particulars. Dretske argues that universal regularities, as opposed to necessary ones, can be accidental rather than nomic, and thus that if there are laws of nature, they must be about relations between properties, not just particulars. Dretske (1988) recognizes the difference between strict and nonstrict laws, between those laws that don't have, and those that have exceptions, and he maintains that "for one thing to indicate something about another, the dependencies must be genuine" (57). According to Dretske, a law about a genuine nomic dependency can have exceptions when the dependency's instantiation rests on the meeting of special conditions. Although he doesn't use this term, this sort of nomic relation would seem to be expressed by a *ceteris paribus* (CP) law. As he says, a token doorbell may be wired in an unusual way such that its ringing indicates that the garage is opening, not that someone is at the door. This would be an example of a local, specially conditioned

[3] In taking intensionality to be the mark of intentionality, Dretske follows Chisholm (1957). Chisholm turns Brentano's claim about intentionality as the distinguishing feature of the mind into a linguistic claim about the characteristics of statements about propositional attitudes (see Brentano, 1874/1973). These statements operate in an intensional, as opposed to an extensional, context, which makes them opaque to substituting some of the terms used in the statements with extensionally equivalent terms; that is, substituting terms in this way affects the statement's truth value. The point is simply that a statement about someone's belief about dogs, say, is different from a statement about dogs, in that statements about dogs are all true no matter how the statements characterize dogs, as long as the properties of dogs support the characterizations. However, statements about a belief about dogs have to be supported not just by the mind-independent facts, but by the believer's perspective on dogs.

regularity. In cases of biological interest, "a sign—some internal indicator on which an animal relies to locate and identify, say, food—will have only this kind of local validity. It will, that is, be a reliable indicator only *in* the natural habitat or in conditions that approximate that habitat" (57). A biological dependency, then, stands between an accidental correlation and a physical dependency that has no exceptions. This must mean that a biological dependency is due to the development of a mechanism that works only under Normal conditions in accordance with a CP law.[4]

While information is found all across nature, semantic relations are relatively rare. Accordingly, Dretske (1988) builds on a distinction made in Grice (1957), between two kinds of meaning. There is the sense of "meaning," called "natural meaning," in which one thing indicates or signals another. For example, certain red spots on skin mean the measles, and smoke means fire. But a semantic relation that causes the behaviour of a semantic system can't be just this kind of meaning. A semantic symbol, as opposed to an informational signal, can misrepresent, which is to say that the symbol token can be caused by something that isn't in the symbol's extension. A semantic system can mistake one thing for another and so semantic content can cause the wrong sort of behaviour. But there can be no *misindication* without a violation of natural law. There can be a mistaken interpretation of a signal's source or ignorance about what a signal indicates, but a signal's indication of its source is an objective relation that depends on the way properties are nomically related. Grice called the type of meaning that can misrepresent "non-

[4] It may be that speaking of CP laws or of biological laws is misguided, that the notion of a nonstrict law is incoherent, and that special sciences are fundamentally different from physics and shouldn't be expected to discover natural laws. I'll continue to speak of CP laws, assuming the received view of special sciences and putting aside these worries, if only because I don't think the objections I'll raise turn on these worries. I'll return to this issue, though, in section 5.4.

natural meaning," and Dretske's project is to naturalize this so-called non-natural meaning.

He does this by building non-natural meaning out of a certain function of natural meaning. Thus, he adds a teleological theory of content determinacy to an informational theory of the metaphysical identity of semantic relations. Reference to the purpose of certain things allows him to distinguish between what he calls three types of information-based representation and between two ways of organizing a representational system. In each case, a representation is an indicator with a purposive function. Two of the kinds of representation have content that derives from other representations, and so they don't have a basic kind of content. Dretske calls these dependent representations *conventional*, and it's worth discussing them here to contrast them with what Dretske regards as the basic kind of representation. First, then, a conventional representation can serve as an indicator with a function chosen *arbitrarily* by a system that already employs representations, as opposed to being selected by a bottom-up process such as natural selection or a certain kind of learning. Call this established representational system that chooses how to organize another system a *designer*. This designer's choice is arbitrary because the function isn't dependent on what the system carrying the representation can do. For example, pieces of popcorn can indicate basketball players in a tabletop game of basketball.

Second, a conventional representation can indicate something by a designer's *nonarbitrary* choice of function. Here, the designer's choice isn't arbitrary because the possible functions the indicator can have are taken to depend on the designed system's own capacities. For example, a fuel gauge can indicate the amount of fuel in a tank,

because the gauge and the tank are connected by a mechanism provided with just this capacity of transmitting the information. Both kinds of conventional representations are the result of design, but the difference is that both the function and the physical capacity of the second type of conventional representation is designed, whereas only the function of the first type is designed.[5]

The third kind of representation, however, is *natural* rather than conventional in that this third sort of indicator has a function not selected by a designer. Whereas conventional representations depend on other representations, a natural representation is configured without the help of other representations and thus has original and objective rather than derived and interpreted content. Conventional representations are assumed to derive ultimately from natural ones, and so Dretske wants to explain natural representations. Explaining natural representations requires explaining semantic properties without referring to symbols whose content isn't also explained by the theory. Dretske considers two ways of producing the mechanism that implements natural representations: natural selection, on the one hand, and conditioning or a type of learning process, on the other. Between the two, he says, only learning can account for a semantic system's behaviour.[6]

The key point about learning is supposed to be that learning is a process of self-organization. If the result of learning is that an indicator is given a role to play in the organism, such as control over a movement the organism can make, the indicator's ability

---

[5] I'll return to this question of arbitrariness in the intrinsic properties of certain symbols, in sections 6.2 and 6.3.

[6] Dretske (1991) says that a genetically determined, involuntary effect will "persist whatever its internal cause happens to mean (if anything) about environment conditions...What explains why the *X* in *this* animal causes *A*, then, is not *its* meaning, but the meaning of corresponding *X*s in remote ancestors" (206-7). It turns out, though, as I argue in sections 3.5 and 3.6, that Dretske's theory of learning as the basis of semantic systems faces a similar objection.

to play this role is due to the organism's structuring of itself by making the indicator an internal cause of the movement. It's important for Dretske that the content of some such internal condition as a brain state be the starting point of a causal explanation, given what Dretske takes the *explanandum* to be, which is an organism's *behaviour*. Behaviour isn't just a bodily *movement* or an effect of some effort made, such as an arm's raising; instead, behaviour is an event's being brought about by a condition internal to the organism. Movements in themselves are explainable in mechanistic or physical, and thus nonsemantic terms, by appealing to natural laws that abstract from the distinction between a system's interior and exterior. Thus, the raising of an arm can be explained as an effect not of an internal condition, as such, but of blood flow, muscle-flexing, and so forth, making reference exclusively to processes that may or may not happen within the organism. For example, the blood-flow may occur in the arm that moves another organism's arm. Behaviour, on the other hand, is in part movement that is characteristic of a type of organism, because the movement is typically connected to some internal condition of the moving organism. Instead of just a bodily movement M, such as the raising of an arm, there is $C \rightarrow M$, an internal condition's causing of the movement. For example, whereas the raising of an arm, as such, could be just a movement, a salute is a type of behaviour. A salute isn't just an arm's being physically moved, but is, say, a soldier's raising of his arm or more specifically the arm's being raised by a condition or structure internal to the soldier.

Dretske calls behaviour a *process*, to distinguish behaviour from the isolated event that is, say, a movement caused not by internal conditions as such, but just by some more general mechanism. A behavioural process is C's causing of M, and thus includes

an internal condition and a movement, or more generally, a bodily change, as the behaviour's proper parts. For this reason, *C* which causes *M* can't also *cause* the behaviour which is *C*'s causing of *M*; however, something about *C* can *explain* the behavioural process of which *C* is a part. A movement, though, may happen *to* the organism even if the mechanism producing the movement happens to be within the organism. For example, the growing of hair is a movement of part of an organism, but not something the organism does. As Dretske (1988) says, "I *get* rashes, I don't *do* them" (6). This suggests that behaviour is *voluntary* bodily movement, but Dretske wants to distinguish between behaviour and action, the latter being a type of behaviour. This is because biologists and other scientists speak of the behaviour of nonhuman species, and Dretske wants to capture the general scientific sense of "behaviour." I think, then, "behaviour" is used here in a loosely anthropocentric way. Although there is such a thing as rat behaviour and not just rat movement, what this behaviour has in common with human action is that some of the rat's movements are the rat's own. The rat may not voluntarily produce these movements, but the movements are typical *of rats*; they don't happen to rats, but are the characteristic outcomes of something within the rat, not of a mechanism which could be relocated or duplicated outside the rat to produce the same effect. Roughly speaking, behaviour is an organism's own movement; the movement flows from the organism, and this flowing is a behavioural process.[7]

---

[7] This issue of natural behaviour as something's own will arise again in my criticism of Millikan's theory of functions, in section 4.4. According to Dretske (1988), "It may be arbitrary whether something should be classified as behavior or not," but he maintains that it's "not at all arbitrary that, once so classified, it is a causal process of the sort" Dretske describes (25). He points out that there is "no hard and fast line separating internal from external causes." This is because internal and external causes themselves have causes (22). For example, as long as a mechanism, or evolved or constructed system, is sufficiently complex, the mechanism may be regarded as having its own internal region, in which case the mechanism will behave rather than just move: the mechanism's effects will flow characteristically from the mechanism itself. But the internal/external distinction is reserved mainly for the relation between the whole organism

Behaviour has a different kind of explanation than does bodily movement, the former being a process with an *internal* point of origin and the latter being an event in a more generally characterized causal chain. The movement of a rat's paw is explained entirely in terms of whatever mechanism causes the movement. This mechanism may be internal or external to the rat; for example, someone may grasp the paw and move the rat's arm for the rat. Were the movement the result of a biological mechanism, the movement would happen because of a special nomic dependency, since the movement would depend on background conditions without which the mechanism could not have evolved in the first place. In any case, a bodily movement has what Dretske calls a *triggering cause* that explains why a token internal condition *C* causes a token movement *M* at a particular time. For example, the triggering cause of a rat's movement might be a stimulus which causes a defense mechanism in the rat to react: the movement would happen when it does because of the timing of the stimulus which triggers the mechanism that causes the movement.

Behaviour may also have what Dretske calls a *structuring cause*. The structuring cause isn't the cause of why the movement *M* occurs or of why *M* or whatever causes *M* happens at a certain time, but is, rather, what accounts for why it's *M* that has one cause rather than another. If *M* is a rat's set of running movements, and *M* is caused by *C*, where *C* is the rat's defense mechanism, the structuring cause is some background condition that explains why *M* is nomically dependent on *C*. In general, a structuring cause is, as Dretske (1988) says, "the cause of one thing's causing another." For example,

---

and its environment, and for relations between organs or other systems within the organism. So the difference between bodily movements and behaviour is a pragmatic one, depending on whether the movement's cause is internal to something sufficiently complex to warrant an explanation that draws the internal/external distinction.

"One puts yeast in dough so that the bread will rise when put into the oven—so that the heat of the oven will cause the bread to rise." In this case, "an event of type *C* causes or brings about an event of type *E only in a certain restricted or special set of conditions*," which can be called "background conditions." If these conditions don't obtain, *C* will not cause *E*. (39). Again, were a thermostat wired to open the garage door whenever the room reaches a certain temperature, the triggering cause of the garage door's opening would be the reason the garage door opens when it does: the room temperature reaches a certain point at a certain time, triggering the mechanism. The structuring cause, though, is the designer's decision to wire the thermostat to the garage door, and this cause explains why the thermostat is causing the garage door to open *at all* (42). In short, a structuring cause is the cause of a nomic dependency. In the case of a mechanistic, as opposed to a strictly physical, unconditional effect, the structuring cause is something about the process that configures the mechanism or that connects one mechanism to another, in the first place. The triggering cause is some temporally specific condition that accounts for when a mechanism works. The structuring cause is that which sets up the mechanism so that the mechanism works in a certain way when triggered.

In the case of an organism's behaviour, the structuring cause is part of what configures the whole organism, so that the bodily movement in which the behaviour terminates is the organism's own typical movement rather than that just of some part of the organism. In the case of most organisms, this structuring cause is genetically determined and therefore naturally selected. But learning is another way of configuring an organism; indeed, learning is a way for an organism to configure itself so that the organism can behave in the most characteristic, voluntary way, by rationally acting. This

sort of behaviour is structured by reasons, by beliefs and desires, that gain control over the organism's capacity to interact with external conditions. This control doesn't happen by accident or by physical law, but by the organism's having been organized by a learning process, which is the process by which certain *C*s become beliefs or desires.[8]

3.3 Dretske's Theory of Content Determinacy

Having laid out some of the background, I turn now to the details of Dretske's theory of how the content of a symbol, used by a semantic system, is determined by a learning process. Dretske's view is that the semantic content of a belief or of a desire is the role the internal condition is supposed to play in an organism, because of what the condition did for the organism during a training period. An organism, *S*, is assumed to be receptive to an external condition, *R*. The receptivity is a rudimentary form of motivation that makes *R* rewarding and thus relevant to *S*, before *S*'s movements are modified by their rewarding consequences.[9] But the use of *R* depends on the intermediary step of coordinating with another external condition, *F*, by *S*'s producing bodily movement, *M*. To get *R*, *S* must deal with *F* by *M*. A necessary condition of *S*'s dealing with *F* is *S*'s use of an indicator of *F*; otherwise, *S* will lack information about *F* and won't be able to

---

[8] Dretske (1991) adds that there is a clue in the relevance of learning to purposive behaviour. "Since minds conveniently appear on the evolutionary and developmental scene when, and only when, learning occurs, when there appears the kind of behavior (voluntary or purposive behavior) that minds are invoked to explain, the suspicion is irresistible that the elements of these explanations—the beliefs and desires we invoke to explain voluntary behavior—have their origin in precisely those transactions (the learning experiences) that gives rise to the behavior needing explanation" (202).

[9] Dretske (1988): What shouldn't be ignored in an account of conditioning is "a qualification having to do with the *receptivity* of the organism. Rewards tend to encourage reproduction of rewarded events only when the organism is in a certain internal condition." The effectiveness of consequences of behaviour "in modifying behavior depends, critically, on the receptiveness of the system relative to the consequences in question" (110).

coordinate its movements to use *R*. *S* has an internal condition, *C*, which indicates *F*. And so *S* will learn to use *C* and *M*, so that *S* can use *F* as an additional means to using *R*. *S* will have both the motivation and the means to satisfy its receptivity to *R*, and the use of *R* will reinforce some of *S*'s rudimentary abilities, making them functional.

The internal indicator of *F* becomes a belief *B* with *F* as its represented, semantic content as soon as *C* is, as Dretske (1988) says, "recruited" as a cause of *M* (98).[10] Moreover, the initial receptivity becomes a desire *D* with *R* as its semantic content, as soon as the receptivity is likewise recruited as a cause of *M*. *S* performs *M* because *S* is motivated to use *R*, and the use of *R* reinforces whatever means *S* takes to succeed, including the use of *C* and the motivation to respond to external conditions in such a way that *S* can use *R*. And *M*'s causal relation to the conditions within *S* becomes a behaviour as soon as *S* is configured (while being conditioned by *S*'s pursuit of *R*) in such a way that *S*'s performance of *M* becomes typical of *S*, or becomes *S*'s own *M*. Dretske appeals to Thorndike's Law of Effect, according to which successful behaviour tends to be repeated since the reward or reinforcing condition *R* increases the probability that the movement that succeeds in using *R* will occur again in the same circumstances (99). This law assumes *R* is relevant to *S* because *S* is receptive to *R*, but the receptivity becomes a desire *for R* only when the receptivity is *supposed* to cause *M* to use *R*; the useful internal conditions of *S* must be given the function of helping *S* obtain *R*, and their functions are given in the conditioning process by which the causal connections are established and reinforced by *S*'s use of *R*. As Dretske (1990) says, "By understanding that both belief

[10] As Dretske (1988) says, "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" (84).

and desire derive their content—what is believed and what is 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" (834).[11]

Although there is no misindication, there is misrepresentation, and this is explained by Dretske in functional terms. *C*'s function is to do what *C* did in acquiring control over *M*, while *S* was being conditioned. For *C* to carry out the function of indicating *F*, circumstances have to be similar to those that occurred during conditioning. If circumstances differ, *C* may malfunction and represent *G* as *F*, as though *C* were indicating *F* whereas only *G* is indicated. In short, the internal condition's work during *S*'s conditioning, which accounts for the condition's function and control over some movement, is the standard by which the condition's later work may be judged. Any performance that falls short of indicating *F* or of motivating *S* in the ways that rewarded *S* while *S* was being configured by its formative interaction with its environment, counts as a malfunction, an effect that differs from the internal condition's optimal work, or its function.[12]

---

[11] With regard to desires, which have motivational rather than representational semantic content, Dretske's theory is meant to apply to what he calls *pure* desires, which are those whose content isn't determined in part by the content of beliefs or of more basic desires. For example, were someone to want to sit on a certain chair, believing the object to be a chair, the person would want to sit on the object even if the object turned out not to be a chair; the person's desire would be guided by a belief about the object, and so wouldn't be a pure desire.

[12] There is a question here of the determinacy of a detector's function. Suppose that a detector has been recruited to indicate dogs, but that on some occasions the detector indicates foxes. The question is whether the detector functions or malfunctions on these occasions. Were the detector's function only to indicate some proximal properties shared by dogs and foxes, there would be no malfunction here. So is the detector's function to indicate proximal or distal types?

Dretske would say that the function is determined by what a token of the type of indicator did when the organism was trained to satisfy its receptivity towards some reinforcing condition. As for the question of whether a detector indicates proximal or distal properties, Dretske (1986) says that a sign can have highly disjunctive informational content but a single semantic content. The key point is that learning combines information about multiple proximal sources into information about the common distal source. While the information about the proximal causes is "time-variant," the cognitive mechanism that unifies this information, through learning, has a "time-invariant function," or the function of indicating the

So *S* is a semantic system, because the structuring cause of $C \rightarrow M$ is an external relation between the internal condition and some external condition, and the internal condition acquires the function of entering into this external relation. *C* is given causal control over the performance of *M*, by the structuring of *S*, because *C* indicates *F*; in other words, *C*'s cognitive role within *S* is granted, as it were, because of *C*'s ability to track *F*. The causal relation between *C* and *M* is established by *C*'s indicating *F*, which is to say that the reason a certain internal condition *C* is given control over a certain movement *M* is because of *C*'s informational content. *C*'s indicating of *F* is the structuring cause of some $C \rightarrow M$; once *S* is conditioned by sufficient reinforcement, and *S* behaves by performing *M*, *C*'s ability to indicate *F* becomes *C*'s function, making *C* a belief, *B*. The difference between a belief and an indication of *F* is that a belief is *supposed* to indicate *F*. *C*'s function is to indicate *F*, because *C*'s ability to do so is why *C* is given causal control over *M*.

The same can be said about *S*'s receptivity to *R* that is conditioned to become a desire, *D*, except that instead of having the function of indicating *R*, the receptivity's function is to motivate *S* to use *R*; the receptive internal condition makes *R* relevant to *S* so that *S* will track *F* and perform *M* to use *R*.[13] Dretske (1988) argues that desires have the same intentional properties as beliefs. For example, an ascription of a desire, like one of a belief, is referentially opaque in Chisholm's sense, and a desire, like a belief, has a satisfaction condition that may not be met. To use one of Dretske's examples, a rabbit trained to lick a spout for water may lick the spout, and yet there may be no water bottle

---

property that is *always* the distal source of the proximal information, no matter which route this information takes to the receiver (170).

[13] I think there's a problem with Dretske's distinction between receptivity and desire, and I'll explore this in some detail in sections 3.7 and 3.8.

presently attached to the spout. Note that at some point the rabbit must have gotten water by licking the spout; otherwise, its receptivity to water could not have been reinforced by licking the spout.

Some of *S*'s internal conditions become wired in the first place to *S*'s responses to external conditions, because the internal conditions already have rudimentary informational or motivational contents, and so these contents explain why *S* is equipped, or why *S* equips itself, with certain mechanisms for interacting with *S*'s environment. *S*'s behaviour is structurally caused by *S*'s relations to external conditions in so far as these relations account for why *S* behaves in certain ways.[14] The rudimentary content of the internal conditions that help configure *S* to cause *M* accounts for the internal wiring needed for *S*'s behaviour to occur, since this content is the wiring's structuring cause. For example, *S* may have an internal structure, *C*, that detects water, *S* may be receptive to drinking water, and so *S* can learn that when *S* moves in a certain way once *C* detects water, *S* is relieved of its thirst. In this way, *C* acquires control over the movement, and the structuring cause of this internal wiring is *C*'s relation to water. *S* is a semantic system, because the extrinsic property of an internal condition—its information or its receptivity—causally explains *S*'s behaviour, as a structuring cause of this behaviour, that is, of why *S* is configured such that a certain internal condition is given control over a certain movement of *S*'s. One internal condition rather than another causes one movement rather than another, because this internal condition was useful during a formative period for the organism, and was useful because of this condition's relation to

[14] As Dretske (1990) says, "By understanding that both belief and desire derive their content—what is believed and what is 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" (834).

an external condition. So reference to this external relation is needed to explain learned behaviour.

3.4 The Replacement Argument

I want to raise two criticisms of Dretske's theory, but first I want to consider a well-known general criticism of etiological theories of content determinacy, which take the content of a token symbol to depend on a relation between an *earlier* token symbol and an external object. The criticism takes the form of a thought experiment and appeals to the critic's intuition that intrinsic properties rather than historical ones have causal powers.[15] Suppose some object were destroyed but recreated molecule-for-molecule. The two objects would then have identical intrinsic properties but different causal histories. The intuition is supposed to be that, despite their different histories the indiscernible twins would have exactly the same causal powers. Following Stich, Dretske (1991) calls this the Replacement Argument (RA). One object is replaced with another, the two are supposed to be identical except for their histories, and yet the historical differences are supposed to make no difference to their capacities to behave or to their actual behaviour. When applied to Dretske's theory, then, the point of the objection would be that his explanation of how there could be a semantic system fails, because a reason, such as a belief or a desire, has no causal power due to its having a structuring cause lying in its past. Causal power is derived solely from something's intrinsic properties, not from any

[15] For some versions of the criticism, see Boorse (1976), Stich (1983), and Davidson (1987).

extrinsic property, such as the etiological property of the system's having been configured by a certain learning process.

Dretske's response to RA is just that the twins might have different functions, given precisely their different histories (209). The conclusion that the twins would have the same causal capacities and would behave in the same way leaves aside the question of what the objects are *supposed* to do. The twins might actually behave in identical ways even though the two are judged by different standards. This is the point about their purposive functions. If certain kinds of functions are historically determined, by natural selection or by conditioning, two objects that are identical with respect to all of their properties that could be measured at a particular time might still have different functions. Dretske considers two plants that have the same intrinsic properties and that behave in the same way, say, by changing their colour at exactly the same times, but that have these properties for different reasons. The one plant changes colour for the purpose of attracting a type of pollinator, whereas the other changes colour to repel certain insects. The plants may be assumed to have grown in different environments, making the functions historically determined; for example, the latter plant may be assumed to have come from Mars.

But this point about functions doesn't yet address the objection, since the intuition is supposed to be that these functions would be causally irrelevant. Dretske doesn't go on to make what I think is the crucial point, that the functional differences would be seen to affect the plants' actual behaviour as long as the thought experiment's scope were widened to include the circumstances in which the different purposes of each plant's behaviour, determined, say, by the plants' formative interactions with different

environments, make a causal difference to the plants' later behaviour. It's easy to see how this could happen were the two objects to retain their different histories at each later moment. Suppose, then, that one person is destroyed but miraculously recreated by a strike of lightning, so that the two persons have the same intrinsic properties.[16] And suppose that they each have present memories of their different pasts, so that only the later person remembers originating from a lightning strike. Then, of course, their behaviour would actually differ; for example, the one with the miraculous origin might believe herself divine and start a religion about her unusual past.

The retaining of different memories changes the thought experiment, though, since memories would have to be neurally stored and this would make for an intrinsic rather than just an historical difference between the twins. So suppose they have exactly the same memories, after all; suppose, for example, that the later person shares the earlier person's memories so that they each believe they have ordinary, nonmiraculous origins. But now the thought experiment's scope can be broadened to cover the circumstances that would test whether historical differences can make a causal difference to otherwise identical objects and to their behaviour. There would be evidence of the miraculous origin of the later person, and were she to acquire this evidence, the difference between her past and that of her twin would catch up to her, as it were: evidence of her distinct past would endure and await discovery whereupon this particular history would affect the later person's behaviour. Had the earlier twin continued to live, and the later one not to have been created, the earlier twin would have behaved differently from the other twin, since the earlier one wouldn't have had a miraculous origin about which she might later

[16] This is roughly the scenario given in Davidson's swampman thought experiment, which adapts Putnam's Twin Earth arguments. See Davidson (1987) and Putnam (1975).

have come to learn. Discovering her miraculous origin would cause the later twin to behave in a way in which the earlier twin likely would not have behaved.

Suppose there were no such evidence of the resurrection because, after all, the lightning strike is supposed to be miraculous. In this case, the thought experiment would lose its relevance, since the twins wouldn't have different *causal histories*, or past interactions with a *natural* environment. The miraculous lightning strike would stand outside of the natural world in which the resurrection could causally affect the one twin and make for a part of her causal history. In other words, the later twin wouldn't originate from a *lightning strike* after all, since lightning isn't miraculous. For an object's origin necessarily to leave no evidence, no way of learning about this past event, the origin would have to be non-natural and so talk of such an "event" in the thought experiment would be obscure. Of course, the two plants would lack both memory and the ability to discover the differences in their pasts. But this means just that the causal powers of the historical properties might be more indirect. Evidence of the differences in the plants' causal histories, such as their different places of origin, could await discovery for botanists, for example, and the plants' different functions could thus cause botanists to treat the plants differently, thus indirectly changing the plants' capacities and actual behaviour.[17]

---

[17] Dennett (2003) points out that there are practically inert historical facts, such as whether the gold in Dennett's teeth once belonged to Julius Caesar (68-9). No one will ever know what the fact of this matter is, because of a lack of evidence due to an historical lack of interest in the gold's source. My claim that any event in a causal history should leave some trace is compatible with Dennett's claim, since his point is an epistemic one about the inability to justify belief one way or the other about some fact of the past, given a lack of evidence. This is separate from the metaphysical point about the transmission of information from one type of natural event to another. Even if some historical facts are actually inert, this doesn't mean they're necessarily so. Our actual justifiable beliefs about the past may be forever limited, given earlier choices about what kinds of evidence were collected, but this doesn't mean there is any historical fact that *necessarily* leaves no trace. So the thought experiment can simply be extended to show what the causal impact would be *had* evidence of some previous event been collected.

Moreover, the natural copying of an original individual is affected by the original's history, since the copying requires a causal connection between the original and the copy, and so this history has an impact on the copy. That is, strictly speaking, the copy's history includes not just the period up to the point at which the individual is produced as a copy, but the original individual's own history. The only way for the twins to have separate histories, to test whether these histories have different effects, given the same intrinsic properties of the twins, is for the twins to be made without any causal connection between them. Thus, the problematic notion of creation by a miraculous lightning strike.

For these reasons, I don't think the thought experiment given in RA shows that historically determined types, such as learned purposive functions, lack causal power.

3.5 The Problem of Local Potency

I want to turn now to what has been called the problem of local potency, for Dretske's theory.[18] The problem is that if a semantic relation is causally relevant as the structuring cause of a behaviour, and this structuring cause lies in the past when the organism is conditioned and the internal structure receives its function, it's not the content of the internal structure at the later time that is causally relevant, but the content of an earlier token of this structure. This seems to give causal relevance to the content only of the earliest tokens of *C*, making the content of later tokens inefficacious. The information carried by an internal condition that has acquired—from conditioning—a

[18] This problem is raised by Dennett (1991b), Cummins (1991), Horgan (1991), and Kim (1991).

function and thus control over some movement of the organism, can be a cause only of movement as such, not of behaviour, and thus *this* information can't be a structuring cause. The cause of behaviour is the cause of some later *C*'s coming to have a functional role in the organism, and this earlier cause, the structuring cause, must lie in the functional *C*'s past, or in the period in which the organism was still being conditioned.[19] But if this is so, not enough of *S*'s internal conditions seem to have causal power because of their extrinsic relations, and so *S* isn't a semantic system. What Dretske wants to show is that when, for example, a person salutes, the arm's movement is caused by an internal condition in so far as this condition is a belief with the content that a superior officer is present. But Dretske's etiological, backward-looking account seems to imply that the only internal condition that has causal power in virtue of the condition's relation to something external to *S* is a previous token internal condition. This previous token could have been a structuring cause only in so far as the token still lacked a functional role in *S*. So a *C* that has a representational function can't have causal power in virtue of its semantic property, because this *C* can't be a structuring cause. This is because this *C* is already structured, and must be so to have a semantic property, since on Dretske's view a semantic property is a functional one and thus the *result* of some structuring.[20]

Dretske (1991) addresses this problem, by endorsing the suggestion in Kim (1991), that tokens and types need to be more carefully distinguished in Dretske's

---

[19] Instead of speaking, say, of C_1 and C_2, when discussing the problem of local potency, I'm going to speak somewhat more loosely of earlier and later *C*s. The early or earlier *C*s are always the internal conditions whose indication of *F* or whose receptivity to *R* cause *S* to use the later *C*s. So the point of division between earlier and later *C*s is *S*'s recruitment of them as causes of *M*, which gives them their function; the earlier *C*s aren't yet recruited and don't yet have a function.

[20] Note that while in this section I focus on the causal power of later *C*s that are supposed to indicate *F*, that is, on the efficacy of beliefs, Dretske's accounts of the contents of beliefs and desires are parallel, so the same objection about local potency applies to his account of the causal power of *D*. In sections 3.7 and 3.8, I'll return to this point about the similarity of his accounts of belief and desire.

account. Dretske thinks that a more careful formulation of his view of representational functions clears away some confusion. So here is the adjusted account of how later tokens of an internal condition acquire representational content. During the period in which an organism *S* is conditioned by an external condition *R* that reinforces some tendencies in *S*, an internal structure *type N* is recruited as a cause of movement type *M*, because of the success some early token of *N*, an early *C*, had in indicating *F*, which helped *S* use *R*.[21] The early *C* indicates *F* in so far as *C* instantiates *N*, and *N* nomically depends on *F*, although the law here is CP. Now, a later *C* (after the conditioning of *S*) represents *F* even if this *C* doesn't indicate *F*, because later *C*s have only the *function* of indicating *F*. Later *C*s have their function as tokens of *N*, and *N* not only nomically depends on *F*, but has the function of doing so, because *N* was given the job of doing so in *S* due to some earlier token of *N* in *S* which did indicate *F*. On the one hand, "*N* is the type of physical condition whose correlation with condition (type) *F* makes tokens of *N* indicate (carry the information that) *F* (when they do so)" (214). On the other hand, later *C*s "have the function of indicating *F* because they are of a type (*N*) that has this function" (215).

So *N* is nomically related to *F*, which is the nomological source of any *C*'s indication of *F*. But in *S*, *N* is also *supposed* to indicate *F*: in *S*, all later *C*s are supposed to indicate *F*, because some early *C* helps to configure *S*, by indicating *F* and acquiring motor control over *M* as a result of that early *C*'s informational content. Once *C*s gain control over *M*, so that *S* is configured to behave in this way, all later *C*s in *S* have that early *C*, which indicated *F*, as their standard. On this account, Dretske says, all later *C*s

[21] Depending on the context, I'll speak of *N* as a *type*, that is, as a set of objects, or as the *property* of being instances of a type.

are "locally potent" because, "given the nature of meaning, local meanings, the fact that *this C* means *F*, is, in reality, a fact about the kind of information that restructured control circuits so as to give *C* a voice in determining [behavioural] output" (215). So "Present meanings explain present behavior, but only because both meaning and behavior (at least the structuring explanations of behavior) are backward looking phenomena" (216).

Now, Dretske's more careful formulation of his view seems only to restate the point that later *C*s aren't locally potent in virtue of the relations they bear to external conditions. For one thing, the notion of *locally* potent content becomes empty if the having of content is a backward-looking phenomenon. The objection is that the content of all of a semantic system's reasons should be causally relevant to *S*'s behaviour, but that Dretske's theory implies that the content of later *C*s, as beliefs, for example, is causally irrelevant. For example, having learned that information about a fridge is needed to satisfy a desire for a beer, *S* has an internal condition *C* whose function is to indicate the fridge. This token *C* causes *M*, a movement that opens the fridge door, and the question is whether an extrinsic, semantic property of this *C* is explanatorily relevant to why this *C* causes *M*. Dretske's response is that later *Cs* with semantic content are causally relevant, because their efficacy is found not in their actual indication of *F*—some later *C*s may not indicate *F*—but in their function of indicating *F*. They have this function in virtue of their instantiating *N*, and *N* is something like the capacity all *C*s have of doing what some early *C* does in *S*, which is to indicate *F*. A later *C*'s representational content is backward-looking, since this content is a structuring cause of all later *C*s' causing of *M* in *S*. *But a backward-looking property isn't a local property.* In saying that representational content is efficacious in *S* as a structuring cause, what is said is that the explanation of any later

C's causing of *M* must refer to some early *C*'s indicating of *F*. This seems to make the content of later *C*s potent only because of their association with earlier *C*s.

Perhaps what makes the content of later *C*s locally potent is their own representational function. All later *C*s have this function by their instantiation of *N*, and representational content, on Dretske's view, is the function to serve as an indicator of some external condition. But a later *C*'s own representational function can't be explanatorily relevant to this or to a later *C*'s causing of *M*, because what makes content explanatorily relevant, on Dretske's view, is content's serving as a structuring cause. Any *C* that is already wired to *M* because of this *C*'s representational function, must have a structuring cause that lies in this *C*'s past, namely in the content of some earlier *C*. Representational content doesn't structure itself; rather, representational content is structured by informational content, and it's informational content that is explanatorily relevant as the structuring cause of some configured *C*'s causing of *M*.

However, as Dretske (1991) says, the "non-natural", representational content of a later *C* just *is* "whatever natural meaning (information) in past *C*s explains the present causal arrangements" (216). This identity of representational content with earlier informational content is supposed to justify calling representational content locally potent. The local causal power of a later *C*'s representational content is a backward-looking property, namely the informational property of an earlier *C* which indicated *F* prior to *C*'s being recruited as a cause of *M*. But this identity could show just as well that the causally relevant token *C*s would have only informational, not representational content, since what explains the wiring of *C* to *M* is only some *C*'s actual indication of *F*. *C* acquires the function of indicating *F*, and thus the representational content of *F*, only

after the wiring has occurred. The price, as it were, of having representational content, of having the function of indicating *F*, is that the function must already have been determined by something that lacks this function.

So if content is explanatorily relevant to behaviour only as a structuring cause, and structuring causes can determine purposive functions, representational content must be explanatorily irrelevant to behaviour. Saying that this content just *is* earlier informational content is to say that representational content isn't itself *locally* potent. Representational content isn't explanatorily relevant as a structuring cause of its own causing of *M* or of any later *C*'s causing of *M*; indeed, the structuring cause of these causal relations is the nonlocal informational content of an earlier *C*. The early and the later *C*s must be different tokens since they have different properties: the early ones lack, while the later ones have, causal control over *M*, and the early ones lack, while the later ones have, the function of indicating *F* in *S*. Saying that a later *C*'s content is locally potent *as* the content of an earlier *C* is like saying that the function of a lion's roar is locally potent *as* the potency of the roar of an earlier lion, of the one whose proliferation of genes eventually produced the later lion and its roar. Presumably, local potency is supposed to be indexed to a token, specifically to a token that occurs at a particular time. If the property of a later token were just the property of an earlier token, but not the other way around, the later token would seem to have no property of its own. In this case, the later token wouldn't have causal power in virtue of that property which it lacks.[22]

[22] Dretske (1990) says, revealingly, that "anything explained by the fact that earlier tokens indicated *F* will be explained (albeit redundantly and indirectly) by the fact that current tokens mean *F*. To explain behavior by current meaning is just to explain what indicational facts (about earlier tokens) *were* relevant in recruiting the current belief (the current token of this type) for this kind of causal service" (831). The explanatory relevance of later *C*s is redundant and indirect, because the working part of a causal explanation of *C*'s causing of *M*, that refers to the later *C*s, is the implicit reference to earlier *C*s. Still, the

3.6 Representational Functions

The confusion here seems to be about whether a so-called backward-looking property is different from a property that exists only in the past, relative to some token *C*. A backward-looking property is supposed to be just an etiologically-determined function. This leaves the question of what Dretske means by "function", so I want to turn now to this question. Responding to a criticism by Millikan, Dretske (1990) says that his theory of content requires only the commonsense notion of function, as something that *X* is supposed to do. Dretske accepts, then, that his theory appeals to the normative aspect of purposive functions.[23] But Dretske's naturalistic theory of content can't use this commonsense notion, without naturalizing the normative aspect. He can do this in a way that is consistent with his backward-looking use of the notion of purposive functions, by explaining these functions in *etiological* terms. The key point of an etiological account is that the function of *X* is *X*'s effect that accounts for the existence or positioning of *X*. For example, although a nose has the capacity to hold up glasses, this isn't the nose's function because the nose isn't there, on the face, to hold up glasses; rather, the nose is there for the person to breathe, roughly speaking, and this capacity can be given an evolutionary explanation. A conditioned semantic function would be the carrying of information about an external condition, which accounts for the wiring of $C \rightarrow M$.[24]

---

extrinsic, semantic property that has causal power is found only in the early *C* that uses this power, as it were, to compel *S* to recruit *C*s as a means of achieving *S*'s goal of using *R*.

[23] In Dretske's words, "Why should I have to give a definition of this word ["function"]? Why isn't it enough if what I say is true on (at least) one of the commonly accepted (dictionary) senses of the word?...Since what I'm trying to analyze is the power of representation (including the capacity to get things wrong), all I need is some process in which an indicator acquires a special status (call it what you will), a status in which there is, among the many things it indicates, some one thing it is now *supposed* to indicate" (824).

[24] In the next chapter I'll discuss etiological accounts of purposive functions in more detail.

Later *C*s may or may not indicate *F*, but in so far as they all have the function of indicating *F*, and this function is just the explanatorily relevant relation of the indicating of *F*, the question of whether the function is *locally* potent is the question of *when* the function is supposed to be found. The indicating of *F* which is the later *C*'s function can't be a relation into which all later *C*s enter, since the point of their having only the *function* of indicating *F* is supposed to account for their ability to misrepresent *F*, which requires that these later *C*s may not indicate *F*. Moreover, those later *C*s that do indicate *F*, under the right conditions, don't indicate *F* in an explanatorily relevant way, since by the time these *C*s indicate *F*, *S* is already configured such that *C*s cause *M*. These later *C*s are supposed to have the function of indicating *F*, but were this function just the explanatorily relevant indication of *F*, namely the *token* relation that configures *S*, only the *earlier C*s would have the function, the ones that must indicate *F* for *S* to have its configuration and thus for *C*s to be wired to *M*. Assuming a function is a token effect or relation that explains why later tokens have the same effect or enter into the same relation, the indicating of *F* which is the function of *C* would be just the informational relation that explains the wiring of later *C*s to *M*. Learning in Dretske's sense offers a backward-looking explanation, and so the explanatorily relevant information is carried by earlier *C*s, which makes their information a structuring cause of *S*'s behaviour. On the present interpretation, though, this structuring cause, which is a feature of an early *C* in *S*, *is* the function, since this early *C* accounts for the later configuration of *S*, the "being there" of *C*'s connection to *M*, and since the function can't be a relation into which all *C*s enter, the function is identified with the relation's early, explanatorily relevant instantiation. But this means that only the early *C*s have the function of indicating *F*.

Having this function is having representational content, on Dretske's view, so only the early *C*s represent *F*.

But on Dretske's view this leads to a contradiction, of course, since the function of indicating *F* is supposed to be the *product* of earlier *C*s' indicating of *F*, and so the function is supposed to be had only by the *later* *C*s in *S*. If the function of *C*s is the explanatorily relevant *C*'s indicating of *F*, the function is just the earlier *C*'s indicating of *F*. In this case, the representational function would be *in* a conditioned *S*'s past, rather than just backward-*looking*, and the later *C*s would have no such function, since they may or may not indicate *F* and their content, in any case, isn't explanatorily relevant in the sense of having an impact on *S*'s configuration. However, the representational function can't be just the early instantiation of the relation of *C*'s indicating of *F*, which sets the standard, since then *C*s that are configured to cause *M* wouldn't have this function or any representational content of their own. A representational function can't be both the cause and the product of some configuration, but on an etiological interpretation of what Dretske means by his talk of a backward-looking representational function, the contradiction seems to follow.

It's no help saying that the semantic property later *C*s themselves have is the *capacity* to indicate *F*, given their similarity to early *C*s, or the probability of indicating *F* under certain conditions, rather than the carrying of information about *F*. The law that *C* potentially indicates *F* would have to be CP, since the generalization would be that under certain conditions *C would* indicate *F*. But, first, were the capacity the *function* of later *C*s, a *C* which misrepresents *G* as *F* would still fulfill its function as long as *C* could have indicated *F* had there not been, say, some defect in the mechanism in *S* that detects *F*. *C*s

that don't indicate *F* might still have the capacity of indicating *F*, and so *C*s that misrepresent *F* might still be functional. Second, the early *C*s, which have the capacity to indicate *F*, would thereby have the function of indicating *F*, which contradicts Dretske's account. Third, *C*'s capacity to indicate *F* is presumably an intrinsic property of *C*, so to the extent that *S*'s behaviour is causally explained by *C*'s capacity, the explanation isn't a semantic one, referring necessarily to a certain relation between *C* and something else. Fourth, if *C*'s function were to have just the capacity to indicate *F*, the function couldn't also be what explains, in an etiological way, the configuration of $C \rightarrow M$. This is because an earlier *C*'s having the *capacity* to indicate *F* doesn't explain the wiring of $C \rightarrow M$ as easily as does an early *C*'s actual indication of *F*. Many internal conditions might have had the potential to indicate *F* under certain conditions, but *S* would have to reward this potential by giving that *C* control over *M* merely for something *C* might or might not have done, or which *C* would do, but which no *C* might ever actually do. Were *S* or some designer of *S*, who controls the configuration of *S*, to recognize *C*'s potential and to anticipate the possible conditions under which *C* would indicate *F*, perhaps an early *C*'s potential could structurally explain later *C*s' causal connection to *M*. But this would make the configuration derivative rather than naturally foundational, since the configuration of *S* would depend on pre-existing representations.

To recap, I'm trying to determine Dretske's best response to the objection about local potency. Dretske needs a way of saying that later *C*s have causal power in virtue of their semantic content, even while he identifies their having this content with the function of carrying the information that was actually carried by earlier *C*s, and identifies that earlier carrying of information with the structuring cause of *S*'s behaviour. What's

wanted, perhaps, is a distinction between *the* function, that is, some crucial *token* relation that is the structuring cause, and the *having* or the *performing* of a function. *The* function would be the token that sets a standard and is therefore explanatorily relevant to that which comes after it and which depends on that token. In this case, the function would be an early *C*'s indicating of *F* in *S*. Later *C*s would only *have* the function, performing it or attempting to live up to the standard, as it were. These later *C*s may or may not succeed, but they would still have the function because they would fall under the control of the earlier *C*s and be judged, as it were, by the standard set by these early *C*s. The difficulty is explaining what it is merely to *have* a function without resorting to these normative terms.[25] The early *C*s would help configure *S* such that the later *C*s cause *M* because of the early *C*s' carrying of information about *F*. And the later *C*s would have the capacity to instantiate the explanatorily relevant informational relation, the capacity being what Dretske calls the property *N*. Only the early *C*'s indicating of *F* would be relevant to explaining *S*'s behaviour, but there would be a more general way of talking about the *C*s, which is to say that later *C*s *can* be similar to the early *C*s, by also indicating *F*, and moreover that the later *C*s *should* be similar in the sense that by indicating *F* they do what accounts for their being in a position to cause *M*.

With this distinction, Dretske's notion of a function could be saved from incoherence, but this distinction still wouldn't show that the representational content of later *C*s has causal power over *S*'s behaviour. The representational content of a later *C* would be this *C*'s capacity to be similar to an early *C*, by doing that which explains why the later *C* is where it is in *S*; that is, the later *C*'s representational content is its *having* a

[25] Millikan's theory of purposive functions faces the very same question about what is involved in the *having* of a function. See section 4.4.

function of indicating *F*. *N*, or the capacity to indicate *F*, is just some intrinsic property of *C* by virtue of which *C* will indicate *F* under certain conditions. Now, representational content can't be a later *C*'s own indication of *F*, since all later *C*s are supposed to have representational content even when they don't indicate *F*. Some later *C*s may be unable to indicate *F* under abNormal conditions. The only relevant constant with respect to the later *C*s is the early *C*'s indicating of *F* without which the later *C*s wouldn't be where they are in *S*, as causes of *M*. The later *C*s' role in *S* is dependent on the early *C*s whose rudimentary content helps to create this role, but this dependence isn't itself a semantic relation. At best, the representational status of a later *C*, which makes this *C*, say, a belief *B*, explains indirectly why *B* causes *M*, because of *B*'s dependence on an earlier *C*, but this dependence doesn't mean *B*'s own content is explanatorily relevant. It's still the early *C*'s relation to *F* that explains why *B* causes *M*. Most *B*s will have intrinsic properties that enable them to indicate *F*, even if sometimes *B*s don't actually do so; with these intrinsic properties, any *B* will be comparable to the early *C*s, but that which causally explains every *B*'s causing of *M* will still be the informational relation between the early *C*s and *F*, which helps configure *S*. *B* may be similar to the early *C*s and may indicate *F*, but *B*'s intrinsic properties which make for its capacity to serve as an indicator aren't themselves semantic relations, and *B*'s own informational relation to *F* comes too late to explain *S*'s behaviour. In any case, *B*'s indicating of *F* isn't a representation of *F*. Representational content is a function *had* by *B*, or by a later *C*, and the function seems to be *B*'s capacity to enter into an informational relation, given that only an earlier instantiation of this relation is explanatorily relevant to any later *C*'s recruited causing of *M*.

On this analysis, then, representational content, or some *C*'s function of indicating *F*, isn't the later *C*'s own relation to *F*, but rather this *C*'s ability to perform similar to the explanatorily relevant, earlier token *C*. The only extrinsic relation that causally explains any of *S*'s behaviour, originating with some internal condition *C*, is the early *C*'s informational relation to *F*. When a later *C*, which is wired to cause *M*, indicates *F*, this *token* carrying of informational content doesn't explain why this *C* causes *M*; rather, the later *C*'s causing of *M* is explained by the earlier *C*'s indicating of *F* and by *S*'s recruiting of all *C*s, which are similar to that earlier *C*, as causes of *M*. Moreover, the *property* of indicating *F* doesn't causally explain why later *C*s cause *M*, since what explains *C*'s causing of *M* is a structuring cause, according to Dretske, and the structuring cause is indexed to an early token *C*'s indicating of *F*. *N*, the type that indicates *F*, or the property *C* has in virtue of which it indicates *F*, is instantiated whenever a certain mechanism operates under special conditions that approximate Normal ones for that mechanism's type. Both early and later *C*s have this property, in virtue of their similar capacity, but this doesn't account for the local potency of representational content, for the above four reasons.

As I said, in talking about *N*, Dretske endorses Kim's formulation, so perhaps what Kim says about *N* provides Dretske with a response to the local potency objection. What Kim (1991) says is the following: "We will think of *C* as a token state of the *F*-detector caused by *F*'s presence in *S*'s vicinity at the time; thus, the *F*-detector registers the presence of *F* by going into state *C*. But this happens only because *C* has a certain neurobiological property, *N*, and *in general* the *F*-detector registers the presence of *F* by entering a state with property *N*" (62). So on Kim's view, *N* is a lower-level,

neurobiological property possessed by all *C*s. All *C*s would, presumably, share some neurobiological property, such as a property that indicates *F*, under certain conditions. In virtue of having this property, an early *C*'s indication of *F* causes later *C*s to cause *M*, and later *C*s are recruited as causes of *M* because they too have this property. But the property of being *N*, that is, the property in virtue of which *C*s are instances of type *N*, is an intrinsic property of *C*s, making for their capacity to indicate *F*. Just because all *C*s do indicate *F* under certain conditions, given that all *C*s are instances of type *N*, doesn't mean the relation between later *C*s and *F* causes these *C*s to cause *M*. Thus, the distinguishing of *N* from token *C*s doesn't solve the problem of local potency.

The upshot of all of this is that, given the necessary differences between the early and the later *C*s in *S*, it's hard to see how all of these *C*s could instantiate the same type in such a way that the later *C*s have their own causal power in virtue of their semantic content. But this undermines Dretske's account of the causal relevance of content to behaviour. The point about a semantic system *S* is that the content of *S*'s internal states cause its behaviour. On the one hand, Dretske needs the later *C*s to have content so these *C*s, as representations, can be used in a causal explanation of *S*'s behaviour. On the other hand, the later *C*s' representational content is just their capacity to carry the same informational content carried by the early *C*s that help configure *S*, and it's that earlier carrying of information that causes *S* to recruit all later *C*s as causes of *M*. Even on Dretske's modified view, when someone believes beer is in the fridge, and the belief causes the person to open the fridge door, the belief doesn't do so in virtue of being a belief, or an internal condition with content of its own. Rather, this belief's status as a representation is its function, roughly, of being similar to an earlier internal condition.

Specifically, the function *had by* the belief is the belief's capacity to indicate some external condition, where an earlier *C*'s indicating of this same type of external condition accounts for the belief's control over the person's movement. This function isn't itself a relation to an external condition, but is, at best, the result of a connection to the early internal condition that carries information, and it's that earlier carrying of information that causes the behaviour of the belief's causing of movement.

Dretske's theory of a representational function is supposed to bridge the gap between the early and the later *C*s, making the content of later *C*s efficacious in virtue of their connection to early *C*s. But the theory is at worst contradictory and at best of no help in explaining the causal relevance of representational content. Either the function is both the producer and the product or the function is the explanatorily relevant, early token indicating of *F*, and the later *C*s have only the capacity to fulfill this function by performing the same sort of task as the one performed by the early *C*. In the latter case, the early *C* is still the only *C* in *S* whose external relation to *F* causally explains any *C*'s recruited causing of *M*, and the early *C* has informational, not representational content. Kim's formulation of Dretske's view is supposed to allow Dretske to say that the content of any *C* in *S* is efficacious in virtue of its instantiation of the neurological type *N*. But the property shared by all instances of this type accounts only for their common capacity to indicate *F*, and this capacity doesn't make for the right kind of cause. Given what content is supposed to cause, namely behaviour, on Dretske's view of behaviour, it seems clear that the only content that could be causally relevant is the information carried by just an early *C*, since the carrying of this information is the only external relation that can serve as the behaviour's structuring cause. Whatever content the later, already-configured *C*s

have, this content can't be a structuring cause of the behaviour of which these *C*s are a part. So if an external relation is supposed to be causally relevant as a structuring cause of behaviour, the relation between later *C*s and *F* tokens must be causally irrelevant. If a semantic system has configured internal states, such as beliefs and desires, that cause movement in virtue of some of their extrinsic properties, Dretske fails to show how such a system could be possible, since he shows only how the extrinsic properties of internal conditions that aren't yet configured could have causal power.

3.7 Natural Selection and Receptivity

I want to consider now another problem for Dretske's theory of content. On Dretske's view, *C*s as tokens of type *N* are recruited to perform a representational function, such as the indication of *F*, because performing this function helps *S* satisfy its receptivity to *R*. But is this recruitment the natural foundation of semantic relations or does it already depend on ones he doesn't explain? Does Dretske explain some beliefs and desires in *S* only by presupposing that there are certain purposive functions that already make for semantic relations? Indeed, what is the source of the *functionality* of later *C*s? Perhaps because of the success of the early *C*s, all token *C*s in *S* come to be causally related to *M*, as Dretske says. But this is not yet to say that the wiring of *C*s to *M*, as it were, has a purpose. Just because *C*s keep causing *M*, helping *S* obtain *R*, doesn't mean the later *C*s have the function of doing this, that the early *C* is a *standard* for later *C*s or that the later *C*s are *supposed* to be similar to the early *C*. Nothing inside or outside *S* is supposed to recognize the capacity of later *C*s in their coming to control some

movement of *S*'s. The later *C*s, of course, can't literally be "recruited" for their similarity to the early *C*, due to any expectation of, *or interest in*, such similarity on *S*'s part, since the configuration of *S* is supposed to be naturally foundational, not directed by a designer or by the use of unexplained symbols.[26] What I will argue in this section and in the next one is that the pre-reinforced, receptive *C* makes for at least as much of an interest in *R* as does what Dretske calls a desire, *D*, or the reinforced *C*. Dretske doesn't explain the receptive *C*'s content, and so his theory is, at best, incomplete.

Dretske seems to have two answers to the question whether the reinforcement of *C* is a process already guided by semantic relations. On the one hand, in the case of representational (as opposed to motivational) functions, he appeals to *C*'s being an instance of type *N* and to *N*'s correlation with *F*. This point, in turn, seems to call upon Dretske's nomological theory of information, and thus on his early, Platonic theory of natural laws (see section 3.2). As long as some *C*s in *S* are instances of type *N*, these *C*s are nomically related to instances of *F*. On the other hand, this doesn't account for the functionality of later *C*s and for the difference between mere information and semantic content. Dretske's theory of how *S* learns to use *C* as a belief or as a desire is supposed to account for this difference. Again, on Dretske's view, later token indicators in *S* come to have a job in *S* due to *S*'s learning to use the early token indicators to get what it wants out of *R*. *S* learns that indicating *F* is needed to satisfy *S*'s receptivity towards *R*, and this learning configures *S* so that *C* comes to cause *M* because of *C*'s indicating of *F*.

---

[26] Recall that Dretske wants to explain how semantic, so-called non-natural relations arise from informational, nonsemantic relations (section 3.2). He's after a reductive theory of semantic content. The objection I'm now raising is that the reduction fails, since his account presupposes that there are certain semantic relations, instead of explaining how they all arise from a nonsemantic base.

However, for the configuration of *S* to be the natural basis of all semantic relations, *S* can't learn to use *C* in this way due to any preexisting *reason* in *S*, such as *S*'s concept of or interest in *C*s as instances of *N*. Any such reason would let *S* recruit *C*s *as* instances of *N*, or recruit them *for* their indicating of *F*. Were the recruitment left to nomic relations, as opposed to being determined by a contingent process that depends on what certain formative particulars do and on the meeting of special conditions, later and early *C*s would equally just indicate *F* or motivate *S* to use *R* in the nomically determined way, since the *C*s would all be instances of the type that enters into the nomic relation. This wouldn't explain what the later *C*s are merely supposed to do, so the recruitment of *C* must be due to some feature of *S*'s learning process. *S*'s *recognition* of *C*'s capacity and *approval* of this capacity would certainly account for *S*'s recruitment of *C*, for *S*'s favouring of later *C*s with a purpose, but this recruitment process would be semantically determined, not naturally foundational. Of course, were *C*'s purpose naturally selected, there would be no need to appeal to preexisting beliefs or desires in this way, but Dretske (1988) argues that natural selection could produce only syntactic systems as opposed to semantic ones. Clearly, *S* doesn't consciously configure its own neural states in learning to turn its signals into representations. But neither can any unconscious representation or motivation help in the recruitment process.

As Fodor (1990) points out, the appeal to a learning process would be useless in a naturalistic explanation of content were the process to depend on an external designer of *S*, such as a teacher who guides *S*'s learning so that the structuring cause of why *C*s cause

M in *S* is a combination of the teacher's own beliefs and desires.[27] This would explain semantic content in terms of other semantic content, and so wouldn't be a naturalistic explanation. The point of Dretske's distinction between types of representations is that only conventional ones have functions that are projected onto the representation by a system that already has representations, whereas natural ones are established without the aid of representations.

In addition, there had better not be a hidden structuring cause *within S* that already has some unexplained content structuring part of *S*. An example will show how this might work. Suppose an organism *T* controls *S*'s conditioning in the following way. *T* has a receptive *C* that isn't yet functional in *T* and that is receptive to an external condition *R*. Suppose that to use *R*, *T* needs to design the way *S*'s internal indicators control *S*'s bodily movement *M*. Suppose also that *T*'s receptivity to *R* somehow wires some internal indicator in *S* to *M*, so that this indicator acquires a representational function. So instead of the indicator's being recruited by *S*'s learning how to use *R*, given *S*'s own receptivity to *R*, the recruitment happens because of *T*'s receptivity to *R*. The key assumption so far is that *T*'s receptivity is just a receptive *C* in *T*, as opposed to a desire *D* with a motivational function arising from *T*'s own conditioning. But suppose, further, that *T*'s receptivity to *R* has some purposive function not established by conditioning, so that the receptivity is still supposed to have some effect or to enter into some relation.

The process by which this representational function in *S* is established by a receptive *C* that is external to *S* would seem to be semantically driven. *S* is wired in accordance with *T*'s functional receptivity, and so the content of *S*'s internal conditions

---

[27] Indeed, one reason to suspect learning might be guided by the use of symbols rather than the natural foundation of symbols is that animals that can learn tend also to be social so that, in practice, *S*'s learning is at least influenced by representations and desires external to *S*.

derives from the semantic properties of *T*'s receptivity to *R*. But what is crucial to the derivative status of the content of *S*'s internal conditions isn't the receptivity's *externality* to *S*. Were *T*'s receptivity made internal to *S*, so that *S* were itself receptive to *R*, the process by which this receptivity wires an indicator in *S* would seem just as semantically driven, the representational function just as dependent on the receptivity's function, assuming the receptivity were to have its own prior function. On Dretske's own view, a receptive *C* bears a relation to *R* and sets in motion the process of recruiting internal conditions that serve as reasons in *S*. Were the receptive *C* already to have a function, albeit not one that results from reinforcement in Dretske's sense, his theory of content wouldn't be reductive. This is because the semantic functions of some internal conditions would depend on the semantic function of a prior internal condition, namely of this receptive *C* which, I'm supposing, acquires a function from some source other than learning. *S* may structure its behaviour in response to *R*, but if the receptive *C* in *S* that initiates this structuring already has motivational semantic content, the behaviour might just as well be structured by *T*, since the structuring would be a symbol-guided process. The structuring of *S*'s behaviour wouldn't be a bottom-up process, since a pre-established functional *C* in *S* would arrange for other *C*s to have semantic functions.

All of this raises the question of what exactly is involved in the process of learning, posited by Dretske. Dretske (1988) points out that the acquiring of beliefs and desires depends on *S*'s receptivity to *R*. Without *S*'s early motivation to use *R*, *R* can't serve as a reinforcer of *S*'s behaviour, and so *C* can't be recruited to cause *M* and be given the function, say, of indicating *F*. Again, Dretske explains beliefs and desires in a similar way. *S* is receptive towards *R*, which means that *S* has an internal condition that,

instead of indicating *F*, is *for R*. In so far as *S* has this condition, *S* has *R* as its goal. The internal condition *C* that is receptive towards *R* is recruited by *S* as a partial cause of *M*, given that *M* is needed to obtain *R*, so both desires and beliefs have functions that are products of *S*'s conditioning. *S*'s early success in obtaining *R* increases the probability that the conditions within *S* that have this result are maintained, and so the receptive and indicative internal conditions become wired to *M*, making for desires and beliefs. The configuration of internal conditions in *S* is goal-oriented, in that the internal conditions acquire functions because the work of some of these conditions in *S* is reinforced by *R*, and this reinforcement can happen only if *S* wants *R* in the first place, or if the use of *R* is rewarding to *S*. Were *S* indifferent to *R*, there could be no early standard *C* for later instances of its type, that wins a place in *S* due to this condition's assistance in *S*'s use of *R*, and thus there could be no representational or motivational functions, and no semantic relations.

As I noted earlier, I think there is a problem with Dretske's distinction between receptivity and desire (section 3.3, n.13). On the one hand, in Dretske (1988), he uses the label "*D*" for the receptive *C* even though this internal condition hasn't acquired the function of being related to *R* or been recruited as a cause of *M*. As he says, "I shall use the letter *D* to stand for the receptivity of an organism relative to outcome *R*" (110). Thus, he says that "behavior that is reinforced by *R* will be behavior in which *D* is recruited as an internal cause of whatever movements the behavior requires. *D* becomes a cause of *M* because *M* results in *R*" (113). But this talk of *D* as that which is recruited, as what I've been calling the early or the receptive *C*, conflicts with Dretske's claim that his account

of desire is supposed to be exactly parallel to his account of belief.[28] In his account of belief, he's more careful in distinguishing between the early, information-carrying *C* and *B*. *B* is the *C* recruited by *S* because of the earlier *C*'s work. "*C* is recruited as a cause of *M* because of what it indicates about *F*" (101). Speaking of the total internal cause of movement, Dretske says that "*B* is, in effect, that *part* of *C*, the internal cause of movement, that represents the current state of external affairs" (110). So the *C* that has yet to be recruited indicates *F*, whereas *B* represents rather than just indicates *F*, meaning that *B*, but not *C*, can make a mistake. Thus, this early *C* can't be identical with *B*. As Dretske says, the learning process "will result in the recruitment of an *F*-indicator as an internal cause of *M*. We have relabeled this internal indicator *B*" (112). Here, *B* is the *recruited* indicator *C*, not the early, pre-recruited *C*. At least part of the reason for any lack of clarity in Dretske (1988), about the difference between certain internal conditions, is his not having then done what Kim would later suggest should be done, which is to distinguish more carefully between types and tokens of these conditions.[29] Even before following this suggestion, Dretske is clearer about the difference between the earlier and the later *C*s—the latter being *B* or *D*—in Dretske (1990), where he says the following:

> In the case of belief, the learning process converts internal indicators of *F* into representations of *F* and, at the same time, makes this fact relevant to an explanation of the acquired behavior. In the case of desire, the learning process converts receptive states for *R* (internal states that make condition *R* reinforcing) into (pure) desires for *R* and, at the same time, makes this fact (the fact that these internal states are *for R*) relevant to the explanation of the acquired behavior. (835)

---

[28] Dretske (1990) says that his theory is committed to the implication that "makes my account of desire exactly parallel with my account of belief" (834). He adds that he came to appreciate the parallelism between belief and desire, regarding the way learning converts informational and receptive states into beliefs and desires, respectively, only after reading Bratman (1990), a review of Dretske (1988). "This parallelism between belief and desire...now seems obvious to me...I am grateful of him [Bratman] for rubbing my nose in it long enough to make me understand" (835, sic).

[29] As I pointed out in section 3.5, Dretske (1991) does endorse Kim's suggestion that this distinction be made explicit.

Here, Dretske distinguishes between the early, receptive *C* and the desire for *R*.

Another possible reason for Dretske's earlier blurring of the line between the mere receptive *C* and the functional *D* is his relatively brief discussion of motivational relations, compared to his detailed treatment of information. His account of belief rests on his account of information, and so he can speak of the early *C* as merely indicating, not yet as representing *F*. But what is the parallel naturalistic basis of his account of desire? Again, the answer Dretske gives is in terms of "receptivity." The function of desires has to rest on a natural relation, for Dretske's account of desire to be parallel to his account of belief. The content of beliefs depends on information and thus on natural laws and relations between properties. But Dretske doesn't give as detailed an explanation of receptivity as the basis of motivational functions, even though, on his view, these functions are necessary to the configuration of *all* internal conditions whose semantic relations to external conditions explain *S*'s behaviour.

One thing Dretske's later work makes clear, though, is that, on his view, the use of mental symbols is governed only by subjective norms that derive from human purposes, interests, and attitudes. Dretske (2000) argues that just as certain weather conditions aren't "necessarily" or essentially" good or bad, but are only subjectively regarded as such, given certain interests or intentions to act, beliefs aren't necessarily or essentially normative, but are only subjectively, derivatively so. For example, even if a belief were true or false, or correct or incorrect, this wouldn't make the belief objectively normative. Truth and falsity themselves aren't normative, but are only regarded as such by creatures who may prefer the truth, given certain goals (247). We need to distinguish,

he says, between the norm-free concepts we intend to use, and the goals generating the norms that govern the use of concepts.

> All intentional acts, in virtue of being intentional, bring the actor under the purview of norms in the sense that the actor is obliged (ought) to adopt the means she believes necessary (in the circumstances) to do what she intends to do...If the act (of applying the concept) is intentional, it will come under the purview of norms, not because concepts or their application is a norm-governed activity, but because the act of applying them, *when it is intentional*, generates a set of norms associated with the actor's intentions and desires. (250-251)

Dretske is addressing here the question whether a semantic relation itself or only a goal-directed use to which this relation is put is normative, and his answer is that the latter is the case. If *S* wants to have a picnic, sunny weather is a means of achieving *S*'s goal, and thus this weather is subjectively good, when viewed from the perspective of someone with the goal. Without *S*'s goal, there is nothing good or bad about the weather. Likewise, the content of mental symbols is only subjectively correct or incorrect, given some goal-directed use of the symbols.

Recall that Dretske commits himself to the commonsense view of purposive functions, according to which a function has mainly a normative aspect (section 3.6, n. 23). Representational and motivational functions are just roles that internal conditions are *supposed* to play. Assuming, then, that his view of the subjective normativity of the content of mental representations carries over to his view of the normativity of representational functions, the source of goals, namely desires, must be the source of these functions. As shown above, Dretske takes the receptive internal condition to be the starting point of desires, of motivations that have semantic properties. The source of *S*'s initial receptivity to *R*, then, must be the source of the functionality of representational functions, in that these functions are only means of satisfying the receptivity. However, if

the receptive *C* already has a function, this *C*'s relation to *R* might already count as semantic, making Dretske's theory of content nonreductive.

Clearly, on Dretske's view, this *C* can't have a function that derives from a learning process since receptivity is a precondition of such a process. But receptivity surely has a naturally selected function. Assuming some internal condition *C* is naturally selected to perform the function of making *S* receptive to *R*, the question, then, is whether *C*'s naturally selected receptivity to *R* could be a semantic relation. Were reference to a past instantiation of *C*'s receptivity to *R* needed to explain the function of a present *C*'s receptivity, being receptive to *R* would be *C*'s function. That is, the function wouldn't be just the proliferation of genes, since in this case the genes would proliferate, in part, because of a previous *C*'s receptivity to *R* which increased the earlier organism's fitness, or the chance of its surviving until the organism could reproduce. This would be consistent, then, with the claim that *C*'s mere receptivity to *R* already has *semantic* motivational content in something like Dretske's sense, since the receptive *C*'s evolutionary function would be to enter into a motivational relation to *R*.

3.8 Receptivity and Intentionality

Now, Dretske (1988) argues that pure desires have four aspects of intentionality, and it seems to me that pre-reinforced, receptive *C*s have three of these aspects, and that the third and fourth aren't relevant. If I'm right about this, the early *C*s would seem to have content that isn't explained by Dretske's theory of how semantic functions are configured by a learning process. So instead of reductively explaining semantic relations,

Dretske would be explaining how some semantic relations arise from others. I'll discuss each of the four aspects and then turn to some other reasons Dretske might offer for denying that naturally selected functions can generate semantic content.

First, movement caused by a pure desire can fail to satisfy the desire, and yet reference to the desire's content accounts for the movement. To take Dretske's example, a rabbit can try to drink from an empty feeding tube, wanting the tube to be full. Likewise, any naturally selected trait can fail to perform its purposive function, due to abNormal conditions, and yet reference to the trait's function accounts for what the organism does with the trait.[30] For example, members of many species have hardwired hunting techniques they use when hungry, and the satisfaction of their naturally selected interest in finding food depends on whether circumstances favour their techniques.

Second, and closely related to the first aspect, a desire can be satisfied *by* something without being *for* this thing. For example, a rabbit can be satisfied by beer even though it has a desire for water. Likewise, assuming *C* is a naturally selected trait with the function of making *S* receptive to *R*, *S* may be satisfied by something other than *R* under abNormal conditions, and yet *C* would retain its function, established by earlier Normal conditions.

---

[30] I'm assuming throughout this section that there are objective purposive functions in the first place, such as naturally selected ones. Both Dretske and Millikan argue that there are and that these functions determine semantic relations without themselves depending on the use of symbols that already have semantic content. However, one of the conclusions of this chapter is that the function to which Dretske appeals is subjective and symbol-guided after all, and I argue at length in the next chapter that the function to which Millikan appeals, developed by a process similar to natural selection, is likewise subjective. My reasoning for assuming in this chapter what I deny in the next is that my overarching criticism of Dretske's theory of content, developed in Chapter 5, is meant to be internal to that theory, and it's Dretske who must assume that a process sufficiently similar to natural selection can produce an objective purposive function. This is because if that sort of process can't do so, neither can the process of learning by reinforcement. I say more about this in section 5.2.2, n.11.

Third, a statement about a desire is referentially opaque in Chisholm's sense (section 3.2, n.3). But again, this is true also of a statement about the function of any naturally selected trait. In the case of a statement about a propositional attitude, the statement's referential opacity is due to the availability of different ways of referring to the same thing. For example, Oedipus wants to marry Jocasta but not his mother, even though Jocasta is his mother. So "Oedipus's mother" can't be substituted for "Jocasta" in the statement, "Oedipus desires to marry Jocasta," without changing the statement's truth value, even though the two expressions are co-extensive. Likewise, there are different ways in which a naturally selected trait performs its function, namely under Normal or under abNormal conditions, and the function is most likely performed under conditions that approximate the Normal ones. Although "Jocasta" and "Oedipus's mother" are co-extensive, "Jocasta" has the connotation, when used in certain contexts such as when uttered by Oedipus, that Jocasta is not his mother. In this case, the context is the set of Oedipus's background beliefs about Jocasta which are cognitively related to his desire to marry her. In the case of a naturally selected trait, the context is the Normality or the abNormality of the conditions under which the trait is used. In either case, referential opacity is due to the dependence of a statement's truth value on the statement's context of utterance. For example, suppose "protect" and "defend" have the same denotation, but "protect" comes to connote something that happens in a Normal situation, while "defend" comes to connote something that happens in an abNormal one. In this case, the truth value of "Turtle shells protect the turtles" wouldn't necessarily be preserved were "protect" replaced with "defend," even though the terms in the statement are coextensive. Thus, assuming for the sake of argument that there are objective purposive functions,

intensionality isn't sufficient for intentionality. After all, an ascription of a purposive function to a naturally selected trait could be referentially opaque, but not intentional.

Fourth, a pure desire for *R* depends on an organism's ability to distinguish *R* from something else. Again using Dretske's example, if a rabbit can't distinguish between iceberg and romaine lettuce, the rabbit can't have a pure desire just for one of the types of lettuce. Unlike the other three aspects of intentionality, it seems that a naturally selected trait fails to have this fourth aspect. For example, even if a rabbit can't distinguish between lettuce and some artificial lettuce substitute, the rabbit may have a naturally selected taste just for lettuce. This is because the actual environment that selected for a token rabbit's receptivity to the leafy vegetable would have contained lettuce rather than the substitute.

The point about this fourth aspect, however, is an epistemic point about the limits of explanation, not about a semantic property of desires.[31] If a rabbit behaves exactly the same way towards romaine and iceberg lettuce, there is no need to assume the rabbit has a desire for just one of the two kinds of lettuce. All that's needed to explain the rabbit's behaviour is to assume the rabbit desires lettuce in general. But there remains the possibility that the rabbit has two desires, one for each kind of lettuce, and that these desires are each wired to the same movement as their means of satisfying the different desires. In this case, the rabbit's bodily movements might not provide evidence of the rabbit's separate desires, but this claim about evidence has to do with the vantage point of the person explaining the behaviour. Recall that behaviour, on Dretske's view, is a connection between an internal condition and an external movement. Were there access

[31] As Dretske (1988) says, the properties of a reinforcer that are relevant to *S*'s desire are those that figure in an *explanation* of *S*'s behaviour (129).

to the internal conditions, such as to the rabbit's neural states, there might be reason to offer two explanations, despite the rabbit's same movement towards either type of lettuce, given a difference in the internal causes of these movements. In this case, there would be reason to posit two kinds of behaviour and thus two desires. Alternatively, there might be minute differences in the outward movements, indicating a preference for one type of lettuce, and these differences might not be detected in ordinary situations. So this fourth aspect has more to do with conditions of explaining behaviour, than with conditions of desires themselves.

Of course, assuming not all naturally selected traits are symbols with semantic content, the three aspects of intentionality that these traits have must be insufficient for intentionality. My point is just that Dretske can't appeal to these aspects in support of a claim that a later, reinforced *C*, or *D*, has motivational semantic content, whereas the early, pre-reinforced *C* doesn't.

Still, Dretske (1988) might support that claim in other ways. There is, he says, a difference between a *drive* and a *desire*. A drive, or an instinct, is a naturally selected internal cause *C* of some movement *M* and the cause is selected because of some beneficial consequence, *R*, of *M*. An organism *S* is driven towards *R* as its goal, and so *C*'s causing *M* is goal-*directed* behaviour. A selectionistic explanation of the behaviour is that the genes that produce *C* were transmitted in the past because an ancestor's drive towards *R* caused *M*, which resulted in *R* and which thus increased the ancestor's fitness. To use one of Dretske's examples, squirrels have an instinctive sequence of arm movements they use to burry nuts, and this sequence was selected because storing food in this way was beneficial to the squirrel's ancestors; the squirrel's ancestors that didn't

bury nuts in that way had a greater chance of dying prematurely than of living long enough to reproduce. All of this Dretske grants. He argues, though, that none of this shows that the drive is like a desire in being goal-*intended*. A desire is *for* its goal, but a drive is just an internal state that sometimes has beneficial results. Perhaps, then, the receptive *C* is a drive which, in some organisms, is converted into a desire by reinforcement. In this case, there would be no presupposition that there's some semantic content, in explaining representational and motivational functions in terms of a process beginning with receptivity.

Dretske seems to have two separate arguments supporting the conclusion that a drive is goal-directed but not goal-intended, although he connects these arguments. I'll treat them as separate and then see how they might be combined. First, a drive isn't "modifiable" by learning, which is to say that the driven behaviour will occur automatically, even in abNormal situations in which the behaviour doesn't have the beneficial effects and may, on the contrary, harm the organism (123). For example, squirrels engage in their nut-burying movements regardless of whether the movements are likely to succeed: squirrels will move as though they were digging and covering up nuts even when on a hardwood floor. To take another of Dretske's examples, blowflies instinctively extend their proboscis when they detect sugar water and haven't eaten in some time. When the nerve informing the blowfly's brain as to how much sugar water is contained in the foregut is cut, and the blowfly sucks the fluid, the blowfly will continue to do so until it bursts. Dretske adds that behaviour that isn't the product of learning can't have as its structuring cause *C*'s relation to a goal, making *C*'s content causally relevant to explaining the behaviour. Thus, a drive can't have the same kind of content as a desire,

or as a product of learning.[32] Driven behaviour may be explained in terms of the tendency for the behaviour to have a certain result, but this isn't to say the behaviour can be explained in terms of what the behaviour is *for*. This amounts to saying that a drive lacks semantic content, in Dretske's terms, because no semantic relation between the drive and its goal is needed to explain driven behaviour. The reason is that driven behaviour isn't modified by learning and thus occurs even when the behaviour will plainly fail to achieve the goal.[33]

This argument seems to assume that movement caused by a desire, rather than by a drive, wouldn't happen unless there were a high chance that the movement will achieve the goal. A goal that isn't modifiable by learning is just one that is less likely to be achieved in as many situations as is a goal that is so modifiable. But the assumption that desires cause movement only when they are likely to be satisfied is surely false, and it also makes the argument irrelevant to the question at hand. An organism's behaviour will be automatic and unsophisticated (futile or counterproductive) if the organism has few behavioural options to suit the variety of situations in which it may find itself. If an organism isn't perfectly adapted to an environment, it will try out the behaviours it can perform even if they will likely fail under the suboptimal conditions, because the

---

[32] If the blowfly "is incapable of *learning*," he says, the cause of the blowfly's proboscis extension "is *not* a desire for sugar water (or for anything else). Unlike a desire, it cannot explain the fly's behavior in terms of what it is for. Though it may produce movements that normally have *R* as their result, it is not *for R*" (124).

[33] I should add that a pre-reinforced receptive *C is* modifiable by learning, since *C* is precisely an internal condition that is configured by a process of learning in Dretske's sense. In this case, the receptive *C* isn't a drive. But I don't think this is what Dretske means by "modifiable." Were *S* receptive to *R*, but to fail to achieve its goal, due to unfavourable background conditions, so that *C* isn't reinforced by *R*, *S* doesn't learn to connect *C* with *M*, and thus *C* doesn't become *D*, the receptive *C* would cause *M* in an automatic fashion, like a drive. The receptive *C* that makes *R* relevant to *S*, even before *S* first encounters *R* and learns to enjoy *R*'s benefits, seems to be a drive or an instinct. At least, I don't know how else Dretske would explain initial receptivity. Also, *S*'s receptivity may not disappear when *C* becomes *D*; the receptive *C* is modified in the sense of being channeled into *D*, but not eliminated. The kind of modification Dretske seems to have in mind is the ability to override some motivation, to stop performing the motivated behaviour, such as when conditions are unfavourable. However, a receptive *C* that becomes a desire is configured, if not eliminated, and so is modified in some way.

organism's range of behaviour is limited. But this mismatch between behaviour and a possible environment has to do with much else besides whether a particular internal condition is something other than a desire in a semantic sense. If Dretske is saying that some necessary conditions of *S*'s having a desire are that *S* be well-equipped to deal with many situations, and intelligent in determining what response a situation calls for, then indeed not all motivational internal conditions have semantic content. Lower organisms such as squirrels and blowflies will be motivated to achieve *R* and won't have a desire for *R*. But this will be because having a desire requires more than having an internal condition whose function is to enter into a motivational relation to a goal. Dretske will have defined "desire" in terms of a sophisticated *use* of motivation.

So from the fact that many organisms have relatively unsophisticated ways of achieving their goals, it follows only that they have drives rather than desires in that they lack, for example, sufficiently intelligent means of putting their motivations to work. What doesn't follow is that reference to a motivational semantic relation isn't needed to explain their behaviour, and that they therefore lack desires in this narrow respect. Dretske needs a reason why a desire in just the sense of having something *for* something else, has to be put to *intelligent* and *flexible* use. Without such a reason, behaviour that is merely driven, in being poorly adapted to some environment, may still be caused by a desire in the sense of having a semantic explanation. In other words, some drives can have the semantic properties of desires even if no drive has all of a desire's other properties, assuming desires must also be put to intelligent and flexible use.

Also, Dretske's argument can be parodied, meaning that the same objection can be raised against saying that a reinforced motivation, or a desire in Dretske's sense, enters

into a semantic relation with a goal. The type of learning Dretske talks about will fail to equip an organism with the means to succeed under *all* conditions. Depending on an organism's intelligence, its flexibility, and its type of detectors, an organism may learn to detect *F* to achieve *R*, but may be equipped to succeed in achieving *R* only under background conditions similar to those of the learning process. When placed in a different environment, the organism may detect *F* and produce movement *M* but fail to get *R*, because the background conditions differ. If the organism has a limited range of movements, it may have no choice but to keep trying with *M*, in which case its learned behaviour will appear to be merely instinctive and driven. But this doesn't necessarily speak to whether the animal has or has not a desire *for R*. In addition, some organisms may break down from shock when faced with a new situation or environment, so that they use their learned behaviours in an automatic fashion. Again, all of this is beside the point at issue, which is whether a certain drive, or naturally selected motivation, is a desire in the sense of an internal structure with certain semantic properties.

Dretske's second argument is more relevant to this point. He argues, in effect, that any content of a naturally selected motivation would lack local potency, since what explains the motivated behaviour is the connection between an *ancestor's* motivation and the beneficial consequence of the controlled movement. Therefore, once again, a drive isn't a desire because reference to a token drive's own content isn't needed to explain the behaviour caused by the drive, whereas a desire's content is locally potent and therefore explanatorily relevant to the behaviour caused by the desire. In Dretske's words, a particular animal

> inherits genes that program *d* [a drive] to cause *M whether or not M* tends to yield *R*. The explanation for the fact that the animal inherited these genes may reside in

> the fact that productions of *M* by ancestors of this animal tended to yield *R*. But what happened to ancestors of this animal says nothing about what the productions of *M* in this animal *did* or *will* yield. (125)

The problem with this argument, given the arguments in sections 3.5 and 3.6, is that Dretske's own theory of beliefs and desires has the same problem explaining the local potency of their content. Of course, in the case of natural selection, the ancestor's particular drive is much further removed from most of the drives that have the biological function of producing the same beneficial results, than is the receptive *C* from the reinforced *D* in the same organism. But there is still the same problem of local potency because both natural selection and reinforcement are processes stretched out over time. Whether many organisms are changed or just one organism is, there will be a difference between the earlier and the later periods in which the change happens. The point, then, is that if naturally selected drives lack semantic content, because of the local inefficacy of their relation to their goals, reinforced desires lack semantic content for the same reason. If, for some reason, local potency isn't a problem for the content of desires, it shouldn't be a problem for the content of drives.[34]

Now, Dretske also combines the two arguments, or more specifically uses the first to support the second. Thus, after pointing to the problem of local potency (without calling it this), he goes on to say that "*This* animal may be in a completely different

---

[34] Dretske can turn here to his definition of "behaviour" as movement having an internal cause, that is, having a cause within a single organism *S*. Naturally selected, driven behaviour is displayed by all representative members of a species. But Dretske grants that the internal/external distinction here is meant to be flexible, so he should have no objection to explaining a species' behaviour by appealing to the content of a naturally selected condition internal to some representative members of the species (see section 3.2, n.7). Instead of explaining an *individual's* behaviour by appealing to an external relation that structures the behaviour, as the individual undergoes a process of learning, behaviour that is typical of a *species* can be explained by appealing to an external relation (between a drive and a goal) that structures the behaviour of the species' members. Instead of some subsystem of *S* recruiting *C* as a cause of *M*, an environment recruits a genotype that codes for a motivation, creates a phenotype, and wires the motivation to some of the organism's movements. The explanations seem parallel in all relevant respects.

environment, one in which tokens of *M* no longer lead to *R*. Still, given the genetic programming, *d* will *still* produce *M*. As long as the behaviour is not modifiable by learning, nothing will change" (125). That is, the animal will continue to behave in a way that fails to achieve its goal. I think this is to say that the ineffectiveness of driven behaviour under suboptimal conditions is evidence that the drive lacks locally potent content, and thus that the animal doesn't have a desire *for R*. But this is to call upon the intelligence and behavioural flexibility that are correlated with learning, not upon reinforcement itself. Behaviour may be caused by the process of reinforcement or by genetic configuration of a phenotype. Either way, the behaviour can actually succeed only when conditions are favourable, and thus only when an animal is fortunate enough to be in a suitable environment or else sufficiently intelligent to recognize when conditions are favourable, and flexible enough to modify its learned behaviour to suit the situation. If learned behaviour has a better chance of succeeding than driven behaviour, this is because animals that learn are also relatively sophisticated in that they have relatively high intelligence, enabling them to recognize when their behaviour will fail to achieve their goal, and a comparatively wide range of behaviours, enabling them to cope with more situations. Again, if the difference between drives and desires is that desires are found only in relatively sophisticated animals, then the difference isn't that desires, but not drives, have semantic content, or are goal-intended rather than goal-directed.

However, Dretske could argue that semantic content is partly a matter precisely of intelligence and behavioural flexibility, and that these are themselves products of learning, so that a mere receptive *C* that hasn't yet been reinforced lacks semantic content. Still, the effective behavioural responses that a reinforced *C* provides for *S*

would thereby make *B* or *D* different only in degree from a pre-reinforced receptive *C*. In being a product of natural selection, the receptive *C* is already the result of a slow, environmental trial-and-error process, giving the organism a head-start, as it were, in dealing with its likely environment. What the organism's own learning process does is to fine-tune this innate capacity for dealing effectively with external conditions. So if the semantic content of *B* or of *D* is partly the result of the advantages given to *S* by the learning process that configures *B* or *D*, the receptive *C* should have the same *kind* of content, because of the advantages given to *S* by the process of natural selection that configures the pre-reinforced receptive *C*.

Of course, by definition, drives are naturally selected rather than learned, so if motivational semantic content is somehow tied to learning, drives lack this content. But whether this content is tied to learning in a substantive, as opposed to a stipulative way, is just the question at issue. I think Dretske succeeds in showing that learning reinforces an organism's attachment to *R*, and thus that learning attaches, as it were, an individual *S* to its goal. As the organism is encouraged by its success, it repeats and thus perhaps perfects its movements, again strengthening the connection between *D* and *R*. There need be no such actual connection between a particular organism's innate behaviour and the achievement of its naturally selected goal, since only its distant ancestors may have succeeded with that behaviour. A vestigial trait, for example, will be generally useless even if early instances of the trait may once have been useful. Moreover, I think Dretske is right to draw a distinction between drive and desire. Clearly, not everything with a naturally selected function has a desire to fulfill the function. Not only would blowflies desire sugar water and squirrels the burying of nuts, but hearts would desire to circulate

blood, which is absurd. Again, assuming there is such a thing as an objective, purposive function, the function may be goal-directed, but this doesn't suffice for a desire. Nevertheless, whatever background conditions of desires there may be, these conditions will be met by an organism with a receptive *C* that eventually acquires certain functions from reinforcement.

Moreover, as I've argued, a naturally selected trait that has a purposive function seems to have also three of what Dretske calls the four aspects of intentionality, the fourth (and the third) being not such aspects, after all. For example, the squirrel's attempt to bury nuts in a hardwood floor may be instinctive and stimulus-dependent, but this attempt is actually a case of *failure*, which is one of Dretske's four criteria of motivational intentionality. The squirrel fails because its movements have a biological function and thus a purpose that is *supposed* to be achieved. Natural selection and reinforcement are both processes by which internal conditions are configured and given a function. Like a desire in Dretske's sense, a driven receptive *C* has the function of motivating an animal to try to achieve a goal. This *C*'s causing of a movement isn't just something that tends to have a certain result, since *C* operates under Normal and not just normal conditions.

If drives have the relevant semantic aspects posited by Dretske, and meet also the background conditions of desires, at least to some degree, Dretske lacks a reason why drives have no semantic motivational content. But if the pre-reinforced receptive *C* is a drive, and a drive has semantic content, in that the drive has the function of being for the achievement of a goal, Dretske's theory of semantic systems explains the semantic content of some internal conditions by appealing to the semantic content of other internal

conditions. It may be that *S*'s initial receptivity to *R* causes *S* to use certain internal conditions as means towards the end of satisfying the receptivity, even without any recognition of these means as indicators of *F* or as stronger, reinforced motives for using *R*. Nevertheless, the semantic functions of *B* and of *D* derive their functionality from that of the receptive *C*. As Dretske says, their functionality is subjective, or derived from some motivation or interest. Whereas Dretske wants to show that a semantic function is objective and not itself dependent on the use of symbols, he actually shows that the learned function of an internal condition of *S* depends on the naturally selected function of a receptive *C*. This *C* already seems to have semantic content, given Dretske's own view of motivational content, and so the semantic functions of *B* and of *D* are subjective. This is to say these functions are guided by a pre-established symbol, by the pre-reinforced receptive *C*. On naturalistic grounds, then, Dretske's theory is incomplete, since it requires a selectionistic theory of the receptive *C*'s content. This is one reason I want to discuss Millikan's theory in the next chapter, since she explains representational and motivational functions by appealing to the sort of reproductive process that occurs in natural selection.

3.9 Conclusion

According to Dretske, learning is a kind of self-organization, and semantic relations are causally relevant to the behaviour of the self-organized *S*. But Dretske shows only how the content of an early, formative state of *S* causes *S*'s behaviour. Moreover, the process of self-organization posited by Dretske requires *S*'s receptivity to the reinforcer

R, and the receptive *C* seems to guide the learning process with its own semantic function. Assuming there is a naturalistic explanation of the receptivity at the root of reinforcement, the best such explanation would seem to be a selectionistic one. For this reason, I turn in the next chapter to Millikan's theory of content, to a theory that seems no worse with respect to the problem of local potency, and that appeals to a process similar to natural selection as the determinant of semantic relations.

Chapter 4

Proper Functions and Isomorphism: Millikan's Theory of Content

4.1 Introduction

A purposive function is a thing's effect that fulfils the thing's purpose. For example, a hammer's function is the holding together of pieces of wood with nails.[1] Millikan (2004) points out that a theory of purposive functions, such as the one she provides, has implications directly for how a symbol's content is determined, and at best only indirectly for the nature of intentionality. So, for example, an informational theory of intentionality might be combined with a Darwinian, etiological theory of functions. Indeed, her account of these functions isn't by itself her account of the nature of intentionality; instead, as I'll argue in section 4.8, her account of intentionality follows from her metaphysical realism, not from Darwinian considerations.

Before I can address her theory of the determinacy of semantic relations, then, I need to summarize, in section 4.2, her naturalistic view of how things in general can

[1] The effect or result of something's activity should be distinguished from the activity as the means by which the result is achieved. A purposive function is something's purpose, which is to say a certain result or end achieved by some means. Still, some devices that have functions, such as a hammer, fulfill their purpose only by a single means, in which case the end seems to run together with the means, and the device's function seems to be not just the result of what the device does when it succeeds, but what the device does as it tries to succeed, as it were. Thus, a hammer's function might seem to be the activity of driving nails through the wood, not just what is achieved when the hammer's work is done. But I follow the literature in assuming that a purposive function has the appearance, at least, of being a teleological, backward-looking cause, which is to say an effect that something tries, or is supposed, to achieve. Thus, a function is, strictly speaking, just the result that achieves a purpose and that explains why something engages in certain behaviour that is, in effect, an attempt to fulfill that purpose.

succeed or fail at fulfilling their purpose, which is to say her view of the objective normativity of purposive functions.[2] I'll focus on her presentation of this view in Millikan (1984), and I'll argue in 4.3 to 4.7 that Millikan lacks a successful naturalistic explanation of these functions. Then I'll argue in 4.9 that there is a conflict between her metaphysical realism and her account of functions, and I'll suggest that, in any case, Millikan misses an important aspect of intentionality, which is the way symbols are used as stand-ins for what they are about.

4.2 Purposive Functions

What Millikan (1984) says about functions is meant to be broad enough to account for biological and for artificial, or intentionally designed or used ones. The account is etiological in that an effect's functionality is thought to lie in the origin of something's capacity to produce the effect. This capacity is explained in terms of how something comes to be a member of what she calls a "reproductively established family," a group whose members exist because they are reproduced. What the members do as a result of their membership is called their "proper" function, which means that the effect is their own characteristic behaviour as distinct from what they may actually do under certain circumstances or from what a user may intend the members to do (2).[3] For

---

[2] Another reason to consider her account of functions in some detail is to set up a contrast in Chapter 5, between two different strategies for explaining semantic relations. I take Millikan's etiological theory rather than, say, Dretske's to be the best representative of one of those two strategies. The other strategy is the one I take up in Chapters 5 and 6.

[3] In section 4.4 I criticize Millikan's notion that an effect can be something's "own." But her underlying point is that there's a difference between what something is supposed to do and what it can do. Only the former could be the thing's function, and this special quality of the effect is marked by calling this effect the thing's own.

example, a heart's proper function is to circulate blood, not to make noise when an organism runs, and a hammer's proper function is to drive nails through wood, not to be used as a paperweight. Thus, a proper function, whether biological or artificial, is objective and intrinsic rather than purely conventional, but the function is also ideal in that the proper effect may not actually be carried out. A member's own function is distinguished from what the member may do under prevailing conditions, and so a member may have a proper function even though conditions don't allow for the member to perform the function. Instead of being determined by what a member does under the conditions that just happen to be met, the function is determined by the member's relation to what an ancestral member of the same type did under what Millikan calls a set of "Normal conditions." These conditions are those that historically were met for a descendant to have been reproduced in part by its ancestor's having successfully brought about the same effect that is the descendant's proper function. More precisely, Normal conditions are those "that must be mentioned" in a "Normal explanation," which is an explanation "of how a particular reproductively established family has historically performed a particular proper function" (33). By "Normal," she means *descriptively normative*, or what she calls "quasi-normative," as opposed to meaning *normal* in the sense of being average, or indeed as opposed to meaning *normative* in a fuller, prescriptive or value-laden sense (5).[4]

So for an item to have a proper function *F*, the item must be a member *m* of a reproductively established family *R*. *R* is such that all of its members have some

[4] I'll have much to say about the distinction between descriptive and prescriptive norms, in this chapter and in the remaining two chapters. The distinction is supposed to be that descriptive norms don't depend on mental processes, whereas prescriptive ones do, so that the former are objective while the latter are subjective.

properties in common, making up their reproductively established character *C*, and *C* is due to the fact that if the earlier members, the ancestors, have these properties, so too must the later ones, the descendants, because the ancestors cause the descendants to exist by copying themselves. Millikan distinguishes between first- and higher-order *R*s, a higher-order one being some *R* whose members don't reproduce themselves but have their *C* produced by the proper function either of some other *R*, as in the case of organs produced under Normal conditions by genes, or of some device, as in the case of mass-produced commercial products. Millikan adds that members of *R* need not have all of the properties included in their *C*; *m* can be malformed and still belong to *R* as long as *m* is in some respects similar to Normal members of *R*, namely to those members that have the most properties that make up *C*, and this similarity can be explained by an approximate Normal explanation.[5] This way of accounting for malformed members is meant to be vague to allow for the vagueness of whether something is, for example, a malformed eye or a blob of misplaced organic matter (25).

Crucially, an item *x* has what Millikan calls a "direct proper function" *F* "if *x* exists having a character *C* because by having *C* it *can* perform *F*. (Notice how close this is to the idea that *x* exists in order to perform *F*.)" She adds that "because by having *C* it *can* perform *F*" should be interpreted to mean "because there were things that performed *F* in the past due to having *C*" (26). The reason *m* with character *C* has some *F* as its direct proper function is that, as a result of *C*'s causing *F*, *C* came to be positively correlated with *F* over a set of items that included ancestors of *m* with *C* and other items that lacked *C*. With respect to biological functions, this point about a positive correlation

[5] So the malformed individuals can still be members of *R* if they share some properties with Normal members of *R*, because all of these members share a more loosely-defined reproductive history.

is a way of speaking about the competition needed for natural selection. The idea is that if some distinguishing features of *m* have some effect, and this causal relation results in the fact that the same features of more and more members of this *R* produce this effect, whereas members of other *R*s don't become similarly specialized, the effect of those features must give these members an evolutionary advantage. The distinguishing features are assumed to be reproductively established, and so in the biological cases, the effect of these features is genetically determined and selected for. Millikan speaks of a "direct" proper function, to distinguish this function from a derived or adapted one. For example, some feature of the chameleon's skin has the direct proper function of varying the chameleon's skin colour with the colour of what the chameleon sits on. In this case, *F* is relational in that the effect is to instantiate a relation of similarity between a certain skin colour and the colour of something else. This is the direct *F* in the sense of being the most general function that accounts for the presence of the chameleon with its reproductively established ability. But since this *F* is relational, *F* acquires the more specific, context-sensitive function of adapting the chameleon's skin colour to the brownness of a tree branch, assuming the chameleon actually sits on a brown branch. So indirectly, the feature has the temporary, derived proper function of turning the skin colour brown.

To summarize, the view is that an item *x* has *F* as a proper function as long as there is a Normal explanation of *x*'s capacity to perform *F*, according to which *x* is a member of *R*. This means that *x* is at least capable of producing the effect *F*, because of *x*'s special origin. In the case of biological functions, ancestors of a later instance of a trait also performed *F* and this helps explain why the later instance is similar to these

ancestors: Normal conditions were met which allowed the ancestors to perform *F*, which in turn led to the ancestors' survival and thus to their tendency to reproduce members of their type which have the trait that has the capacity to perform *F* under Normal conditions. Earlier performances of *F* by earlier members of *R* cause later members to have the capacity to perform *F*, by causing these members to exist as a result of a reproductive process, and so the later members exist in part because of *F*, which makes *F* their own, proper function.

4.3 Descriptive and Prescriptive Normativity

Millikan (1984) explicitly defines the normative notion of what a functional thing is "supposed to" do in terms of her technical notions of "proper" and "Normal." As she says, a naturalistic account of "supposed to" can be "accomplished if we can show that 'supposed to' can be unpacked in terms of proper or Normal rather than actual relations." She says this because she defines "proper" and "Normal" "as straightforward, causal-order, natural-history categories" (88-89). Thus, Millikan argues that these functions, that determine the content of symbols, are normative but not in any evaluative, value-laden sense. I want to argue, however, that there is no descriptive sense in which something is supposed to happen, that there are no purely descriptive normative terms.

On the contrary, Millikan (2002) says, "normative terms are not always evaluative." Instead, these terms are used more generally

> to indicate any kind of measure from which actual departures are possible. For example, a numerical average is one kind of norm, as is any sort of regularity: 'With that kind of sky in the west it ought to be sunny tomorrow'. (Proper

> functions do not correspond to averages or regularities either, of course. They define a standard of their own kind.) (7)

Millikan wants to distinguish between at least partly subjective and purely objective norms. A naturalistic account of functions should take these functions to be objectively, descriptively normative, and thus to be independent of any observer or of any prescription of what the functional object should do.

She gives three examples of descriptive norms: departures from a measure, averages, and regularities. I'll discuss each in turn, and the point I want to make in doing so is that what makes any of these three normative is prescriptive, and so, contrary to what she says, none is descriptively normative. The problem with saying that something's departure from a measure is only descriptively normative is that the measure itself is conventional. Granted, once a measure, or a reference point or standard is chosen, an objective comparison can be made between objects and the standard. Take, for example, the definition of a meter as the distance traveled by light in an absolute vacuum in a certain fraction of a second. Suppose the length of a certain tree branch is shorter than this distance. This *difference* in length is objective, but this isn't to say the same for the *departure* of the branch's length from that of the distance traveled by light in a certain amount of time. Again, the speed of light is objective, but the choice of this speed as a measure of length is subjective and conventional. Once the standard is stipulated, the standard can be used to discover objective differences, but there would be no incorrectness of something, relative to the standard, without the conventional practice of regarding something as a standard in the first place.

Moreover, the incorrectness or departure of the branch's length would depend on the use of the branch as something that is supposed to be a meter long. Just because a

standard is chosen for the length of a meter, and some physical object somewhere differs in length from that standard, doesn't mean the object departs from the standard. The tree branch also has to be chosen as something—say, as a walking stick—that is supposed to have a property in common with the standard. Anything at all that isn't the same length as the distance traveled by light in a certain period may differ objectively in this respect. But departure from a measure is more than just this sort of objective difference, since the former is normative whereas the latter is not. The problem is that the norms seem at least partly subjective: the epistemic and pragmatic norms having to do with the purposes of measurement are evaluative in that they govern interests in performing certain actions.

The same is true with regard to a numerical average. The single value that represents some feature of the data set depends on the chosen means of calculating the average and on the choice of the data set. For example, the arithmetic mean is calculated by summing the numbers in a list and dividing by the number of items in the list. The numbers in the list can then be compared to the average number. But there are other ways of finding the central tendency, such as the calculation of the median or of the mode. Moreover, just as the length of the tree branch has to be deemed relevant to the standard of a meter, for the branch's length to be said to depart from the standard, the data points have to be grouped into a set for their average to be calculated. For example, the data points might be the marks on a set of final exams. Once these two choices are made, a number in the set can be said to depart, positively or negatively, from the average. But there would be no such departure without the evaluative norms governing the initial choices.

The third sort of descriptive norm to which Millikan refers in the above quotation is just any natural irregularity. For example, there might be a *ceteris paribus* law that under certain weather conditions, the sky is sunny. These conditions may or may not be met, but the sky is regularly sunny when they are met. For a natural regularity to differ from what happens normally or an average, there has to be a way in which the regularity holds even when the conditions are rarely or never actually met. The regularity seems to be an *idealization*, in that assuming the explanation is justifiable and useful, given some prescriptive (epistemic and pragmatic) norms that govern the choice of explanation, the regularity would hold under the stated conditions.[6] I won't argue this in detail here, but will simply assert that, at a minimum, it's unclear whether *ceteris paribus* laws have a purely descriptive status. The point is not the antirealistic one that there would be no regularity under special conditions without an observer interested in explanation; rather, the point is that, given that there is such a regularity, any normative sense in which what happens under different conditions is irregular or a departure does depend on the interest in explanation, which is governed by evaluative norms. If actual weather conditions cause the sky to be cloudy, whereas under special conditions the sky is sunny, cloudy skies depart from the standard, provided by the *ceteris paribus* law about conditions for sunny skies, only if focusing on sunny skies is justifiable or useful for the purpose of explaining the weather.

None of these three ideas, therefore, captures the idea of a non-evaluative, descriptive norm. The reason this is a problem for Millikan is that on her view of what a naturalistic theory of content should do, which I criticize in Chapter 5, there should be no

[6] By "epistemic and pragmatic" norms, I mean the norms of rationality and, more broadly, the values of a person or a culture. Epistemic norms determine the reasonableness of a belief, pragmatic ones the value of an action.

appeal to prescriptive norms in the theory's *explanans*. As I'll argue in section 4.4 to 4.7, the purposive functions that Millikan posits end up being prescriptive rather than descriptive.

4.4 What an Etiological Theory of Functions Explains

I want to discuss now what exactly an etiological view of functions is supposed to explain. Millikan's account has in common with Wright (1973) the assumption that when a function is ascribed to something, the ascription rests on an explanation of why the thing exists. As I'll argue, this assumption is at the heart of an etiological account of functions, but it is also problematic on the naturalistic view that ascriptions of biological functions are only (somehow) descriptively normative. Problems with that assumption are easier to see on Wright's simpler presentation, so I'll begin with Wright's claims and then turn to Millikan's.[7] According to Wright, saying that one of the nose's functions is to facilitate breathing is to answer the question of why the larger system, the person, has a nose, or why the nose is there in its position with respect to the person as a whole. A nose may be good for holding up glasses, but this effect doesn't explain why someone has a nose. Wright then argues that if the explanation is taken to be etiological, or a backward-looking causal one, something's functional effect can be distinguished from its accidental one. Saying that the nose holds up glasses doesn't explain why a nose is there on the face, because this doesn't explain how the nose historically *got* to be there. So Wright proposes that a necessary condition of *X*'s having the function *F* is that "*X* is there because it does

[7] There's an enormous philosophical literature on biological functions. See, for example, Lewens (2004), Neander (1991), and Bedau (1991).

F," with the understanding that "because" is used in an etiological, causal and explanatory sense.[8]

As Wright says, this point about explaining why *X* "is there" is ambiguous, since what might thereby be explained is either why *X* has certain distinguishing features, such as why *X* is positioned where it is in a system, or why *X* exists at all. Wright (1976) says that "is there" is a "general place marker that takes on different significations in different sorts of cases" (81), and Wright (1973) provides some illustrations. The phrase "is there" can mean "is where it is," which is the point about *X*'s position. It can mean " '*C*'s have them,' as in 'animals have hearts because they pump blood.' "[9] Also, it can mean merely, " 'exists (at all),' as in 'keeping snow from drifting across roads (and so forth) is why there are snow fences' " (46). He adds, though, that " 'is there' can only sometimes, but not usually, be rendered 'exists (at all).' So, contrary to many accounts, what is being explained, and what *F* is the result of, can very often *not* be characterized as 'that *X* exists' *simpliciter*" (55, n.19).[10] The other two meanings, *X*'s position and *X*'s being had

---

[8] In speaking of that which has a function in terms of "*X*" that has "Z," Wright might be thought to be speaking only of the function of types, not of tokens. Although he doesn't clearly distinguish between types and tokens in his account, however, he's explicit about the etiological nature of his functional explanation, and this sort of explanation implies the distinction between earlier and later tokens of a type. Millikan (1989) points out that in Wright (1976), he says, " 'because *X* does *Z*' does not reduce to 'because things like *X* have done *Z* in the past' (pp. 89-90)" (121, n.1). According to Millikan, this suggests that Wright's theory isn't etiological after all. But as is made clear in Wright (1976), and as Godfrey-Smith (1994) points out, what Wright means is that if a present instance of a type has a function, the ascription of this function must rest on a law about a causal relation, and this law must not apply *only* to earlier instances of the type. Thus, Wright speaks of functions in a tenseless way. Although Millikan carefully distinguishes between token ancestors and descendants, there is a sense in which a proper function also should be characterized without regard to temporal differences, since a present descendant must have the same function as its ancestor, given their equal membership in some *R*. What produced this membership, the reproduction of earlier members, lies in the past, but what is produced, the similarity between the reproductively established properties of each of the members, doesn't lie simply in the past, relative to some descendant. The similarity relates a descendant to its ancestors, just as a nomic relation connects all the instances of a type that have the same causal power.

[9] Here, Wright emphasizes the *having* of *X*: "If to specify the function of quills is to explain why porcupines *have* them, then the function must be the reason they *have* them" (43).

[10] I would add that saying that *X* exists at all in order to have, or for the purpose of having, a certain effect, is too much like saying that someone lives to perform *F*. The latter is usually a way of saying

by a system, seem to me closely related. For example, part of what is meant by asking why someone *has* a liver is to ask why the liver is *where* it is in the body. The having of a liver is the having of the organ in a certain useful position. In so far as someone is said to have a liver, the point isn't that the liver is kept in a place where it can't be immediately used, such as in a cold storage bin in a hospital. Someone who had access to this sort of liver would be said to have only a spare liver, not a liver as such. And when the functional feature is located in a certain position, such as in an internal or otherwise useful place, the system can be said to *have* the feature; what the system has is at least access to the feature.

Wright raises these questions about *why something is there*, for two main reasons. First, he wants to stress that "the ascription of a function must be explanatory in a rather strong sense." The question, "Why do animals have livers?" is deeper than the question, "What is the liver good for?" in that the latter question can be answered by taking into account accidental benefits of having a liver, as distinct from the liver's function. Indeed, he says, the question, "Why do *C*s have *X*?" is equivalent to the question, "What is the function of *X*?" (43).[11] So the first of Wright's reasons is that an explanation of a function is an answer to a question about a natural regularity, not just about an accident. Wright's second reason is that he wants to link questions about

---

hyperbolically that the person would rather not live than live without performing *F*. For example, someone might be said to live to ski. Skiing might then be considered a function in the sense of being good for the person, because the person's ability to ski is, in effect, a means of keeping the person alive, and living is itself assumed to be good. But if "existing to perform *F*" is just a peculiar way of talking about a conventional, interest-relative kind of function, there is little reason to suppose that the question, "Why do *X*s exist at all?" can be answered by pointing to a biological function that need not be of interest to anyone.

[11] Wright (1973) raises another why-question he says is equivalent to, "What is the function of X?" This question is, "Why do *X*s do *Y*?" as in, "Why do hearts beat"? But clearly this *type* of question isn't equivalent to one about *X*'s function, since the activity might just as well be nonfunctional, as in, "Why do hearts make noise?" Were this why-question equivalent to one about *X*'s function, all effects would be functional since all effects are carried out by something.

function with questions that can be answered by a causal, historical account. On the one hand, there is a ready-made explanation of a feature's origin, such as an evolutionary one of a biological trait. On the other hand, there are commonplace ways of asking about something's function, such as asking about why a certain feature *is there*, or why something *has* a certain feature. An account of the feature's origin can answer these common questions by showing how the feature *got* to be there, and how the thing *came* to have the feature. Thus, on Wright's view, the etiological theory of functions is well-motivated. And some motivation is needed, Wright says, since "Functional and teleological explanations are usually *contrasted with* causal ones, and we should not abandon that contrast lightly: we should be driven to it" (44).

Explaining why something "is there," in the sense either of why the thing "exists (at all)" or why the thing "is where it is," can be done, of course, without any appeal to etiology or to functions. On the deductive-nomological (DN) view of scientific explanation, an explanation of any particular event has to account for the existence of the object whose behaviour is part of the event, by deducing the statement that it exists from other statements. So what makes an ascription of a function to some feature can't be just that the ascription accounts for why there exists an object with the feature that causes the behaviour; otherwise, all DN explanations of particular events would also be of functions. Moreover, the spatiotemporal context in which the event occurs can be explained, along the way, in DN terms, when the *explanandum* is a particular event. Information about the context in which an event is observed can be included in the description of the *explanandum* and derived from laws, observation statements, and statements about initial conditions, in the *explanans*. For example, the statement that a liver is connected to two

large blood vessels in a particular human body, can be derived from suitable biological generalizations and observation statements, such as that any individual undergoing certain embryonic development has a liver connected to two large blood vessels, and that the individual in question has had that development.

My point here isn't that the DN model of explanation is the correct one, but that if the *explanandum* of an etiological account of functions is as broad as that of a DN account of particular events, the etiological explanation can't be correct.[12] Just because evidence of something's origin can be used to explain why the thing exists at all or why it exists in a certain position, doesn't mean this sort of explanation suffices as one of a function. Whether the functional thing's history is necessary to explaining the function is another question, but even were the history just necessary, the etiological theory would lose some of its motivation. Moreover, as quoted above, Wright takes at least one of the ways of asking why something is there, the way being why something *has* a certain feature, to be *equivalent* to asking about its function.

This brings me to Wright's third meaning of "why *X* is there." I think that answering why a system *has* a certain feature does address the question of the feature's function, but that what connects the two questions is an analogy between biological and artificial functions that neither Wright's nor Millikan's naturalistic accounts can use.[13] An artifact, which is to say a product of intended design, *has* its features, in the sense of

---

[12] Buller (1999) points out that the main philosophical framework for explaining functions, that preceded Wright's etiological approach, was the DN model of explanation, assumed by the theories of functions found in Nagel (1961) and in Hempel (1959). In defending an etiological account of functions, Wright needs a way of showing that such an account provides a genuine scientific explanation, given that an etiological account differs from a DN one. Again, Wright shows this by showing how an etiological account can answer commonplace, but deep, why-questions about functions.

[13] Again, by "natural" and "artificial" functions, I mean, respectively, functions that don't or that do depend on how the functional item is used or designed by some organism or someone with intentions towards the item.

possessing them, only derivatively, at best, since the owner or user has the artifact as a whole in a way that depends on intentions to produce or to use the artifact. For example, a hammer can be said to have certain features such as a wooden handle and a blunt metal end, which are usually those the producer or seller wishes to highlight. The hammer has these features, but the hammer in turn is used by someone with intentions towards the features. In the case of a type of artifact, the answer to the question, "Why do *C*s have *X*?" is equivalent to, "What is the function of *X*?" because the answer depends on an answer to another question, which is, "Why do the users have *C*s?" The answer to the latter question refers to the users' intentions towards *C*s. For example, some hammers have wooden handles because people have hammers, and the reason people have hammers is because they wish to swing them and thus need to grip the hammers. The handle's function is determined by that intention, and a function that depends on the intention to use something is subjectively normative, not objectively or descriptively so.

So an organism's having of a trait would indicate that the trait has a function, were the having of a trait analogous to a user's having of an artifact with intended features. But no such analogy can be part of a naturalistic theory of biological functions, because such a theory assumes that these functions are unintended effects of an evolutionary process. Even were the analogy to appeal only to an organism's intention to use its own traits and not to any designer's intentions, no proponent of the etiological theory could make use of the analogy, because the source of the trait's function would be the intention to use the trait, not the trait's origin. This, of course, would count against an etiological theory of functions. Millikan, in particular, has no use for such an analogy, since she wants the notion of a biological function to serve in her theory of content, and

so that notion can't refer to the content of intentions towards functional features. Moreover, she doesn't want to posit prescriptive norms in this theory.

Wright (1976) suggests that "teleological expressions in most nonhuman applications represent dead anthropomorphic metaphors" (21). "A metaphor dies," he says, "when the metaphorically extended use of a term becomes established more or less independently of the original paradigm" (19). Also, the substantive proposition put forward by any metaphor with empirical content can be translated into nonmetaphorical language (18). Wright might respond, then, by saying that he can make use of the analogy between biological and artificial functions as long as the analogy isn't a "living" one that does any work in the naturalistic explanation of biological functions. Thūs, when someone asks, "Why does the turtle have a shell?" this question might be equivalent to one about the shell's function, on the assumption that the turtle has the shell in the same way that someone has an artifact, with intentions to use the artifact in some way. This, though, would be a dead metaphor, in that the *explanans* of the etiological theory wouldn't refer to anyone's intentions in explaining the shell's function, although someone operating under the original paradigm might have assumed there are such intentions, such as intentions on the part of a designer of the turtle.

The literal sense of "Why do *C*s have *X*?" would have to be something like, "Why are *C*s identified or distinguished by *X*?" as in, "Why are hearts identified or distinguished by the thumping noise they make?" But this sense of the question isn't equivalent to "What is the function of *X*?" since, evidently, not every feature that can be used to identify a type is a functional feature of the type. This suggests that the why-question an etiological theory is supposed to be able to address, by explaining how *C*s

came to have *X*, is equivalent to a question about *X*'s function only when the why-question inquires about an artificial function. Given that *C*s are ultimately users with intentions towards the artifact *X*, explaining how *C*s came to have *X*, without appealing to *C*s' intentions, doesn't explain *X*'s function.

To summarize, Wright claims that asking about a function is the same as asking for an answer to a certain why-question, or for a deep explanation. A causal account of how something with a function came to be *at all*, or came to be *where* it is or *had* by a system is supposed to answer the why-question, provide the deep explanation, and thus account for the function. This is how Wright motivates an etiological theory of functions, which is to say a naturalistic theory that speaks only of a causal relation that accounts for the origin of the functional feature. Functions can be causally, and thus naturally, explained if having a function is a matter of something's "being there," and the thing is caused to be there. But the ambiguity of "why *X* is there" conceals a problem with Wright's account. Some meanings of the phrase don't indicate a function of *X*, and the meaning that does, *C*'s having of *X*, indicates only an artificial function determined by someone's intentions and thus by a prescriptive norm.

Now, as quoted above, Millikan (1984) says that *F* is *x*'s direct proper function "if *x* exists having a character *C* because by having *C* it *can* perform *F*," that is, "because there were things that performed *F* in the past due to having *C*." She adds, "(Notice how close this is to the idea that *x* exists in order to perform *F*)," taking for granted that "*x* exists in order to perform *F*" is a statement about a function. This is the less thorough of her two definitions of "direct proper function," but here, at least, she seems to adopt Wright's way of motivating an etiological theory of functions. Her *explanandum* is why

"*x* exists having a character *C*."[14] For Wright, something has a natural function if the thing is there because of an earlier effect similar to the one that is the thing's function, and the thing's "being there" can mean its being had by a system. For Millikan, something has a natural, proper function if the thing exists with an ability because of an earlier effect similar to the one caused by the thing's ability, which is the thing's proper function, and the thing is said not just to exist but to "exist having" this ability. Moreover, Millikan (1984) defines what she means by "proper" in "proper function," as a thing's "own" function rather than an effect that depends on actual, and thus possibly abNormal, conditions or on someone's intentions (2). This raises the question whether a function can be a feature's own, in the sense of belonging to the feature, without the function's being artificial or dependent on an intended use of the feature. So Millikan's less thorough account seems open to the objections raised above against Wright's theory. The objection is that what lends plausibility to Millikan's claim that a function can lie at the end of an historical chain of events is the presupposition that the function is somehow interest-relative and prescriptively normative, and thus explainable without the appeal to the functional feature's objective history.

4.5 Does Millikan have a Theory of Functions?

Like Wright, Millikan does not mean to explain biological functions as being dependent on anyone's intentions, but unlike Wright, she eschews analysis of pre-

[14] In the more thorough definition, she changes the *explanandum* to "the fact that *m* exists," removing any reference specifically to the *having* of *C* (28). Again, there are non-etiological explanations of why something exists at all, and explaining why something exists isn't necessarily the same as explaining its function. But Millikan has much more to say about which things do and which things don't have functions, as I'll soon discuss.

reflective why-questions about functions. She says emphatically, " 'Proper function' is intended as a technical term. It is of interest because it can unravel certain problems, not because it does or doesn't accord with common notions such as 'purpose' or the ordinary notion of 'function' " (18).[15] She motivates her explanation by showing that it can account not for the properties functions are naively thought to have, but for analogies between diverse categories which she argues for on independent grounds. She uses her definitions to talk about analogies between "body organs, tools, purposive behaviors, language elements, inner representations, animals' signals, customs, etc" (38). So the above objection, about having or owning a functional feature, might seem irrelevant to Millikan's account even though in her definition of "direct proper function" she uses some of Wright's language.

But it's not the case that the definitions she uses to explain the functions of various things are entirely technical, in the sense of having no burden of capturing any of the ordinary properties of functions. After all, as quoted above, she says that she offers a naturalistic account of how something is *supposed to* have a certain effect. And she does this by unpacking "supposed to" in terms of "proper" and "Normal," which are in turn defined as causal-order categories. Although she doesn't analyze Wright's why-questions, she assumes the functions she explains are in some way normative, which is what is ordinarily assumed about functions.[16] For her account to be naturalistic, she can't explain the normative in terms of the normative, and so her definitions should indeed

---

[15] Millikan is an externalist about content, so ordinary intuitions about the meaning of concepts, such as the concept of functions, have no authority for her. Thus, she says, her "program is far removed from conceptual analysis" (18).

[16] The reason this is ordinarily assumed is that the functions with which most people are familiar are the artificial, conventional ones of things that are intelligently designed, and these functions depend on interests in certain uses and thus on norms of instrumental rationality, of using something as a means of achieving a goal.

refer only to causal-order or to other non-normative categories. So Millikan's *explanation* of biological proper functions would not be reductionistic were her definition of "direct proper function" to refer to something that applies only to artifacts, namely to their being *had* in a way that gives rise to functions. Moreover, her reference to etiology would be superfluous, since her account would imply that the normative aspect of biological functions, which is to say the functionality of certain biological effects, is somehow already prescriptive and intentional.

The threat of this problem might be overlooked as a result of Millikan's belittling of the task of explaining any commonly-assumed normative aspect of functions in non-normative terms. She uses scare quotes to refer to a functional feature's being designed or to the way the feature is supposed to have a certain effect.[17] Thus, she speaks as though she were not in agreement with the pre-reflective notion of functions as being subject to norms, as though some of the functions she explains were only apparently normative and thus in no need of reductive explanation. Moreover, she speaks of "unpacking" the notion of "supposed to," and of "defining" it in naturalistic terms (17). But definitions can be arbitrary, whereas reductive explanations cannot be so. It's not enough for definitions that figure in a naturalistic explanation of functions to refer only to causal-order categories; these categories must be used to *explain* the normative aspect of these functions, among

[17] She says, "Having a proper function is a matter of having been 'designed to' or 'supposed to' (impersonal) perform a certain function" (17). Her reference to the impersonal nature of the norm indicates that she wants to speak only of a nonprescriptive norm. But here are some other examples. In summarizing her view, she says, "Put roughly, the meaning of a sentence is its own special mapping functions—those in accordance with which it 'should' or 'is supposed to' map onto the world" (9). And again: "False beliefs will then appear merely as things that were 'supposed to' have had such and such relations to the actual world" (18). And again: "The intentional is 'supposed to' stand in a certain relation to something else" and "Intentional icons are devices that are 'supposed to' map *thusly* onto the world" (95). In part, she might be using the scare quotes to criticize talk specifically of supposition as a euphemistic way of talking of the normative aspect of functions. But, as quoted, she speaks in the same way of *design* and of what a functional feature *should* do.

other things. Finally, she claims that her definitions are technical and that she doesn't analyze ordinary notions of function, even though she agrees that functions are in some way normative.

Millikan (2002) says that not only "proper function" but "supposed to" are technical terms in her account of functions. Thus, she says, in Millikan (1984) "the term 'supposed to' was defined naturalistically and, indeed, entered in the 'Glossary of Technical Terms' " (n.7). So Millikan might argue that her account of functions is not meant to reductively explain even this normative aspect of ordinary functions, namely that something *has* a function even when the thing doesn't perform the function, because the function is *supposed* to be performed. She might say that her definitions of such terms as "reproduction" and "direct proper function" don't amount to a theory of functions at all, but just to a way of talking that makes sense of similarities between certain biological, conventional, and other items. In this way, she would let the similarities speak for themselves, as it were: she would argue just that if there are conventional functions, and if biological effects are similar to conventional functions, then there are also biological functions. But she wouldn't take up the task of explaining how there could be biological functions that are comparable to conventional ones.

There are two problems with this interpretation of Millikan's project. First, even if Millikan's definitions aren't meant to explain functionality, but are offered purely for the linguistic purpose of providing a way of talking about comparisons between certain items, her choice of labels to attach to her definitions shows that she presupposes that biological functions have a normative aspect. Otherwise, she need not connect the notion of Normality with that of what something is supposed to do; she could use any label at all

for mere linguistic purposes, and needn't speak of unpacking "supposed to" in terms of Normality. What she says about proper functions would be merely definitional not because no explanation is needed of why the analogies hold, but because her own talk of functions doesn't provide the explanation.

In fact, Millikan (1984) speaks of her "theory of proper functions," while adding that despite this "grandiose title," her intention is the humble one of giving her a precise way of speaking about the analogies she finds (38). Thus, she says that what she offers are "definitions," as though she were interested in making only a linguistic point, offering a way of talking that may or may not be useful. However, anything can be compared to anything else in certain respects. The definitions would be most useful were the similarities to have an underlying cause, even were this cause just the ability of various mechanisms to realize some higher-level properties. But if there is an underlying cause, there should be a naturalistic explanation of the resulting similarities; that is, the things that are similar must then be instances of a natural kind, and so there should be a theory of some properties shared by the instances. Thus, if there is an interesting analogy between body organs, tools, purposive behaviors, language elements, and the rest, and what they have in common is their functionality, Millikan needs a *theory* of their functionality, not just a precise way of talking that would be justified were these phenomena really all functional. A less charitable interpretation of her project is that, by switching from talk of theories to talk of definitions, Millikan claims the advantages but not the responsibilities of having a naturalistic theory of functions. One advantage is that, assuming she had a theory of functions, her definitions would be motivated by an account of the usefulness of her generalizations about body organs, tools, and so forth. The chief

responsibility of putting forward a naturalistic theory, though, as opposed to just any way of talking, would be, of course, to *explain* in natural terms how some things could have in common what they are said to have in common. Millikan needs, then, a theory of functions and not just a set of definitions.

A second problem is that, regardless of whether Millikan (1984) is supposed to explain proper functions, her theory of how content is determined, which I've yet to discuss, is committed to the claim that there are these functions even in the biological domain, and this means she owes an explanation at least of the possibility of such functions. She wants to use her account of functions to explain how a symbol token can refer to something even if the token isn't ever directly related to its referent; that is, she wants to solve the old problem of the independence of semantic and of simple causal relations. In naturalistic terms, this is the problem, roughly, of finding the more complicated natural relations that determine content. Thus, Millikan wants to say that the elements of a symbol system have determinate content, even if these elements aren't themselves directly related to their referents, as long as the system reproductively descends from a token system that is so directly related.[18]

For content to be determined in this way, the direct connection between an element of an ancestral symbol system and its referent must be *relevant* to the element of a descendent symbol system. If something with a symbol system is reproduced by some natural process, the descendant probably has the capacity for a comparable symbol system. But this doesn't mean the descendant has mechanisms that actually generate the same symbols as those generated by the ancestor's mechanisms, so that the descendent

[18] As I'll explain later, Millikan takes symbols to refer only when combined to form symbolic complexes such as sentences, so I'll speak loosely here of the need for a symbol system as opposed to an isolated symbol.

semantic relations are fixed by the ancestral ones. As Millikan points out, a token mechanism's performance depends on which surrounding conditions are actually met, so if the descendent and the ancestral mechanisms operate under different conditions, the two won't actually generate the same symbols. The only way for the symbols, or for the mechanisms that are the proximate causes of the symbols, to be the same, given the reproductive connection and the difference in actual conditions, is for there to be a normative analysis of the descendant's relation to its ancestor. Only were the descendant's mechanisms supposed to operate in the same way as its ancestor's, despite the fact that they might not actually do so, given a change of surrounding conditions, and only were this function to determine the symbols' content, might the reproductive history be relevant to a theory of content determinacy. An etiological theory of content needs for there to be Normal and not just normal or average conditions, since the latter don't make the ancestor's behaviour relevant to the descendant's, given a change of surrounding conditions. If the ancestor is a sort of standard for the descendant, so that the content of the descendent symbols is determined by what happens under the ancestral conditions, and the reproduced mechanisms at issue are biological, there must be biological functions and thus there should be a reductive explanation of them.

For these two reasons, I treat Millikan's definitions as offering a naturalistic, reductive explanation. Even if she means for these definitions to address the linguistic and not the theoretical issue, seeing whether her definitions could address the theoretical issue would be a start to seeing whether any other definitions could do so.

Before I move on, I want to point out that in her more recent work, Millikan (2004) does offer just the sort of reductive explanation I think she needs to offer. She argues,

> the purposes of genes, of unlearned behaviors (smiling), of learned behaviors, of conscious intentional actions, of at least some cultural products (greeting rituals), and of artifacts are all purposes in exactly the same sense of "purpose." In all cases the thing's purpose is, in one way or another, what it was selected for doing. Moreover, the purposes we attribute to whole persons, rather than just to various of their aspects or parts, are composed of no more than the purposes of these parts and aspects, and of the way these have been designed to work together. (13)

So all real purposes have the same cause: "adaptation by some form of selection." Although there are "levels of selection," as opposed to just the purposes of genes, all purposes are selected in a Darwinian fashion. For example, citing Popper, Dennett, and others, she claims that "experimental thought attempts to reach consciously projected goals by trial and error," (11), that is, by a similar process to the natural development of different species.

For this reductive explanation to work, the purposes of the parts of organisms can't in turn depend on the purposes of the whole organism; otherwise, the explanation wouldn't be reductive. But as I argue in section 4.6, Millikan's theory of direct proper functions does appeal to epistemic and pragmatic norms that govern inferences to the best explanation and judgments about which conditions are causally relevant. Instead of reducing the prescriptive norms that govern what whole individuals do to the descriptive norms that determine the functions of parts or aspects of these whole individuals, she explains the latter as being dependent on some of the former. In a genuine reductive explanation of prescriptive norms, there can be no reference to these norms in the *explanans*, but that is precisely what Millikan does in her account of Normal conditions.

And when this reference is removed, there is no reason to think that what she explains is any kind of purposive function.[19]

So far I've only raised some preliminary questions about whether Millikan actually explains biological functions in naturalistic terms. To see if she does, I need to consider the pertinent details of her explanation. If the kind of etiology she has in mind does account for biological functions, any comparison she makes between biological and artificial functions is incidental; the appeal to etiology would do the work of reducing any full-blown normative way of speaking about biological functions to a non-normative or pseudonormative way. But if the details of this etiology don't account for functions that are somehow only descriptively normative, my objection about the similarity between her theory and Wright's comes into play. In short, if her theory doesn't show that there are descriptively normative functions, what makes her theory plausible at all as a theory of functions would seem to be that she speaks as if she were explaining artificial functions that are subject to prescriptive norms. In this case, her theory would only wrongly be taken to explain an aspect of biological effects.

[19] It's worth recalling also that Dennett's talk of natural selection as a designer with goals is, in metaphysical terms, only mildly realistic since he assumes the goals are part of a pattern that must be discernable by an intentional stance, or by a form of explanation subject to its own norms (Dennett, 1991a). According to Dennett, the reality of intentional properties, such as the having of beliefs or goals, depends on whether there is a useful perspective for understanding the properties. Thus, although any intentional system has its intentional properties only in this mildly realistic sense, still *any* system whose behaviour can be predicted using intentional vocabulary has those properties to the same extent. This forces Dennett to speak of the rationales of Mother Nature, or of natural selection (Dennett, 1987). Millikan seems to want to defend a less mild form of realism, but then her challenge is to account for the objectivity of biological functions, given only so-called descriptive norms and her identification of semantic relations with relations that have a well-established place in naturalistic ontology, as I'll show later in this chapter.

4.6 Normal and abNormal Conditions

I think there are two main relevant parts of her explanation of functions, these being the distinction between Normal and abNormal conditions, and the appeal to the cyclical process of reproduction. I'll discuss each in turn. As I point out above, Millikan takes her definitions to be of secondary importance, since she takes her etiological theory to be supported by analogies between various functional categories. What the behaviours of body organs, tools, and so forth have in common mainly is the likelihood of their success under Normal conditions. Functional features have Normal, or (somehow descriptively) normative, rather than just normal or average effects that happen under prevailing conditions; proper functions, then, can be explained only by distinguishing between conditions for the success or for the failure of the features, which Millikan calls Normal or abNormal conditions.[20]

For the distinction between Normal and abNormal conditions to help explain biological functionality, the distinction must itself be reductively explained. Millikan (1984) says that what counts as Normal or abNormal depends on a type of *explanation*, which she calls the Normal explanation and which is "an explanation of how a particular reproductively established family has historically performed a particular proper function." As she says, "The conditions that must be mentioned [in such an explanation] are 'Normal conditions' for the proper performance of members of *R*" (33). So why some

[20] In some cases, the abnormality of a proper function, or the failure of the function to be performed on average, is exaggerated. Millikan (1984) says, for example, that "very few sperm actually serve their direct proper functions" (29). She says, later, that this is because "Most never find an ovum and have to call it quits" (34). This is an oversimplification, since there are at least four different kinds of sperm cells: some race to fertilize the ovum, some block sperm from other sources by joining their flagella, some kill sperm from other sources, and some kill any sperm present. Sperm cells work as a team, and this significantly decreases their failure rate.

effect is a direct proper function is supposed to be explained, in part, by the distinction between the two types of conditions: something has this function as long as the thing is a member of a reproductively established family *R*, the member exists with the capacity to produce this effect because the member's ancestors sometimes produced the same sort of effect, and the explanation of the historical producing of this effect requires a distinction between Normal and abNormal conditions. That is, the effect was actually produced under certain conditions, called the Normal ones. For example, mammalian hearts have historically had the effect of circulating blood under such conditions as the oxygen supply to the heart and the presence of a circuit of blood vessels connected to the heart. These are Normal conditions because they must be referred to in an explanation of how hearts have actually carried out their proper function, or of how some hearts had an effect that accounts for the existence of descendent hearts with the capacity to have the same effect. Were earlier hearts not to have circulated blood, the mammals would not have survived and passed on the genes that are crucial to building organs such as hearts.[21]

The Normal explanation is of a reproductively established effect that accounts for the existence of later members of some *R*. Thus, part of what must be explained is the reproduction that establishes the relation between ancestors and descendants. Millikan does say specifically what this part of the Normal explanation consists in: there are "laws *in situ*" that "explain" why two things have some "specifiable range" of properties in common, by correlating these properties such that whatever characterizes the one thing is

[21] In Millikan's terms, hearts are members of higher-order reproductively established families since their functions are Normally explained by reference not to their own ability to reproduce—an ability which they lack—but to the ability of a family of first-order replicators to do so, and to the proper function of these replicators (the genes). In a similar way, a whole organism should be a member of a reproductively established family, which raises the question of what the whole organism's proper function is, on Millikan account.

shown to have to characterize also the other, the direction of causality going straight from the one thing to the other. A "law *in situ*" is "a special law that can be derived from universal natural laws by adding reference to the actual surrounding conditions," such as the conditions surrounding the production of a descendant (20). She speaks also of the Normal explanation as something that "deduces, i.e., shows in detail without gaps" how a certain setup of laws and conditions leads to performance of a proper function (33). The kind of explanation Millikan has in mind, then, is a covering-law one. Something's membership in *R* is explained as long as a member of *R* is caused to have at least some of the same properties as its ancestor, and the special law stating that there is this causal relation follows from a more general law and from reference to some actual conditions.

But there are few philosophers of science today who would regard scientific explanation as just the application of deductive logic to some general laws and to a list of conditions. As has been well-established in the literature, the deductive-nomological model of explanation was supposed to show how a scientific explanation proceeds like a mathematical proof, without any epistemically relevant pragmatic aspect. This model has been shown to be highly problematic, and over the last several decades attention has turned to the social and other pragmatic aspects of explanation. Kuhn (1962/1970) and van Fraassen (1980) are especially influential in this regard. Without here defending any particular theory of scientific explanation, I do want to say that explanation has a pragmatic aspect and that this is problematic for the claim that Millikan's distinction between Normal and abNormal conditions helps to reductively explain biological functionality.

After all, a causal explanation requires a judgment about which factors are *explanatorily relevant.* This judgment in turn seems to depend on background assumptions that are governed by epistemic prescriptive norms or values such as a theory's testability, scope, simplicity, fruitfulness, and unifying power. For example, Millikan (1984) herself speaks of "one among the *legitimate* explanations that can be given" refers to the reproductive process (28, my emphasis). Her own claim, that her etiological theory of functions addresses interesting similarities across a variety of phenomena, implies that one sign of a good theory is its power to offer a unifying explanation. Deductive rules of inference will not specify *which* historical conditions are causally relevant to a process. To the extent that this judgment of relevance depends on a theory that is itself subject to epistemic and other background, pragmatic norms, a biological function, explained in terms of Normal conditions and thus in terms of a Normal explanation, isn't shown to be only descriptively normative. On the contrary, in so far as the Normality of certain conditions determines a biological function, and this Normality is just the role played by certain conditions as these are understood in a way governed by some prescriptive norms, the function seems itself mind-dependent and prescriptively normative. At least, biological functionality would not here be explained in reductive terms, since for Millikan's explanation to be plausible, she would have to posit prescriptive norms to account for the distinction between Normal and abNormal conditions. It's one thing to say that *understanding* Normal conditions requires prescriptive norms governing the attempt to offer the explanation, but Millikan says more than this: she *defines* what it is to be a Normal condition itself in terms of the giving of a type of explanation. Thus, she makes the distinction between Normal and abNormal

conditions subjective and dependent on the prescriptive norms that govern the giving of any explanation.

Again, Normal conditions, for Millikan, are just those that must be posited to give a certain etiological explanation of something's capacity to have a certain effect. Where there are Normal conditions, there are functions, on the etiological view of functions, since functions are the result of reproduction and Normal conditions are those that sustain the reproductive process. There is an unstated assumption here, though, which is that, besides being just a *possible* explanation of the feature's capacity, a Normal explanation, which posits Normal conditions and functions, is a *good* explanation. The capacity must call for this sort of explanation, which is to say either that the feature must seem to have a function or that the explanation must have some utility. Either way, talk of Normal conditions calls on epistemic and pragmatic prescriptive norms. If something seems functional, this is because a judgment is made about the relevance of certain features, such as those taken to be analogous to features of artifacts, and the assumption that biological traits are relevantly similar to artifacts is part of a theory of design that is itself deemed good according to certain prescriptive criteria. If the positing of biological functions has its own utility, the Normal explanation that posits them is itself prescribed.

My main point here is not that the truth of any theory requiring epistemic or pragmatic evaluation is subjective rather than objective, nor that a proponent of such a theory has to be an antirealist or an instrumentalist about the theory's subject matter. My main point, instead, is the following, much more limited one. In the special case of Millikan's theory of direct proper functions, she posits a type of *explanation* in her account of Normal conditions, and then uses this account in her definition of what it is to

be those functions. Thus, Millikan needs to posit epistemic and pragmatic prescriptive norms since there is no explanation without these norms. Therefore, one way she cannot reductively explain the purposive functionality of certain biological effects is by appealing to the distinction between Normal and abNormal conditions, since her account of that distinction must posit prescriptive norms.

4.7 The Cyclical Process of Reproduction

To summarize, the question I'm raising about Millikan's etiological theory of natural functions is whether her theory successfully explains their normative aspect in naturalistic terms. I've argued in section 4.4 that her theory shares some key assumptions with Wright's and is thus open to the objections I raised against Wright's theory. Millikan doesn't mean to rely on an analogy between biological and artificial functions or to posit any evaluative norm as the determinant of biological functions; on the contrary, she is explicit that she intends to do otherwise. But she defines "direct proper function" in Wrightian terms, linking the ascription of a proper function to an explanation of a feature's existence, even specifying that what must exist is something that "has" a capacity to produce some effect, and that a function is "proper" in the sense of being something's "own." These ways of talking make more sense with regard to an artifact's effect, in which case the having and the owning depend on the intention to use the item.

The conclusion that should be drawn, however, is just the following *conditional* one. Suppose that no other part of her theory reductively explains the functionality of biological effects. In that case, what lends her theory any plausibility it may have as a

theory of functions must be that she presupposes that the functions she explains are somehow artificial, intentional and value-laden. Talk of the having of a capacity and of a function as something's own would be a clue that she has no hold of an alternative conception of functions as natural as opposed to artificial effects. The burden that has to be taken up, then, is to explore the relevant parts of her theory to see whether, instead, the *unconditional* conclusion should be drawn that her theory of functions does justify talk only of artificial, prescriptively normative functions, not of so-called natural, descriptively normative ones. So far, I've argued that the distinction between Normal and abNormal conditions can't sustain the reductive explanation.

The cyclical nature of the reproductive process, however, would seem more promising. After all, her theory is an etiological one that links functionality to the origin of the functional feature, by referring to a process or to a history. Wright assumes the process is evolutionary, in the case of biological functions, but Millikan is more explicit. She carefully lays out what she means by "reproduction," distinguishing between token members of a reproductively established family, between ancestors and descendants. So perhaps this is the crucial element of her theory. Perhaps what makes x's effect, F, a proper, natural function isn't just that the existence of x, with its capacity C, can be explained in historical terms, but that the nature of this history is special. Specifically, what explains the descendant's capacity is an effect of the very same type of capacity, possessed by the descendant's ancestor, and so what makes an effect a function might be that the capacity to carry out the effect is the result, roughly, of copying.

With regard to terminology, Millikan (1984) says, "The ordinary word 'copy' probably expresses what I will define better, but that term has problems of its own; so I

have settled on 'reproduction' " (19). She doesn't say what problems these might be, but one such problem seems clear: the notion of copying applies more readily to artifacts than to biological traits. The word "copy" derives from the Latin, "*copia*," meaning abundance, or an overflow in quantity. An overflow is more than *enough* of some quantity, and thus the root meaning of "copy" is *generous reproduction*, or the production of an extra quantity, given that there is already sufficient quantity to achieve some purpose. Were the notion of copying used in a reductive theory of functions, the extent to which the theory succeeds would thus be unclear, at best: while her theory would indeed be one of functions, these functions would be explained as though they were the intended effects of artifacts. A reproduction may be understood as just another production, without any such assumptions about suitable amounts. It's unclear why Millikan thinks "copy" better suits what she actually defines using "reproduction." She emphasizes the way an ancestor makes its descendant similar to itself, so that the descendant is an imitation or a duplicate, but either term, "copy" or "reproduction," carries this meaning of similarity.

Turning to the substantive issue, I want to compare natural cycles that don't generate proper functions, with the reproductive cycles Millikan says do, to see what part of reproductive cycles might account for the functionality. Here are two natural cycles. First, in the Earth's water cycle, the sun heats water in the oceans, causing it to evaporate and to accumulate in the atmosphere; when the temperature cools, this water vapour condenses, forming clouds that produce rain; this precipitation may return directly to the ocean or be collected in lakes or rivers, or in plants where it transpires and eventually returns to the ocean where again it can evaporate. Second, in the rock cycle, rocks on the

planet's surface are eroded by wind and water, forming sediment such as sand; sediment tends to be buried under the Earth's surface where the sediment changes its form, depending on the temperature, and can find its way back to the Earth's surface. For example, the sediment may melt, becoming magma which is ejected as lava from volcanoes, to cool and form igneous rocks which are subject to weathering or erosion, producing sediment. Scientists generally don't regard anything in any stage of these cycles as having a function, but these cycles are similar to the ones Millikan says are needed for historically determined, natural functions.

According to Millikan (1984), the relevant kind of reproduction is such that there is a law *in situ* explaining the similarity between an ancestor and its descendant, by correlating their similar properties and positing a causal relation as holding straight from the ancestor to the descendant so that the descendant must share the ancestor's properties. The law *in situ* is derivable from universal laws that make reference to surrounding conditions of the reproductive process. Minimally, reproduction in Millikan's sense is mechanically produced similarity; the mechanism, or the causal relation, accounts for the necessity that the output resemble the input, and since either the same object can serve continuously as input or else the output can become a new input, the similarity is reproduced. Millikan gives as an example of reproduction the copying of text on paper run through a photocopier. The ink marks on the input page are correlated with those on the output page, and the input page causes the similarity—but only by initiating the copying process. The mechanisms in the photocopier, of course, are causally responsible for the similarity.

Likewise, all generations of clouds are similar to each other, at least with respect to the properties that make them clouds at all, because each generation is formed by the same process of precipitation, evaporation, and condensation that links each generation. That process ensures that had one generation been different, the next generation would have corresponding differences. Of course, cirrus clouds, for example, don't necessarily come from cirrus clouds. But were clouds generally made of some substance other than water, and thus were their surface reflectance and other such properties altered, those differences would be preserved by the way clouds are formed. This is because the process changes only the form of the substance, from rain, to larger bodies of water, to vapour, to clouds. Thus, the general form of clouds is preserved from one iteration of the cycle to the next. For the same reason, generations of igneous rocks have similar properties, such as their texture and mineral composition, because these rocks are formed by a repetitive process.

Millikan speaks of the causal relation's being "straight" from ancestor to descendant, and so one difference between reproduction and these other natural cycles might be that the latter are complex or indirect, whereas the former is more straightforward. But this can't be so, since genetic replication is just as complex and indirect as the other cycles. The similarity between generations of genes is brought about not just by the ancestral genes, but by the proteins that make up the replication machinery.[22] At another level, the similarity between generations of genes depends also

[22] Genes don't reproduce themselves, but are reproduced only by being run through a sort of factory with a workforce of several kinds of proteins. Helicase proteins separate the strands of a DNA helix, each strain is coated with SSB to prevent reannealing, RNA primer is synthesized by Primase, the new DNA strand is extended by Polymerase with the help of sliding clamps, the RNA primer is removed by RNase H, and the remaining short DNA strands are connected by Ligase. These proteins can be called the surrounding conditions, needed to derive the *in situ* law.

on the interaction of organisms (in the case of sexual reproduction). This certainly complicates the causal relation and thus the *in situ* law, with respect to the reproduction that Millikan says is needed at least for biological functions.

There is, however, one clear difference between the rock or water cycle and the sort of reproduction Millikan talks about, which is that the ancestors and descendants that reproduce in Millikan's sense are *tokens*, not types. The ink marks on each sheet of paper are reproduced by the photocopier, just as the genetic sequence carried in the cells of each token organism is reproduced by a battery of proteins. The similarities between generations of clouds, however, hold only between the properties that make up clouds in general, not between token or even subtypes of clouds. When the water particles of a cirrus cloud precipitate, the raindrops tend to return to an ocean, but the part of the ocean that evaporates and forms another cirrus cloud need hardly consist of the same token water molecules that once constituted those raindrops. Moreover, no cloud need have the same specific properties as those of the earlier token cloud whose molecules do constitute the later one. The same sort of point is true of rocks.

Now, I want to ask whether this difference between the natural cycles accounts for the functionality of biological effects. I think the answer is that this difference makes no difference. Granted, a reproductive process that shapes tokens rather than types is more fine-grained. Thus, Millikan can say that the existence of a token descendant's capacity can be explained in terms of an effect performed by a token ancestor of this descendant. For example, a particular heart has the capacity to pump blood because the heart descends from a heart that had this effect. Had some ancestral heart not pumped blood, there would have been no descendent hearts, with their distinctive properties

enabling them to pump blood. In other words, had the later heart not been produced, directly or indirectly, by a particular heart that actually pumped blood, there would have been no later heart. The existence of the descendant can surely be explained in these reproductive terms. But this doesn't mean the descendant is supposed to have the same effect as its ancestor. Just because there is a fine-grained reproductive process that produces members of a type, with the *capacity* to have the same effect as an ancestral member of the type, doesn't mean this effect is a special one *for the descendent members*. Just because an appeal to the earlier-performed effect is needed to explain the existence of the later members with the distinctive capacity, doesn't mean this effect becomes a function in any normative sense.[23]

The coarse-grained cyclical processes that produce clouds and rocks can't be used to explain why particular clouds and rocks have their more specific features. Thus, the existence of a type of cloud can't be explained in terms of the general cyclical process that produces clouds of all types. But were the processes more fine-grained, so that cirrus clouds, for example, would come indirectly from cirrus clouds, the relevant implication would be just that had the earlier cirrus clouds differed, the later ones would likewise have differed. There would be a reproductive process here in Millikan's sense, and so there would be an explanation, for each token cirrus cloud, that links its existence, with its distinctive properties, to an effect had by some earlier cirrus cloud. But were this effect to depend on the meeting of certain surrounding conditions, and were these

[23] A possible response is that the distinctive capacity of later members would be functional in a normative sense were the existence of these members to count as a kind of *success*. The problem with this response, for Millikan, is that the success would be subjective, depending on the *desires* of these members to survive and to reproduce. Thus, the functions would be prescriptively rather than objectively normative. Moreover, these desires would already have content, and so her explanation of content in terms of a proper function would presuppose certain other content, and so her explanation wouldn't be naturalistic in the sense of being informative, reducing talk of content to talk of something else.

conditions not met for a descendent cloud, there would as yet be no reason to say this cloud is supposed to do what its ancestor did. The historical explanation that refers to what the ancestor did, by way of explaining why there are descendants of the same type, and that refers to a reproductive process that links the members of the type, doesn't have any normative implication for any of the members. The clouds would be reproduced but they would have no function. What's needed for the normative implication is the additional, Wrightian assumption that explaining why something exists, or why it is there, is the same as explaining its function. This is the assumption I've criticized in section 4.4.

So just because the workings of the reproductive process that links the members of the type can be specified, in Millikan's theory, doesn't mean the historical explanation, as such, has normative implications for the historically determined capacity. A purposive function is an effect with a normative aspect, an effect that *should* happen, to fulfill a purpose, but appealing to something's reproductive history doesn't explain the thing's functionality. Neither does the distinction between Normal and abNormal conditions reductively explain descriptively normative functions. Thus, it seems the unconditional conclusion I spoke of earlier in this section is warranted: Millikan's theory of purposive functions rests on an analogy between natural and artificial functions and design processes, and implies that the prescriptive norm is the source of all genuine purposive functions. Assuming what I've argued, that there are no descriptive norms or at least that neither Millikan nor Dretske shows that there are, the so-called descriptive normativity of biological effects doesn't make these effects functions.[24]

[24] Strictly speaking, my objections above count not against the general notion of descriptive norms, but only against Dretske's and Millikan's assumptions that there are these norms. I assume, though, that

My arguments against Dretske's and Millikan's assumption that there are objective norms span most of this chapter and much of the last one. The main point is that they fail to show that there are these functions, and thus fail to explain content in terms of them. Indeed, they each posit prescriptive norms without meaning to do so. For example, a proper function is prescriptively normative to the extent that a trait's *having* of the function derives from the use of the trait as a sort of instrument to achieve the user's goal (see section 4.4). Also, the function is prescriptively normative in so far as what it is to be the function is to be *explained* in a certain way (4.6). Moreover, Millikan offers examples of what are supposed to be objectively normative functions, and in each case the norm is prescriptive or at least not obviously descriptive (4.3). It remains to show how Millikan's lack of a naturalistic theory of purposive functions affects her theory of content.

4.8 Intentional Icons and Representations

My above criticism of Millikan's theory of objective (non-evaluative) but purposive (descriptively normative) functions is that there are no functions in Millikan's sense. I now want to summarize how she uses that theory to explain how each symbol has its own content. I will then apply my criticism of her theory of functions to that use of her theory. Lastly, I will raise an internal criticism of her overall theory of content. The internal criticism is that her theory of the nature of intentionality conflicts with her theory of content determinacy.

Millikan's sort of Darwinian account of objectively normative functions, in particular, has the best chance of succeeding, since the theory of natural selection is the most successful account of how what appears to be mind-dependent, such as the design of species, is instead a mind-independent, objective process. Millikan's theory of functions is widely regarded as the most detailed such theory, so if hers fails, that does indicate that norms are likely only prescriptive.

Like Dretske, Millikan distinguishes between general and more specific types of symbols, or what she calls "signs." Although she says there is only a family resemblance between types of signs, she presents a theory of the content of what she calls "intentional icons," which she says are signs that have features in common with sentences and with more primitive signs such as beaver tail splashes and honey bee waggle dances. She defines a subclass of these icons, which she calls representations, and says that this subclass includes certain sentences and concepts. Thus, the types of signs that humans use are comparable to those used by other species. I'll summarize here her theories of intentional icons and of representations.

Millikan (1984) says that "Intentional icons are devices that are 'supposed to' map thusly onto the world in order to serve their direct proper functions; that is, Normally they do map so when serving these functions. And they are devices that are supposed to be used or 'interpreted' by cooperating devices" (95). An intentional icon Normally stands between two devices or systems, a producer and an interpreter, and these systems have a stabilizing function, or a "critical mass of cases of actual use" which keeps the producers and interpreters responding in standard ways (32). In the case of an indicative sentence, which Millikan treats as a type of intentional icon, a speaker uses a sentence to communicate with a listener, because the sentence, when spoken or otherwise made public, can be used, of course, by the listener as well. Without this cooperation in the case of communication, sentences wouldn't continue to be produced; that is, they wouldn't continue to exist as members of reproductively established families. According to Millikan, sentences are indeed such members, although the producer and interpreter devices themselves are learned programs.

A sentence has invariant and variant aspects, its basic syntactic form and the concrete words and syntactic form of phrases. In so far as sentences have the same overall syntactic form, they belong to the same reproductively established family, and the direct proper function of this form is to adapt the users to the variant aspect of the sentence, and thus to the content and to world affairs. More specifically, as Millikan says, "the Normal explanation of how the sentence adapts the interpreter device such that it can perform its proper functions makes reference to the fact that the sentence maps conditions in the world in accordance with a specific mapping function" (97). The syntax adapts interpreters to certain surrounding conditions in the same sort of way that a chameleon's skin adapts the animal to the colour of a specific surface. The skin has a general proper function of providing camouflage, but also an adapted, derived, or context-sensitive function of changing into specific colours, depending on the surrounding conditions. Likewise, the general function of a sentence is to facilitate cooperation between speakers and hearers, by mapping onto some set of external conditions, and the sentence's context-sensitive function is to permit this cooperation in specific cases, with some variation of concrete words, depending on the situation.

The mapping just reflects the fact that changes of the icon's variable parts can correspond to changes in world affairs. In her words, "There are operations upon or transformations (in the mathematical sense) of the icon that *correspond one-to-one* to operations upon or transformations of the real value [the mapped world affair]" (107, my emphasis). The mapping is facilitated by Normal mapping rules, such as rules for substituting nouns for other nouns in a sentence. These rules are Normal conditions that have to be referred to in a Normal explanation of the success, and thus the proliferation,

of the reproductively established family of intentional icons. For example, an English sentence has an invariable, syntactic part and variable terms, and sentences are reproduced, or repeatedly used from one generation of communicators to the next because, by using conventional mapping rules, a sentence adapts the interpreter to a situation. This is because, if the right terms are used, transformations of the sentence will correspond to transformations of part of the world. A Normal explanation of why there are descendent forms of sentences has to refer to this sort of correspondence between the sentences and world affairs.

Millikan (2005) contends that sentences and bee dances that correspond to world affairs are intentional because they "display the characteristic trait of the intentional; namely, they can be wrong or false. They can fail to correspond as they should" (97). A bee can waggle in the wrong way, sending the other bees away from the nectar, thus failing to fulfill the dance's proper function. The wrong terms can be used in a sentence, failing to fulfill the purpose of the communication of specific content. Thus, as far as Millikan is willing to generalize, the nature of indicative, as opposed to imperative, intentionality is found in an icon's being *true* with respect to a state of affairs, not in a term's reference to some element.[25] The intentionality of a whole claim is primary, and that of the parts of the claim, the terms, derives from their role in the claim. The reason for this, on Millikan's view, is that intentionality is a Normal condition of the

[25] Millikan (1984) says that "a referential term is supposed to appear in the context of a sentence," since this is "a Normal condition for its proper performance...A word, taken by itself, does not map or fail to map onto anything...The most basic or most direct kind of correspondence, then, is the correspondence between a true sentence and a world affair" (104). This doesn't yet mean that this kind of correspondence is the basic, or most general, kind of intentionality, but this is what Millikan maintains. Millikan (2005): "Portions and aspects of sentences that make a systematic contribution to truth-conditions can be considered to be intentional in a derivative way. That is, the intentionality of the complete representation which sports a truth-condition is prior to the intentionality of any of its parts or aspects. Truth-conditions are not built up from term references. Rather, term references are abstracted from truth-conditions" (n.1).

performance of a proper function, and what is used mainly, in an intentional icon, is a whole claim, to coordinate behaviour with the state of the environment. Millikan (2000) argues that indicative

> intentional representations always make claims. If they did not make claims, they could not be such as to guide activity appropriately given the existence of their extensions. About *danger*, for example, there is nothing to be done, nor is there anything to be done about *here* or about *now*. But about *there is danger here now*, there may well be something to be done. (198-9)

Millikan (1984) explains a more specific kind of intentional icon, which she calls a "representation." A representation is supposed to *identify* the elements to which the icon's terms refer, meaning that the representation is used as a way of understanding these elements. A bee dance maps part of the bee's environment, but bees don't thereby understand what is mapped, because the bee dance isn't a representation (96). Millikan (2000) focuses on a kind of representation, which she calls a "substance concept." A substance concept is, in part, an ability to re-identify, or to track, a substance, or something that retains some properties and thus, as Millikan says, some potentials for use, over numerous encounters with it.

> This makes it possible for the organism to store away knowledge or know-how concerning the thing [the substance] as observed or experienced on earlier occasions for use on later occasions, the knowledge retaining its validity over time...Substances are, by definition, what can afford this sort of opportunity to a learner, and where this affordance is no accident, but is supported by an ontological ground of real connection. (2)

Millikan argues that some external signs, such as spoken sentences, are representations in that they are used to identify their extensions. Indeed, she interprets natural language as a form of perception, as a way of picking up, focusing, and translating information in the

environment—just as a tree itself can be recognized from either a televised image or a spoken sentence about a tree.[26]

Without getting into the details of her theory of substance concepts, I want to call attention to two points. The first point is that she uses her theory of primitive content to explain the content of mental representations such as concepts. A substance concept isn't just the internal item that correlates with some external item, but is the ability to use such an internal item to fulfill a reproductively established purpose.[27] Second, she takes her theory of content to converge with a form of metaphysical realism, which is why she talks, for example, of substances and of the real value of veridical intentional icons. Indeed, the subtitle of her main book on content, Millikan (1984), is *New Foundations for Realism*, and even at the end of that book she works out an ontology that is compatible with her theory of content. The ontology is realistic, as opposed to antirealistic, in that an intentional icon can correspond to a situation not constructed by the icon or by its users. For example, an icon can correspond to a world affair or to the state of a substance, giving the icon a real, objective value. Thus, her realism makes at least two claims: there

---

[26] This raises the question of whether sentences are representations in her sense or whether they feed into an internal process of representation. Millikan (1984) suggests the latter: "for an outer term to precipitate an act of identification of its referent, it must be translated into an inner term that has the same referent, for the act of identification of the referent is an inner act" (198). Millikan's anti-Cartesian goal, though, is to show that there is no need for a sharp distinction between inner and outer acts. The later work, Millikan (2000), does this by explaining substance concepts, which are representations, as *ways* of identifying substances. Thus, she says that "learning a word for a substance can be learning a way to identify it." More specifically, she says, these ways are media for the manifestation of substances. Cameras, radios, microscopes, and spoken sentences are all ways of picking up, focusing, and translating information into a different medium, and when this is done for the purpose of understanding the information in such a way that it can be re-identified, the media are representational (89).

[27] As she says in Millikan (2000), applying her theory of content, in Millikan (1984), to her theory of representations, including substance concepts, "Intentional representations represent what they would need to for their interpreting mechanisms to use them productively in accordance with design." Moreover, "intentional representation requires only that there be a mechanism designed to produce items bearing a certain correspondence to the distal environment" (197). This correspondence, she says, "can be thought of as an abstract isomorphism" (198).

is indeed a mind-independent world, and we can have knowledge of this world because truth is a correspondence relation.[28]

Millikan's biological argument, that truth is the basic kind of meaning because in an evolutionary context terms by themselves rather than as parts of whole claims would be useless, provides independent support for her claim that intentionality is a kind of isomorphism. But this biological argument is weak. Her reasoning is similar to that used in anti-evolutionary arguments that some organic systems are irreducibly complex. For example, the bacterial flagellum might seem to serve no useful purpose on its own and to work only as part of a whole bacterium. Removing any part of the whole organic machine would cause the machine to break down. Thus, the argument runs, the flagellum couldn't have developed gradually, in a stage prior to the development of the whole bacterium.[29] One problem with this sort of argument is that it overlooks the possibility of exaptation: a precursor to the flagellum might indeed have increased the fitness of the bacterium's ancestor, under ancestral conditions, or might have been a spandrel, a temporarily useless byproduct of some more useful trait, and might then have been naturally selected under later environmental pressures, becoming the flagellum.

In the same way, terms might have been used prior to the development of the sentence, but for a purpose other than that served by the sentence, that is, for a purpose other than expressing truth for the coordination of group action. Moreover, this earlier purpose might have been semantically relevant, and so the isolated terms might have had

---

[28] Explaining the sort of realism she wishes to defend, Millikan (1984) refers to an opposing view that "entails rejection of correspondence...not only as the 'test of truth' but also as the 'nature of truth'—that realism, in one rough sense of 'realism,' must go" (6-7). Instead of being correspondence, truth on this antirealist view is a sort of coherence. Millikan's theory of content is intended to support realism of that kind, the kind that makes a claim about the nature of truth. If this kind of realism doesn't imply the more fundamental claim, that symbols relate to things that aren't symbols, she takes this latter claim to be part of the standard naturalistic picture.

[29] See Behe (1996).

content. For example, the terms, such as "wolf" or perhaps the mental representation WOLF, might have been used in such a way that they stood in for something else in a semantically relevant way. Alternatively, terms might have been spandrels that still substituted for something else even though this capacity wasn't at that earlier time used. There is relatively little known about the origin of language, and I won't argue here for a particular scenario. My point is only that just because a term on its own might be useless for the purpose fulfilled by a whole sentence, doesn't mean the sentence is semantically prior to the term. On the contrary, as I've argued, the reverse might be true, and there are evolutionary scenarios that make sense of this possibility. For this reason, I discount this particular argument of Millikan's, and maintain that her view of the nature of intentionality is motivated best by her metaphysical realism.

4.9 Proper Functions and Content Determinacy

Millikan's reference to the needs of explanation, in saying which conditions are relevant to something's purposive function, provides for a potential solution to the disjunction problem that on a causal-historical theory, content is indeterminate. Fodor (2008) raises the problem for a Darwinian theory of content:

> either natural selection is a species of 'selection for…', and is thus itself a kind of intensional process; or natural selection is a species of selection *tout court,* and therefore cannot distinguish between coextensive mental states. In the former case it may, but in the latter case it doesn't, provide an explanation either of the teleology or of the intentional content of the frogs' snapping. (4)

As long as frogs survive and reproduce because they tend to snap their tongues at flies, or because the little swirling dark dots in their environment tend to be flies, the function of

the snapping-mechanism can be to hunt either flies as such or little swirling dark dots. If a Darwinian theory, such as Millikan's, implies that the mechanism is specifically selected *for* the job of hunting flies, and not just a side effect of something else that's selected, the theory employs an intensional notion, and so can't explain the intensionality of ascriptions of mental states. Otherwise, a Darwinian theory can't distinguish between mental states that have different intensions but overlapping or identical extensions, under a description of these states.

Millikan can reply as follows. Proper function is determined not just by the actual reproductive process that builds members with a certain trait, but by the *best explanation* of the existence of these members. Frogs exist because the tongues of their ancestors snapped at flies and not just at certain dots; frogs had to eat to survive and to reproduce members, and it's flies in particular that would have fed them. Thus, the ingesting of flies as such is a Normal condition of the proliferation of the frog's snapping mechanism, since reference to that effect is needed in a Normal, etiological explanation of the existence of this mechanism.

The problem with this response is that what determines which explanation is best is in part a set of prescriptive epistemic and pragmatic norms, and the normative aspect of biological functions can't be descriptive if these functions are—as I argued in section 4.6— determined in part by Normal conditions, which are those picked out by a prescribed explanation. Recall that Millikan defines "Normal condition" as a condition that plays a certain role in a Normal explanation. Suppose Millikan were to say instead that a Normal condition is whatever condition plays a role in the objective reproductive process that accounts for the existence of the descendant produced by that process,

regardless of whether this process is understood or explained. In this case, Millikan would lack the above response to the disjunction problem: content would be indeterminate, assuming there is nothing out in the world by itself to favour one over another way of understanding the conditions of a certain evolutionary process.[30]

I should add that this question of how a symbol's content is determined isn't just the epistemic question of how a theorist *knows* which content is involved. The problem for an etiological theorist is that the historical facts may underdetermine the content of symbols used by an evolved system. Granted, the dots swirling around protofrogs may have really been flies, regardless of whether there is anyone to understand how this fact came to be relevant to the existence of later frogs with a mechanism for hunting those swirling objects. A realist is surely right to say that the act of explaining a process doesn't necessarily affect the process. But the question here is how a symbol bearing a semantic relation to some set of objects is determined, and this semantic relation has an intensional aspect. The frog's prey is both a fly, a type of living thing, and more generally a small, swirling dark dot. Even were there no one to explain how the frog thinks of its prey, there would have to be some such way of thinking on the frog's part, assuming the frog has a symbol bearing a semantic relation to flies or to certain small dots. On the more realistic version of Millikan's etiological theory, which doesn't refer to the theorist's need to understand the reproductive process, there seems no reason to say that the frog thinks of flies (by way of the frog's own conception of them) rather than little swirling dark dots or that the frog thinks in the inverse way. However, saying that the frog has one thought

[30] Another problem is that frogs are unable to distinguish between flies and little, swirling dark dots, so that the Normal explanation attributes to the frog a symbol with content that is too specific to be useful to the frog. An externalist about content, such as Millikan, can maintain, though, that a symbol's content doesn't depend on the symbol-user's dispositions or ability to identify the referent.

rather than another, as dictated by a theorist's standards of explanation, such as the standard that prescribes the comparison of natural selection to an optimizing designer who selects *for* certain traits, makes the content of the frog's thought dependent on the theorist, which is itself problematic.

Moreover, the point isn't the trivial one of picking between at least two ways of thinking about the very same thing, since in this case the extensions of the two possible mental states aren't identical: not *all* small swirling dark dots are flies. The question isn't just the psychological one of how frogs think of their prey or indeed of how theorists think of the objects. Instead, the question is which set of objects helped determine the content of the descendent frog's thought about its prey, the set of what we call flies or the broader set of small, swirling dark dots. The problem for Millikan is that members of both sets of objects were at work in the evolution of frogs, so the content of the frog's thought is underdetermined on the unpragmatic etiological theory of content. There are at least two ways of understanding the frog's evolution, and the historical facts themselves don't favour one way over the other. Again, if prescriptive standards of explanation are needed to determine the content, the frog's thought about its prey isn't the product just of the evolutionary process that preceded the theorist's attempt to understand this process. This is to say not that content must be prescriptively normative, but that content determined by a proper function would seem to be so, since such a function would be so.[31]

[31] As I argued in section 4.6, without some prescriptive norms such as those that govern an act of explanation, a condition met when the ancestors of some reproductively established family have a certain effect has no relevance to their descendants. It's the act of explaining something about the descendant, such as its existence, that unites descendant and ancestor in the way that justifies talk of the descendant's function. Talk of the objective, causal-historical connection between ancestor and descendant has no normative implications of its own, and thus no functional ones.

I'd like to go over again the problem of content determinacy for Millikan, in more general terms. Millikan (1984) uses her theory of proper functions to naturalistically explain the determinacy of content. Given her theory of functions, she says, a false sentence proves "no more problematic than, say, the color pattern of a chameleon that is 'supposed to' match what it sits on, but doesn't" (89). The chameleon's skin has its function regardless of whether a change in the skin succeeds or fails to perform it, because either way the change is supposed to have some specific effect. But my criticism of her theory of functions, given in sections 4.3 to 4.7, applies to this use of her theory: if there are no descriptive norms, and she doesn't account for prescriptive norms, Millikan has no naturalistic explanation of how content is determined. What she wants to say is that a certain state of a token chameleon's skin is related to some effect even when it's not actually so related, because the state has an objectively determined proper effect. But this assumes there are objective, descriptively normative functions. Without these functions, all Millikan can say is that some state of the skin has a result that depends, in part, on the surrounding conditions.[32] Likewise, if a sentence or any other sign has no

[32] Given a CP law about the skin type, what could be said is that counterfactually, were conditions different, a state of the skin would have some effect, such as successfully camouflaging the chameleon, even though under other conditions, such as actual ones, the state does not have this effect. There are pragmatic and realistic interpretations of nomic relations. On a pragmatic view, what the skin's state would do under certain, ideal conditions would count as the state's own, proper, or essential effect only were the CP law part of a *good explanation*, and the standards for this explanation would have some role in determining the typicality of the skin's camouflaging of the animal. In this way, prescriptive norms would determine the typicality of the effect, and thus its functionality, by determining in part what counts as a good explanation.

On the more realistic view that only conditions and nomic relations in the world determine the difference between biological types, and the act of explaining what happens isn't an essential part of what is thereby explained, what happens under actual conditions wouldn't be bound by what happens under different conditions. There would simply be the comparative point that different events happen under different conditions, and also the statistical point that some set of conditions might be the most common and thus *called* the set that determines the nomic relation or the function. The realist needs something like the highly problematic idea of normative universals, or of abstract entities that are *better than* concrete individuals, to support the claim that what happens under actual conditions is *objectively subject to* what would happen under other, ideal conditions.

purposive function, for the same reasons, Millikan has no explanation of how a "false" sentence nevertheless has a fixed semantic relation. Without a standard for a token sentence, such as a stage in a process that reproduces the sentence, the sentence has only its actual and its possible relations to world affairs, the latter being just its relations under different conditions.

Granted, a chameleon with skin that doesn't protect the animal may actually descend from a chameleon with skin that protected the ancestor, and picking out this historical relation between the chameleons' skins may be needed to explain how the descendant came to exist by a reproductive process. But if this historical explanation has no normative implications, there is no reason to say the descendant is, as it were, bound more by the past meeting of so-called Normal conditions than by the later meeting of so-called abNormal ones, the ones that surround the descendant. There is no reason to say the ancestor's effect is the descendant's own, proper effect. The two skins may be members of the same reproductively established family, but without a norm governing this family—a stern parent, as it were—there is no reason to say one member is bound by what another member does even when the members face different environments. Ancestral effects must be normative for them to be relevant to explaining a descendant's own, proper behaviour.[33]

An historical perspective doesn't reveal what a reproduced trait is objectively supposed to do. Granted, a trait can be explained in terms of the Normality, or descriptive

[33] A descriptively normative function seems comparable to the magic that is supposed to justify certain tribes' worship of their ancestors, in that in both cases the past is treated as though it had a magical impact on a later time. The natural reproduction of a trait isn't magical, but the way a descendent trait is *supposed* to have some effect, given the past-performed effect, together with different later conditions and no prescriptive norms, is indeed inexplicable. There just seem to be no objective, non-evaluative conditions under which an effect fulfils something's purpose or is supposed to be performed.

normativity, of some of its surrounding conditions, and thus some of its effects. In the case of an etiological theory of biological traits, what seems understood in this theory is that a biological trait is comparable to an artifact. This analogy may have explanatory value, given some prescriptive epistemic and pragmatic norms from which the so-called descriptive norms derive.[34] Were the analogy between the biological and the artificial shown to depend not just on the interest in explanation, but on a similarity between the *causes* of biological traits and of artifacts, a similarity that accounts *reductively* for any norm that governs their behaviour, the traits would indeed have objective functions. But I don't think Millikan shows there are any such similarities. She says natural selection and the design of artifacts are both kinds of selection, but she doesn't show that natural selection by itself gives rise to functions in any normative sense. The part of her theory that comes closest to accounting for non-derivative descriptive norms is the Wrightian assumption that explaining why something exists is the same as explaining any function the thing has. In section 4.4, I argued that this assumption presupposes the work of prescriptive norms, so that once again there is no accounting for those descriptive norms.

4.10 Proper Functions and the Nature of Intentionality

In section 4.8 I summarized Millikan's view of the nature of intentionality. What she says, roughly, is that the relation of aboutness is a Normal condition of the performance of certain proper functions, such as the function of communication. In the case of linguistic symbols, the relation is one of isomorphism, since this relation has to be

[34] One such norm would be that anthropocentric metaphors are to be favoured, since they can be used to explain the less in terms of the more familiar.

referred to in a Normal explanation of the reproduction of these symbols. As it stands, though, this teleological account of intentionality is circular, since the relevant Normal conditions are said to be those needed to explain not just any function, but a function that is already defined in terms of the use of symbols. Saying that the isomorphism relevant to intentionality is the kind needed for *communication* presupposes some notion of intentionality, unless the notion of communication can be explicated without reference to the use of symbols. That is, there's no reductive account of intentionality if intentionality is a condition of communication, and communication is just a certain use of *symbols*, which have intentionality.[35] The relevant function might be more general (and less semantic) than communication, such as the function of cooperation between producers and *interpreters*. Again, though, the notion of an interpreter seems to presuppose the notion of the use of symbols; indeed, interpretation is just the explanation of something's *meaning*, and even the notion of explanation makes little sense without the notion of content.[36]

Perhaps the relevant sort of isomorphism, or correspondence between sets of changes, as Millikan specifies, is just the sort needed for *cooperation*. But then Millikan no longer has a sufficient condition of intentionality. Take any reproductively established biological family. Most of the descendants have the same phenotypic structure as their ancestors, and reference to that sameness is needed to explain the way all of these members cooperate, in the broad sense of working together, to proliferate the genes that

[35] And were communication the use of signs carrying informational rather than semantic content, semantic relations wouldn't be Normal conditions of performing the proper function just of communication.
[36] Millikan (2000) says, for example, that information "embodied in an *intentional* representation is produced or channeled in accordance with the proper functioning of some designed mechanism, where a further proper function of that mechanism is to cooperate with a corresponding 'interpreting' mechanism to guide that interpreter in accomplishing some (ultimately practical) function" (197). This takes the relevant designed mechanism to be one engaged in a sort of interpretation, but this makes for only a circular account of the intentionality of representations.

produce the phenotype. Changes of the ancestral structure would correspond with changes of the descendent structure, since the two are linked by a reproductive process; at least, counterfactually, were the ancestral structure different, but still such as to make the ancestors fit their environment, the descendents would differ in a corresponding way. But this correspondence between ancestors and descendants wouldn't be a case of intentionality.

Instead, the kind of correspondence that is relevant to intentionality is the relation that holds between a world affair and a suitably arranged set of *symbols*. This is to say that Millikan probably reverses matters when she says that the content of terms is derived from the content of a whole claim.[37] The notion of truth presupposes the notion of terms, and thus the notion of intentionality, not the other way around. Take the example of the bee dance. The isomorphism between the dance and a certain world affair is only a byproduct of a special use of the dance. Whatever the terms or variant parts of the dance are, the dance is used as a map of an environment, and to this extent the dance stands in for what it maps. This asymmetric relation of substitution, of one thing standing in for another, which holds for each term, is more central to intentionality than is the symmetric correspondence relation, which Millikan thinks holds when the map succeeds. What needs to be explained is the way something can be used as a stand-in for something else. I offer an explanation of this relation in Chapters 5 and 6.

---

[37] Millikan is far from alone in taking a sentence's truth value to be the primary semantic property. Frege is often interpreted as saying that a sentence is the primary bearer of content, and Quine and Davidson have taken the same view. I take the opposite view, that the way a certain part of a sentence, such as a noun, is about something is the deeper fact that a theory of content needs to explain. To some extent, the difference between these views doesn't matter for my purposes, since when I propose my own account of content in Chapter 6, I focus on mental representations and I don't assume these representations have linguistic structure. Still, I take mental representations and terms to have the same sort of substitutionary aspect, which I think is crucial to their intentionality. These symbols are used as stand-ins, or as substitutes, for other things in such a way that they are said to be about these other things.

So that's one criticism of Millikan's view of intentionality. I want to turn now, though, to an internal criticism. On the one hand, Millikan says content is determined by proper functions. On the other, she says the nature of content, at least with respect to intentional icons such as sentences or mental representations, is a kind of isomorphism. I think these two claims are in conflict. Given the etiological theory of functions, it's unlikely that there is a mathematically one-to-one relation between an environment and the part of an organism used to identify items in the environment. There are evolutionary reasons to doubt, at least, that a mathematical function, taking input from the domain of an organism's possible states and giving as output states of an environment, is *surjective*.[38] For biological reasons that are consistent with Millikan's etiological theory of content determinacy, it's unlikely that a reproductively established trait can perfectly mirror its environment, in the mathematical sense of being isomorphic to it. The reason is that the Normal explanation, which is a crucial part of Millikan's etiological account of functions, takes the environment, that is relevant to an organism with a functional trait, to likely contain more than what the organism can identify.

To get a sense of my meaning here, take the relation between changes in the chameleon's skin and surfaces of the chameleon's environment. The chameleon's ability to mimic colour shades and patterns is limited, and so the chameleon isn't an ideal mimic. The evolutionary reasons for this are instructive: the chromatophores in the animal's skin cells, that change the skin's colour, also perform a communicative role, and are responsive to the chameleon's temperature and stress level. These other roles are likely to detract from the skin's ability to mirror all of its surrounding surface patterns.

---

[38] A mathematical (as opposed to a purposive) function F is surjective when for every element of the function's codomain, there is at least one element of its domain such that F maps the domain's elements onto those of the codomain.

The chameleon's capacity to mimic surface patterns is the result of evolutionary tinkering over thousands of years, and a naturally selected trait often has vestigial structures that impede its ability to perform its selected effect.[39] Moreover, chameleons are only as effective as they need to be to survive. In evolutionary terms, perfect mimicry doesn't mean that the possible states of an animal are in one-to-one correspondence with possible states of its environment. Instead, a perfect mimic is just one that can blend with possible surface patterns to such an extent that the mimic *tends* to be safe from predators so that the mimic survives and reproduces. Sometimes the mimic fails and doesn't survive, because its capacities for mimicry are limited compared to the possible changes of things in its environment. Creatures that are actual products of an evolutionary process usually have to struggle to survive, and so mimics don't always succeed, because their environment throws up new obstacles which the mimics aren't well-suited to overcome.

None of this is to deny, of course, that a reproductive process, such as natural selection, can build effective systems, such as teeth that can chew, hearts that can pump blood, and skin that can mimic colour patterns. A reproduced system may have to interact with the environment to increase an organism's fitness. For example, a bird's wings have to negotiate with air currents, and the bee dance might have evolved to map the distance and direction of a food source from a hive. For the systems producing the dance to have proliferated, and for the dance to work as a map, the dance must enter into some relation

[39] This is the case also with regard to honeybee dances. As the biologist O'Dea points out, bees rely on odour to track food sources. O'Dea (2000) hypothesizes that while the bee dance may contain information about the food's location and quality, this information isn't used by other bees, since the dance is an idiothetic movement used mainly to get the attention of other bees so that they can follow the odour trail brought in by the dancing bee. (An idiothetic behaviour is a miniature version of some behaviour, providing internal cues for dead reckoning.) Even if the evolved bee dance serves as a map, the dance seems to have vestigial effects. A dance used at least at one time for its internal cues or for its ability to call attention to the dancer isn't likely to be isomorphic to some relevant part of the environment. Instead, the dance will lack the resources to map all of this relevant part.

with the mapped variables. If "isomorphism" is taken loosely to mean similarity of structure, I think changes to a symbol system may be said to be isomorphic with possible values of the relevant variables in an environment. But if "isomorphism" is taken in the strict mathematical sense to mean one-to-one correspondence, which is the sense Millikan speaks of, I maintain that, for Millikan's own sort of Darwinian reasons, an organism's evolved symbol system isn't likely to be isomorphic with the organism's environment.

Millikan may seem to have a ready response to this objection. She can turn to the idea of an *ecological niche*. According to the ecologists, Begon *et al* (2006), a niche is the way "tolerances and requirements interact to define the conditions...and resources...needed by an individual or a species in order to practice its way of life." A niche "is defined by the boundaries that limit where it [a species] can live, grow and reproduce" (31). In short, a niche is what an organism needs from its environment so that the organism can perform its adapted role. Thus, Millikan might say that the *relevant* environment is defined by a niche, and this environment shouldn't contain more than what the organism can identify since the organism is adapted to this environment. The bee dance, for example, can't be expected to map parts of the bees' environment that are irrelevant to the bees' way of life. Instead, the bee dance is like a road map that focuses on useful, interesting features of an environment. Millikan is aware that correspondence relations can be trivially obtained, by redefining the domain and codomain, which is why she says that those correspondence relations relevant to intentionality are natural ones, meaning those obtained by applying Normal correspondence rules. In the case of sentences, there are rules for replacing nouns only with other nouns, and thus there must

be some assumption about which words are nouns. In a similar way, there may be a Normal mapping rule that stipulates what counts as an organism's environment. The correspondence is saved from arbitrariness if the rules for mapping are needed in a Normal etiological explanation, and ecological niches can supply these rules.

I take this point, that Millikan needs to claim only that changes to an evolved symbol system are isomorphic with changes to the relevant environment—relevant, that is, to explaining how the symbol system evolves. Now, the notion of a niche may no longer be viable in ecological studies.[40] In any case, the ecologically relevant environment isn't just that which is defined by the organism's niche, or by what is needed for what an organism likely succeeds at doing. The environment that is relevant to an etiological explanation of the diversity of organisms, for example, contains everything that accounts for the organism's *tendency merely to survive* and to reproduce. This is the Darwinian axiom that drives this sort of etiological explanation. Thus, from the etiological viewpoint, the relevant environment likely contains competitors and predators that threaten an organism's survival, and many of these opposing organisms have their own niches that define different parts of the same environment. Whenever different species or members of the same species—some of whom have mutations—relate differently to the same environment, this environment provides for multiple niches, or ways of life. A Normal explanation of the traits of one species usually has to account for certain inter- and intraspecies relations. So if each species can be expected to identify only the parts of an environment given by the species' niche, and yet a Normal, etiological explanation of the traits of that species usually refers to the wider environment

[40] Hubbell (2001), for example, argues that ecologists should explain biodiversity from a neutral perspective from which all species have the same general needs, not just the specific needs that delimit niches.

defined by other niches, there is likely an asymmetry between this environment and what a species can identify. Indeed, the point about relations between organisms with different niches is part of the modern definition of "niche," given by Hutchinson (1957). Hutchinson distinguishes between a fundamental and a realized niche, between a species' possible ways of relating to environments and the species' actual, more limited ways, due to pressure from competitors and predators.

For example, supposing that honeybees can map the distance and direction of a food source, the bees might be tricked by an opposing species into following a certain trail. Suppose humans were to want honeybees to consume a special food that would alter some of their traits, by altering the DNA they pass on to their offspring. This could be done by sending out robotic bees designed to perform dances that entice the swarm to follow towards the specially prepared nectar. A Normal, historical explanation of the change of honeybee phenotype would have to refer to the introduction of robotic bees as a Normal condition of the reproductive process. But the honeybees would have no way of distinguishing between organic and robotic foragers. So the symbols used in the bee's map wouldn't capture all parts of the environment that are relevant from the etiological perspective. The same sort of point applies to the narrower case of whether the possible changes of a single symbol, or intentional icon, capture all of the ecologically relevant possible changes of something in an environment. Suppose that a bee dance can identify the direction, distance, and quality or quantity of a food source. And once again, an opposing species chemically alters the food source, altering the DNA of the bees that consume it and thus affecting the bees' reproductive process. Given that members of a species want to reproduce, or produce members of their own kind, this tampering with the

food source is ecologically relevant to the bees, but a bee dance would likely be incapable of discerning the change in the food. So changes to the bee dance wouldn't be isomorphic with relevant changes even to one relevant item in the bees' environment.

There's one other response I want to consider: take whatever mapping capacity a species does have, define the species' semantically relevant environment in terms of just that capacity, and the result is an isomorphism between the mapping capacity and the relevant environment. Now, if a *capacity* to map is just the *chance* an organism has of identifying something, there will at best be a potential isomorphism between what the capacity can produce and the codomain. After all, there may be a natural possibility that some bees, for example, have a way of dancing that tracks the alteration of the food source even though no bee actually has such a dance. So the talk of a mapping capacity is consistent with saying that there's no isomorphism between the map and the relevant environment, because the environment can include items the organism has only a remote chance of identifying. In other words, there might be an isomorphism here were there a CP law that under certain conditions, an organism does map any item in the environment that corresponds to the capacity. But if all that establishes the isomorphism, on the organism's part, is its capacity to map, there's no limit on whether the conditions are ever actually met or likely to be so. The mapping rule that follows from this CP law isn't a Normal or a natural one in that it doesn't take into account any established limits of the mapping capacity for a particular species, owing to the capacity's actual evolution. Such a mapping rule would be trivially obtained rather than informative of what actually happens in nature. So Millikan should want to take into account the evolved limits of a

mapping capacity, but once these are taken into account, talk of an organism's actual mapping ability must replace talk of any abstract potential to map.[41]

Suppose, then, the response is that there is surely an isomorphism between what an organism can actually map, given all of the evolved constraints on its mapping ability, and an environment defined in terms of just those constraints. I grant that there is an isomorphism in this case. But Millikan can't avail herself of this account of intentionality and still hope for the account to reduce talk of intentionality to talk of something else. This is because the evolved constraints on the parts of a fully-identified environment include the organism's hardwired symbols, symbols the organism tends to learn due to a hardwired learning process, and the interest in identifying distal types to respond appropriately to what confronts the organism. These symbols already have semantic content; otherwise, the notion that there is an environment defined just in terms of what an organism can *identify* or *track* is empty.

What Millikan would want to say is that an intentional icon has content in virtue of a proper function to establish a one-to-one correspondence between changes of the icon and changes of something else. I've already suggested, at the beginning of this section, that as an account of the nature of intentionality this account is at best incomplete, since the function can't be specified without presupposing that there are symbols that are already used as substitutes, and thus that already seem to have semantic content. Now, there is the same problem in specifying the changes of a relevant codomain that are supposed to be isomorphic to changes of a set of intentional icons. Saying that intentionality is isomorphism between changes of subsentential *symbols* (that already enter into substitutionary and thus semantic relations) and changes of some other type is

[41] This distinction is similar to Hutchinson's distinction between fundamental and realized niches.

circular, and so doesn't by itself make for an informative, naturalistic theory of content. Likewise, saying that intentionality is isomorphism between changes of intentional icons and changes of types in an environment defined in terms of a limited ability to identify types—surely by means, once again, of symbols with semantic content—is circular.

There is, then, a trilemma for Millikan's account of intentionality. Intentionality may be isomorphism, in which case the Darwinian assumptions of Millikan's theory of content determinacy imply there likely isn't any intentionality, because the codomain that is relevant from the etiological perspective tends to be larger than the domain given by an evolved ability to map. On the contrary, there is intentionality, so intentionality isn't isomorphism. Alternatively, the domain and codomain may be equal in size if the codomain is specified in terms of a set of evolved constraints that include a set of symbols and a preference for things to fall under them. These constraints would already have content, and so this account of intentionality wouldn't be naturalistic in a reductive sense. But Millikan wants such an account, and so she can't identify intentionality with isomorphism in this way. Finally, intentionality may be more than just isomorphism, in the sense of a truth relation, in which case Millikan's theory of the nature of content is at best incomplete. Indeed, in my view, a theory that identifies intentionality with one-to-one correspondence misses an important asymmetry of semantic relations, namely the way a symbol stands in for something else and not the other way around. In any case, in the next two chapters I want to focus on the substitutionary aspects of certain symbols.

4.11 Conclusion

I've argued that Millikan's reductive theory of reproductively established functions fails. The theory would have to explain prescriptively normative effects, using only descriptive terms. Millikan's theory doesn't do this. She claims to ignore any prescriptive aspect of purposive functions, but she presupposes prescriptive terms when she speaks of explaining why a trait exists, or of why a system has a certain capacity for its very own, and when she speaks of the relativity of function-ascriptions to a Normal explanation. Moreover, she doesn't show how any kind of norm arises from a reproductive process. She uses this theory of functions to explain the determinacy of semantic relations, and so this latter explanation fails. I've also criticized her view that intentionality is isomorphism, and argued that even if her account of content determinacy worked, this account would contradict her account of the nature of intentionality.

Moreover, even were there no such conflict, a truth conditional account of content passes over a symbol's substitutionary aspect, the way a symbol is used as a stand-in for something else. Indeed, neither does an informational account of content, such as Fodor's or Dretske's, account for this substitutionary aspect. In addition, these three theories all have internal contradictions, which I've exposed in my discussions of them and which I diagnose in the next chapter, tracing the contradictions to a strategy shared by the three theorists. I then propose a different naturalistic strategy, one that accounts for the substitutionary aspects of symbols and for the prescriptive norms that seem to determine semantic relations.

Chapter 5

Two Naturalistic Strategies for Explaining Content

5.1 Introduction

In the last three chapters I critiqued three influential naturalistic theories of content. In this chapter I offer a diagnosis of a defect they share. The defect lies in the theorists' strategy of combining a naturalistic metaphysics of intentionality with some account of how semantic relations are determined. Given that the determinants are prescriptive norms, and that such norms have at best a problematic position in naturalistic ontology, a theory of content that follows this strategy is sure to be internally inconsistent (see sections 5.2 and 5.3). I then argue that the prescriptive norms pose a problem not for any naturalistic theory of content, but just for any that follows this metaphysical strategy. I point an alternative strategy, which focuses on naturalistic methodology rather than ontology and which promises to be naturalistic and unified, and to account for the substitutionary aspects of symbols and for the norms that determine semantic relations (5.4 and 5.5).

5.2 Basic Naturalistic Ontology and Theories of Content

Each of the three theories of content I've considered has its own problems, as I've

tried to show, but I think they also all have a common failing, which is the approach they take. They each answer directly the question about the nature of intentionality, by using the resources of basic naturalistic ontology, and then find some way of answering the question of how symbols have determinate content, consistent with reference to those resources. By "basic naturalistic ontology," I mean the most general set of things that have to be posited to make sense of scientific knowledge, on a sympathetic, optimistic interpretation of science.[1] So, for example, nomic relations of some sort might be posited, since scientists speak of natural laws. Fodor, Dretske, and Millikan each use some element of a naturalist's basic ontology in explaining what semantic relations are, and then add an explanation of the relations' determinants. The underlying problem is that norms are needed to explain this determinacy, and there is apparently no place for norms in basic naturalistic ontology. Thus, the two answers each of the three theorists gives are in conflict.[2] I want to show here how these conflicts work in the three cases.

5.2.1 Fodor

Fodor (1998) says that "Meaning is information (more or less)" (12). This would be his answer to the question, "What is intentionality, anyway?" He identifies the relation between a symbol and its referent with a set of nomic relations that permits the

---

[1] Strictly speaking, naturalistic ontology includes everything that exists in nature at all levels, including perhaps prescriptive norms, at some emergent level of explanation. But "*basic* naturalistic ontology" refers only to what is fundamental in this ontology.

[2] As is clear in the literature, numerous problems have been raised for these three theories of content. Indeed, Fodor, Dretske, and Millikan have criticized each other's theories. These criticisms may or may not have left one of the theories in a better position than the others. My view is that, regardless, the criticism I'm raising here, and that I've raised in most of the preceding three chapters, affects them all equally and points the way to an alternative approach, which I'll take in the present chapter.

interpretation of the symbol as a signal indicating something else. When the symbol, "dog," is tokened, someone can learn that a dog is probably responsible, directly or otherwise, for the symbol's use, assuming there is a CP law that uses of that symbol are naturally dependent on instantiations of *dog*. Now, information is already naturalized in that it has a well-established place in naturalistic ontology. The work of early information theorists, such as Shannon (1948), was concerned with engineering problems of long-distance communication. The point was to quantify communication in terms of what can remain constant despite a change of media, such as a conversion of sound waves into electric currents running through a telephone wire. Information theory was instrumental in developing certain technology, and this technology was made possible by the understanding of certain natural regularities, or nomic relations. By saying that meaning is information, Fodor is unifying something relatively rare with something much more widespread, that is, symbols and their intentionality, with the natural dependency of some properties on others. I want to emphasize that the appeal here is to something in basic naturalistic ontology, namely to the nomic dependencies that make possible the learning of often hidden reality from apparent, observed indicators.[3]

Fodor's broader theory of mind appeals to another part of this ontology, namely to the mechanism that implements a nomic relation. There must be a means by which dogs cause "dog" tokens. Thus, in the first place, "dog" symbols must be acquired, and so Fodor posits a locking mechanism by means of which this happens. Fodor's reasoning

[3] The appeal isn't to science itself, nor just to the way some types tend to be instantiated along with other types; instead, the appeal is to information, which is the content of a signal sent on the basis of a natural dependency among types, where this dependency is the sort of thing that can be understood and exploited. There would be no informational content in a property that no one could understand as being nomically related to some other property. The anthropocentric aspect of information is due to the original, practical goals of information theory.

here is that of a computationalist who assumes that symbols are efficacious. The idea is that a mind is a sort of computer program, such that the physical workings of the neural hardware are semantically interpretable. Symbols can cause bodily movement because symbols are physical particulars that are also organized in a higher-level way.[4] A scientific ontology doesn't include necessarily uninstantiated relations between abstract properties; these nomic relations must be realized in a way that provides empirical evidence of them. The locking mechanism is supposed to be a biochemical process that accounts for species-relative ways of thinking about external types on the basis of perceiving some stereotypical features of some objects. This is the sort of process of which there could be empirical evidence gathered in an experiment. And so Fodor's ontology must include not just free-floating nomic relations, but mechanisms that realize these relations and indeed emergent levels of realization. While the locking mechanism isn't meant to be part of Fodor's theory of content, the latter theory's computational and other naturalistic assumptions imply that there is a process by which the physical particulars are acquired. Symbols must be physical particulars for them to enter into syntactic relations, but at the higher, psychological level, symbols are used only after they originate in some fashion. According to Fodor, primitive nonscientific symbols can't be learned, but there is a mechanism by means of which a type of creature acquires these symbols with certain content.

---

[4] As Pylyshyn (1984) says, "what the brain is doing is what computers do when they compute numerical functions; namely, their behavior is caused by the physically instantiated properties of classes of substrates that correspond to *symbolic codes*. These codes reflect all the semantic distinctions necessary to make the behavior correspond to the regularities that are stateable in semantic terms. In other words, the codes or symbols are equivalence classes of physical properties which, on one hand, cause the behavior to unfold as it does, and on the other, are the bearers of semantic interpretations" (39).

To summarize, Fodor's answer to the question, "What is the nature of intentionality?" can be given as follows. Intentionality is a kind of information. This information is what a signal indicates about its source, due to a nomic relation between the property of being an instance of some type, such as *dog*, and the property of causing a type of symbol token, such as "dog." The nomic relation holds between properties about which there tends to be empirical evidence found by observing how the nomic relation is realized by a mechanism. One such mechanism is the locking mechanism accounting for how symbols are acquired. All of these parts of the answer to the question draw on basic naturalistic ontology.

Fodor's theory of content addresses also a second question, "How is the content of a symbol determined?" In trying to account for what I've called the metaphysical robustness of content, Fodor appreciates this question's difficulty. To see this, consider the parallel question, "How is a physical relation, such as *X*'s being heavier than *Y*, determined?" Take, for example, a watermelon's greater weight than an apple's. There is an epistemic question of how it's known that the one fruit is heavier than the other, since someone might make a mistake and think the apple is heavier. But this isn't the question at issue, of how the relation is determined. The relevant question is what makes one thing weigh more than another, so that the general relation, *Hxy*, can be differentially instantiated. Assuming the relation is individuated, so that it differs from other relations, the relation must have limits that apply differentially to whatever bears the relation to something else. Only were there such limits would the epistemic question make sense, since there can be no mistaken measurement of a relation that has no distinguishing features. Now, that which sets the limits of *Hxy* is the force of gravity acting on a mass,

since weight is the extent of the force that must be applied to an object to hold it at rest in a gravitational field. Roughly speaking, this physical relation is determined by how things interact in space, given certain properties of space and of the interacting things. The important word here is "interact." Given the physical principle of locality, any macroscopic physical relation is a form of interaction between certain things, and so certain conditions of interaction must be met for the relation to be instantiated. Moreover, when these conditions are met, the relation *must* be instantiated, but when only some of the conditions are met, the relation can't be instantiated. In a similar way, a game of hockey, defined by conventional rules, can take place only between two teams of potential players, and when these players interact in a certain way, the result is necessarily a hockey game, but when only one team is present there can be no such game.

Semantic relations aren't physical or interactive in that sense. As Fodor points out, a symbol token can have a wide variety of causes—indeed, so many causes that explaining the semantic relation requires more than even a CP law allowing for variation due to interfering systems: at least two CP laws are needed to account for the robustness of content. A semantic relation must be strong enough to remain standing, as it were, despite the lack of any pattern among the symbol's potential causes. A hockey game might be similarly robust were the game playable by a team of human players against a team of squirrels, raindrops, or virtually anything else. The robustness of semantic relations is found also in the possible lack of any interaction at all between a symbol token and an item in its extension, as in the case of a symbol about a fictional character. These differences between semantic and physical relations are what give the appearance that the limits of a semantic relation are set by a prescriptive norm. After all, a mere

prescription or recommendation, as it were, is applicable only when the interaction at issue need never occur. For example, even were no one ever perfectly rational, it might still be that people ought to be so rational. Rationality might be prescribed, and people might bear a relation to rational actions, by way of having an epistemic obligation to perform them, even though the actions are never performed and there are no conditions under which people would be physically compelled to perform them. That is, there are no objective conditions that would necessitate the performance of an action that is merely prescribed as opposed to being scientifically predictable.[5]

The way in which a symbol can be about something even though the symbol never physically interacts with the thing is like the way someone may be rationally or, for that matter, morally obliged to perform some action even though the person never actually performs it and there are no circumstances under which the person is physically compelled to do so. In either case, the relation is robust in the sense that the relation can be instantiated even without all of its *relata*. When the physical conditions are not all met for one object to be heavier than other, it cannot be that the one is nevertheless heavier. But even if a fox causes a "dog" token, or a symbol is about a nonexistent object, or someone is obligated to perform an action the person would never perform, the symbols have content and the person has the obligation. The apparent work of norms on semantic relations is one of the challenges of explaining content in naturalistic terms.[6]

---

[5] For this example's sake, I'm assuming a non-instrumentalistic view of rationality. In so far as rationality is just an instrument, the point about rationality is that some means are better than others at achieving a goal and that, given that someone has the goal, a conditional, goal-dependent imperative declares that a rational person would choose the more efficient means of achieving the goal. Arguably, there is nothing irreducibly prescriptive about this imperative. By contrast, a non-instrumentalistic view of rationality addresses the value of the *goal*, as opposed to calculating the efficiency with which some means achieves some goal, and describing what rational people do, without recommending rationality.

[6] There may be physical laws about uninstantiated relations, but they pose no counterexamples to what I've just said, since were the conditions met for the instantiation of a physical relation, the interaction would

Fodor explains the robust determination of content, by positing a higher-order asymmetric dependency. Not only do "dog" tokens depend, *ceteris paribus*, on dogs, but the dependency of "dog" tokens on foxes itself depends on the dependency of these tokens on dogs. For each symbol type, there must be an arrangement between possible worlds, such that any possible world in which items not in the symbol's extension cause the symbol tokens asymmetrically depends on the world in which items in the extension cause the symbol tokens, and not the other way around. But this means there's a world in which *only* items in the extension cause the symbol tokens.[7] In effect, this world is what Fodor calls a Type One, or Normal, situation. The only difference between this solution and the functionalist's is that the functionalist, such as Dretske or Millikan, tends to say that the Normal situation is *historically* prior to the abNormal one, whereas Fodor implies that the one is *metaphysically*, or intuitively, prior to the other. According to Fodor, the setting in which only dogs cause "dog" tokens, and there is no mistaking of the cause's identity, need never be realized, but there must be an intuitive grasp that such a possible way of causing the symbol tokens is asymmetrically independent of the other settings in which non-dogs cause the symbol tokens. As it happens, though, Fodor's account of how symbols are acquired by a locking mechanism implies that this Normal situation is both realized for each acquired symbol type, and temporally prior to any case in which the acquired symbol is used, say, to misidentify its cause. The latter is so because any symbol used in a misrepresentation must first be acquired.

---

have to occur. That's what it means to say physical laws aren't mere prescriptions or commandments. Under the appropriate conditions, things have no choice, as it were, but to physically interact as set out by the applicable natural law. But there are no objective conditions under which a "dog" token must be caused by a dog, nor are there just such conditions under which someone must perform rational actions.

[7] Were foxes to cause "dog" tokens in any world in which dogs do, the worlds in which foxes cause these tokens wouldn't asymmetrically depend on the world in which dogs do, since a world can't asymmetrically depend on itself.

Thus, even though Fodor sets out to explain the independence of semantic and causal relations, to distinguish his theory from a functionalistic one, some of his ontological assumptions conflict with the claim that semantic relations are robust. Fodor affirms that symbol tokens can *always*, under any conditions, be caused in all sorts of ways, and claims that this is the main problem with theories such as Dretske's and Millikan's. According to Fodor, these other theories are unable to explain the fact that a semantic relation isn't a causal one, since these theories pin a semantic relation to the causal relation instantiated under conditions favourable to the performance of some purposive function. But Fodor also claims that symbols are physical particulars that enter into nomic relations, and that any special nomic relation, corresponding to a CP law, is realized by some mechanism. This mechanism is a complex system that behaves according to the law, under certain conditions, such that there are empirical reasons for asserting the law in the first place. The locking mechanism is a reliable way of having the symbol token, that becomes a "dog" token, be caused only by dogs, assuming dogs are the animals that tend to have the surface properties that trigger the mechanism and cause the symbol's first tokening in someone. Even were some non-dogs to have exactly the same surface properties as dogs, under conditions that are ideal to perceiving dogs, which seems unlikely, at best, the content of the acquired symbol would still be mechanically determined. So on the one hand, on what I think is the best interpretation of it, the robustness claim is that there *can't* be a Normal situation for the tokening of a symbol, since a symbol can *always* have a variety of causes. And Fodor should be taken as affirming this robustness claim. On the other hand, the same set of assumptions Fodor

uses to explain what intentionality is lays out just such a Normal situation for each symbol type.

I think the root problem here is that, while semantic relations appear to be determined by prescriptive norms, there is no obvious place for such norms in a basic naturalistic ontology that includes only what exists at the most fundamental level; instead, from that ontological perspective, norms must be reducible to something descriptive and objective. A prescriptive norm is a subjective standard that is supposed to be followed even though it may never actually be followed. But there are only objective entities in a naturalist's basic ontology, and so anything that appears irreducibly subjective and value-laden has to be explained as though it were entirely objective.[8] The problem is that a so-called descriptive norm isn't entirely objective; indeed, to the extent that this norm is normative, and that it's a standard that need never, but still should, be met, this norm is given by a prescription (see sections 4.3 and 4.4). In effect, then, the appeal is to a prescriptive norm which is made to seem as though the appeal were to something else. But since prescriptive norms aren't well-established in a materialistic ontology, a theory that posits so-called descriptive norms will conflict with one that makes explicit use of this ontology.

In explaining content determinacy, the later Fodor doesn't refer explicitly to descriptive norms, nor does he address how an asymmetric dependency between nomic relations might account for the apparent normativity of content determinacy. But on what I've called the metaphysical interpretation of it, the robustness claim introduces the normativity of symbol use, and the asymmetric dependency claim appeals to objective

[8] I'm assuming a naturalist who addresses metaphysical questions is, roughly speaking, a monist, taking matter to be the primary substance. Minds and their subjective properties, such as consciousness, values, or prescriptive norms must therefore be fundamentally material or else somehow illusory.

relations that are supposed to account for the robustness.[9] There is no such account. Instead, using the limited resources of basic naturalistic ontology, Fodor speaks of a certain relation between nomic relations which requires that there be, after all, a Normal situation for the causing of any symbol token. His theory of how symbols are acquired posits a mechanism that works precisely in such a Normal situation, when conditions are met for items in a symbol's extension to cause the symbol token. Thus, these theories of Fodor's amount to just the teleological story about symbols against which he argues.

Moreover, these theories conflict with each other, as can be shown by the following. The asymmetric dependency claim is an answer to the question about content determinacy. Fodor states this question in terms of the robustness of content, a statement that has the normative implication that some uses of a symbol are correct even though they need never, under any set of conditions (including so-called Normal ones), occur. The theory of the locking mechanism posits a set of conditions under which one of these uses of a symbol *must* occur, given a CP law satisfied by the mechanism's work. Were there no such conditions, Fodor would lack a solution to the doorknob/DOORKNOB problem, which is the problem of why the objects in a symbol's extension tend to be the

---

[9] After laying out his theory of content, Fodor (1990) adds that a complete theory should account for the wrongness of some uses of symbols. He distinguishes between the normative aspect of symbol use and the robustness claim that "captures the point that some ways of using symbols are ontologically dependent on others" (128). Were this summary of the robustness claim accurate, the claim would be equivalent to the asymmetric dependency claim. And were this so, accounting for the robustness couldn't be a causal theory's *desideratum*, whereas Fodor also says otherwise. So there must be more to the robustness claim.

What I've called the metaphysical, as opposed to the empirical, interpretation of this claim is that a Normal situation for symbol use is impossible, since a semantic relation isn't any causal relation. This interpretation implies that a semantic relation need never reflect any of the causal relations between a symbol and anything else. But this is just another way of stating the point about the normative aspect of symbols, that the uses of a symbol are subject to a standard which need never be realized even though some of the uses are correct or incorrect according to the standard. Were a prescribed relation just a causal relation that must be instantiated under Normal conditions, the relation wouldn't be prescriptive after all; moreover, the relation wouldn't be metaphysically robust. Thus, I maintain that Fodor's point about the robustness of content does already introduce the normative aspect of symbol use, even though his asymmetric dependence theory doesn't explain this aspect and isn't intended to do so.

ones that trigger the mechanism for acquiring the symbol type. The conditions that enable the mechanism to lock to a distal type-determining property, on the basis of a perception of the object's surface features, provide for just the Normal situation that a teleological theorist can say determines the symbol's content. Thus, Fodor's robustness claim points to apparent prescriptive norms determining semantic relations, but his theory of content determinacy appeals only to an objective metaphysical dependency, and he doesn't show how this dependency gives rise to the prescriptive norms, or to the robustness of content. At best, his theory of concept acquisition appeals to a mechanism that might have an objective purposive function, but he doesn't explain any such function, and he takes ADT to be an alternative to a teleological explanation of semantic robustness.

Unlike Dretske and Millikan, the later Fodor doesn't claim to be offering a reductive explanation of prescriptive norms. For example, he doesn't explain the asymmetric dependency as the result of an objective purposive function. But his robustness claim amounts to the claim that there are such norms. Far from accounting for these norms, or for the apparent way in which content is determined, his theory of content determinacy, together with his theory of symbol acquisition, imply that semantic relations aren't metaphysically robust and thus aren't prescribed. In Chapter 2 I spoke of a conflict between these two theories, between ADT and LMT, taking the robustness claim to be assumed by ADT. More precisely, what might be said is that in Fodor's case, the internal conflict isn't between his two theories, but between his statement of the *explanandum* of his theory of content determinacy (the robustness claim) and an implication of part of his broader theory of mind (LMT), derived from basic naturalistic ontology. His asymmetric dependency theory is supposed to explain the robustness of content and thus the apparent

prescriptive norms that determine content. But this theory implies that content isn't robust in the metaphysical sense, that there must be a possible situation in which symbol tokens are caused only by objects in their extension. On top of this, his theory of symbol acquisition implies, not just that this situation is possible, but that for every acquired symbol there is a situation in which precisely that causal relation is instantiated, and that this situation is temporally prior to any situation in which the acquired symbol is used in a misrepresentation.

5.2.2 Dretske

Having just made some of my main points regarding the flawed approach to explaining content, I can be briefer in giving my diagnosis of the conflicts in Dretske's and in Millikan's theories, starting with Dretske's. Both Dretske and Fodor take semantic relations to emerge from information, and so the relevant part of Dretske's ontology also includes nomic relations. But Dretske emphasizes the need for symbols to have causal power in virtue of their content. He wants to give not just a naturalistic explanation of semantic relations, but a realistic one, according to which these relations must causally interact with the rest of nature.[10] Dretske's explanation rests on the distinction between structuring and triggering causes. When someone wires a garage door to close whenever the doorbell rings, the person's intention for this to happen is the reason why the bell causes the garage door to close, which makes the intention the structuring cause. Likewise, when a creature learns to use an internal indicator of *F* to satisfy an interest in

[10] Note that on Fodor's computational view of the mind, symbols have causal power in virtue of their syntactic, but not their semantic properties.

R, the indicating of *F* is the reason why the indicator *C* is linked to the movement *M* needed for the creature to obtain *R*. The relation between *C* and *F* is the structuring cause of the way the creature wires itself during the learning process, and this relation becomes semantic when *C* acquires a function of indicating *F*, thus allowing for possible misrepresentation of *F* under certain conditions.

For this explanation to be naturalistic, Dretske can't refer to any semantic relation on which the learning process might depend. In Chapter 3 I argued, however, that his reference to *C*'s initial receptivity to *F* does just this, since this receptivity is as intentional as is any desire *D*, on Dretske's own account (see sections 3.7 and 3.8).[11] Moreover, I argued, Dretske doesn't show that a symbol has any causal power in virtue of its semantic, functional content, since what is efficacious on his account is, in the case of beliefs, the mere information—*C*'s indicating *F*—that structures *C*'s relation to *M* prior to *C*'s acquiring a function and becoming a belief. This is the problem of local potency (3.5 and 3.6).

But having discussed purposive functions at some length, in Chapters 3 and 4, I can now offer what I think is the deeper reason for these problems with Dretske's theory. This reason, again, is that a naturalistic answer to the metaphysical question of the nature

---

[11] This criticism of mine assumes, however, that there is some evolutionary account of functionality. In Chapter 4 I argued there is no such account. Thus, Dretske could argue that *D* has semantic content, whereas the receptive internal condition *C* does not, because *D* acquires a function through learning, whereas natural selection doesn't equip any trait, such as *C*, with a function. I would respond as follows. If normative functions don't emerge from natural selection, they surely don't from learning by reinforcement. Thus, if there are no purposive functions due to natural selection, neither *C* nor *D* nor *B* would have semantic content. Dretske must assume that what he calls functions in the ordinary, and thus normative sense can emerge from a mindless process such as natural selection; otherwise, there is no reason to think they could emerge from what he assumes is the mindless process of the reinforcement of drives. Thus, he can't adopt the objections I raise against the biological account of purposive functions. So although I think there is a common underlying reason for the internal conflicts in Fodor's, Dretske's, and Millikan's theories, the specific objections I raise against them are independent of each other. After all, I'm claiming the conflicts are *internal* to each theory, even though it's instructive that there are such conflicts in *each* theory.

of intentionality restricts what can be said about the determinant of semantic relations. Specifically, there seem to be no prescriptive norms in basic naturalistic ontology; that is, assuming there is fundamentally the natural, scientifically explained world, such norms are at best illusory. But the limits of semantic relations appear to be merely prescribed as opposed, say, to causally determined. Thus, Dretske seems to think, there is a need for another kind of norm, for a so-called descriptive, value-neutral norm that does the same work as a prescriptive norm but that is compatible with what is found in basic naturalistic ontology. Although Dretske doesn't give a detailed discussion of the type of function at issue in his theory of content, he says the function is the ordinary purposive sort, as I point out in section 3.6. This means the recruited, functional *C* must be subject to some sort of norm or standard such that this *C should* do something under certain conditions. He doesn't argue that this norm is dependent on someone's interests or is otherwise subjective. Thus, this norm must be descriptive, not prescriptive. *But there are no descriptive norms*. Nevertheless, some norms seem necessary to explain how content is determined. Thus, in so far as Dretske's theory works at all, the norms and so the functions at issue must be prescriptive after all, even though there is no room for these on his metaphysical account of the informational nature of intentionality.

Specifically, in so far as *C* acquires a function upon being "recruited" by the system or creature *S*, this function must be somehow subjective. And this is the role of *C*'s initial receptivity to *F*. Prior to *S*'s learning how to behave, *S* wants *R*, and if *C*'s being subsequently linked to *M* gives *C* a function to cause *M* (because of *C*'s indicating of *F* which is needed for *S* to obtain *R*), this function must be like any artifact's subjective function. When *C* is linked to *M*, *C* becomes, in effect, an artifact with a function

dependent on *S*'s prior interest in *R*. Likewise, a hammer, for example, has a function only relative to someone's interest in doing a certain job. At least, this interpretation of the functional *C* as an artifact and a means of achieving the goal of using *R* is the only way to account for any normativity or ordinary functionality of *B* or of *D*, given Dretske's account of learning. Dretske needs for *C*'s function to be prescriptively normative, but an appeal to basic naturalistic ontology—and specifically his assumption that for something to be naturally real, the thing must have causal power—won't permit any explicit reference to this normativity.

With regard to causal power, a hammer interacts with the world in virtue of its physical properties. But the hammer's status as a tool, and not just as a physical object, is a matter of how the object is valued by a user. A tool can be better or worse at performing a task, approaching or departing from a standard, whereas a physical object, as such and on its own, can do nothing of the kind. Moreover, any causal power the tool has as a valued object appears subjective. For example, a hammer's trustiness doesn't add to the hammer's objective causal power, although someone who believes a certain hammer is trusty might be more inclined to use that hammer. A hammer's trustiness depends on an interpretation of experience with the hammer. Similarly, *C* as a neural particular has causal power, in its relation to *M*, but as a recruited, functional particular, *C*'s causal power must depend on some prior mental state, such as *S*'s receptivity to *R*. Reference to the true, subjective source of *C*'s functionality—to *C*'s prior receptivity to *R*—is only implicit in Dretske's theory, since this reference conflicts with Dretske's metaphysical assumptions about intentionality. Moreover, the structuring cause of *C*'s causing of *M* can't be just *C*'s indicating of *F* and the desire for *R*, since the full reason for the internal

wiring must include the receptivity. Thus, at least one of the structuring causes is the equivalent of someone's intention to use an artifact. Part of the reason *C* becomes connected to *M* is that *S* is already drawn towards *R*: the internal wiring becomes a means by which *S* reliably obtains *R*. What has structuring power on Dretske's theory isn't *B* or *D*, since these internal conditions are already structured, and in any case there is no such local potency. Instead, what helps configure *S* is *S*'s prior interest in *R*, and this is just as well since the only source of *C*'s functionality must be some such prior interest. This interest makes the function genuine, albeit prescriptive, goal-oriented, and subjective rather than natural and objective in Dretske's sense.[12]

I'll go through this line of argument again. When stating what intentionality is in natural terms, Dretske turns to basic naturalistic ontology: intentionality is information, which depends on nomic relations and on properties that make a real difference in the world due to their causal power. When stating how semantic relations are determined, Dretske turns to purposive functions, explaining these functions—at least superficially—as the result of a primitive type of learning, or a certain objective process. But there can be no such functions, and thus no such determinacy; there are no descriptive, objective norms or functions. Instead, the functions must be governed by prescriptive norms, and Dretske's account of how the internal conditions are recruited actually implies that there are those norms. But those norms won't make symbols efficacious in virtue of their semantic properties. Those norms are subjective and evaluative, and so anything governed by them, such as an artifact with a purposive function, could have causal

[12] This is slightly different from saying, as Dennett (1971) and Davidson (1980) do, that a creature's beliefs and desires are normative because they have roles in the rational project of getting what the creature wants. This assumes there are rational standards against which beliefs and desires are measured. The problem with Dretske's account isn't that it presupposes norms of rationality, but that it presupposes the subjectivity and instrumentality of the functions of beliefs and desires, because it presupposes a creature's receptivity.

power, as something with normative status, only in the user's view. That is, this causal power must be subjective and interpretive. *C* ought to cause *M*, from *S*'s point of view, because *S* is already receptive to using *R*, and *C*'s status as a recruit isn't reducible just to *C*'s prior objective indicating of *F*. *Of course* a functional *C*, such as *B*, doesn't make a causal difference just because of its physical capacity to be related to *F*; only as a neural particular does *C* have this causal power. *C*'s prescriptive, goal-directed function makes *C* a sort of internal artifact, and the structuring cause of this *C*'s role in *S* isn't any objective information, but the equivalent of a designer's interest in a task *C* can perform, namely *S*'s receptivity to the use of *R*. The causal power of anything naturally fundamental, figuring into a naturalist's basic ontology, is objective, since at this level what is real is material. Informational relations are efficacious in this way, since they reflect nomic relations between physically realized properties. But functional items lack this sort of efficacy, since their status as such is subjective, due to the subjectivity of the prescriptive norms that determine the function's limits. So as naturally real, an information-carrying signal can make a causal difference, while as functionally determined, a semantic symbol can make no such difference.[13] Thus the problem of local potency and the internal conflict in Dretske's theory of content.

5.2.3 Millikan

With regard, finally, to Millikan's theory, recall that she takes the nature of

[13] This is similar to the criticism of Davidson's account of the efficacy of mental events, found in McLaughlin (1993), according to which Davidson inadvertently makes these events out to be epiphenomenal, since what has causal power on his account is only the events in virtue of their physical properties.

intentionality to be isomorphism, since this is the nature of truth, given a correspondence theory of truth. Her best reason for claiming this is that this a view of intentionality converges with a form of metaphysical realism that is independently supported. In so far as the best interpretation of scientific knowledge is realistic in Millikan's minimal sense, once again Millikan's claim about the nature of intentionality seems to draw on basic naturalistic ontology. Whereas Fodor emphasizes the need for mechanisms that realize special nomic relations, and Dretske emphasizes the need for naturally real objects to have causal power, Millikan emphasizes the potential for realistic knowledge of the natural world. And once again, semantic relations seem prescriptively determined, whereas a naturalistic account of content, on this first, flawed approach, can't appeal to prescriptive norms. Millikan argues that there are descriptive norms, but her arguments fail in my view and she appeals, after all, to prescriptive norms. In particular, she follows Wright in regarding the ascription of a purposive, "proper" function to be an answer to the question of why something *has* the functional feature, but this amounts to taking the feature to be owned and thus used as an intended artifact. Also, she takes Normality to depend on a type of explanation, and thus on the prescriptive norms that govern any explanation.

These two types of prescriptive norms play different roles in the conflict between her theory of content determinacy and her theory of the nature of intentionality. On my interpretation of Wright's analysis of function-ascriptions, the key idea is that something must *have* its functional feature, where this idea is put forward in answer to the question, "Why is that feature *there*, or why do those things exist with that feature?" But only one sort of *having* can account for any such functionality, and that is the user's ownership of

an artifact. Only in that case is there clearly a function at issue, and only by extension from that case can an answer to the above question be counted as a function-ascription. Of course, any such function is prescriptively normative, which is to say, interest-relative or otherwise subjective. So if a heart has the function of pumping blood, the heart must *have* the capacity to pump blood, and this having must be derivable from some user's intention, such as that of the organism with the heart. This ownership of a trait is comparable, say, to someone's having of a hammer which gives the hammer the purposive function of performing whatever effect must be performed to satisfy the user's interest.

Were this analogy pushed further in the case of biological functions, the user of the functional features would have to be either the organism with the trait or some designer of the organism. Assuming some organisms have functional traits even though the organisms don't own their traits in the sophisticated way needed for the traits to become, in effect, artifacts, the owner must be some sort of intelligent designer. But this is ruled out by the second sort of prescriptive norm in Millikan's account, namely by the norms of etiological explanation which determine the relevance, or Normality, of conditions to a trait's function. Millikan would argue that these epistemic and pragmatic norms are naturalistic, and so a Darwinian explanation of the evolutionary process is best. There should still be a user of an organism's functional trait, since were there no such user the trait would have no function. The implication, then, is that the user is a mindless tinkerer, or selector of genetic building blocks, instead of an intelligent designer.

But this implication is in conflict with Millikan's answer to the metaphysical question, "What is intentionality, anyway?" Her answer, as given above, is that

intentionality is isomorphism. As I've shown, a mindless tinkerer, such as the process of natural selection, won't likely build a system that is isomorphic to its environment; instead, there will be an asymmetry between what is included in the relevant environment, and what an organism has the resources to identify in the environment. From the etiological perspective, the relevant environment is the one in which an organism is adapted to survive and reproduce. Thus, an organism must be able to identify *at most* whatever it needs to occupy its niche. But there are still no grounds for saying that an organism is likely able to identify *everything* in this niche-defined environment. This environment is also just a place in which the organism *tends merely to survive* and reproduce, and thus is a place in which the organism can be affected by something it can't identify. From the etiological perspective, the mindless tinkerer that adapts organisms to certain environments is frugal, as it were, meaning that organisms are given just enough ability to tend to survive long enough to reproduce. This is because a mindless tinkerer builds not what is ideal, but only what can be built under actual pressures. So an organism likely lacks the ability to identify everything in the environment to which a Normal explanation of the organism's inherited traits refers. Rather than structurally mirroring the changes in an environment, the possible transformations of a set of symbols, for example, will provide only a simplified interpretation of these environmental changes. Human scientists have acquired the ability to identify not just a host of things relevant to many niches on this planet, but things in places that aren't suitable to any form of life. This is the exception that proves the rule. So while an organism may have the ability to identify everything in *part* of its environment, that which is relevant to the Darwinian, etiological perspective is the *whole*

environment, which is likely to contain things the organism cannot identify. And since the relevant part of the environment, that makes for an isomorphism with an organism's set of intentional icons, would have to be defined precisely in terms of what the organism can actually identify, the claim that intentionality is isomorphism wouldn't reduce talk of intentionality to talk of something else. The isomorphism would require that the codomain be limited by some symbols that already have semantic contents, such as those things in the environment for which the symbols stand in as substitutes.

Millikan turns to the potential for scientific knowledge of the real world, in explaining the nature of intentionality: a symbol's having of content, in the most general case, is the symbol's capacity to link with other symbols to bear an isomorphism with an environment. That is, meaning is generally truth in the sense of correspondence. But Millikan turns to etiology and, fundamentally, to biology in explaining how content is determined: a symbol has its specific content, because the symbol is subject to a purposive function that determines what the symbol is supposed to do. As shown above, this etiological account of purposive functions implies that if intentionality is isomorphism, there is likely no intentionality, or that if there is intentionality, because there is the isomorphism in question, Millikan's theory of content isn't reductive. But Millikan's etiological theory of purposive functions assumes that content can be reductively explained, which is why she means to posit only so-called objective functions that don't depend on prescriptive norms, since these norms might themselves depend already on the use of symbols. Thus, once again, an appeal to basic naturalistic ontology provides for an account of what it is for a symbol to have content, only at the expense of a conflicting account of what it is that determines the content. The determinants would

seem to be prescriptive norms, but there are no such norms in basic naturalistic ontology. To get around this conflict, Millikan tries to naturalize the norms, but the norms to which she appeals are prescriptive, after all, and the mindless tinkerer of natural selection that would have to generate some of them produces creatures that are likely unable to identify all parts of their etiologically relevant environments. This means that on Millikan's account of the purposive functions that determine content, content isn't isomorphism because there's no such isomorphism between organisms and their environments, despite her account of the nature of intentionality as isomorphism. And Millikan can save the latter account only at the cost of making this account nonreductive, and thus on her view, non-naturalistic.

So no matter what element of basic naturalistic ontology a theorist uses when answering the question, "What is intentionality?" the theorist seems able to answer to the question, "What determines a symbol's content?" only in a way that conflicts with the answer to the former question. Assuming the limits of semantic relations are only prescribed, and assuming prescriptive norms aren't found in basic naturalistic ontology, any adequate theory of how content is determined should conflict with any theory of the nature of intentionality that appeals to that ontology. This is why, one way or the other, regardless of their particular answers to the above two questions, each of the three theorists, Fodor, Dretske, and Millikan, offers a theory of content with the internal conflict. Each assumes the way to naturalize content is to appeal to basic naturalistic ontology, and to add an account of content determinacy that is consistent with reference to some element of that ontology. If content is determined by prescriptive norms, there will be no such consistency, because there is apparently no place for these norms in what

a naturalistic philosopher assumes there is at the deepest metaphysical level, to make sense of what scientists say there is at the empirical level. Neither is there such consistency in the case of so-called descriptive norms that seem to have a place in basic naturalistic ontology, but that aren't just descriptive after all, because they prescribe a relation's limits.

5.3 The Necessity of the Internal Conflicts

Before I turn to my alternative approach, I want to address a possible response to my claim that there are, and must be, these internal conflicts. Addressing this response also presents an opportunity to explore further the source of the internal conflicts. The response is that the questions of the nature of intentionality and of content determinacy are independent of each other, and so answers to them can't be in conflict. For example, what might be said is that Dretske's point about causal power applies to things that are naturally real, from a metaphysical viewpoint, but his point about purposive functions applies to how certain real things are arranged, or to what sets the limits of semantic relations. Assuming the point about functions is historical or otherwise empirical rather than metaphysical, there should be no need for the account of functions to draw on basic naturalistic ontology. Indeed, there would seem to be no possible conflict between the two accounts, assuming the functions *emerge* from a lower level of nature such that the vocabularies needed to give these accounts are autonomous.

The problem with this is that prescriptive norms couldn't emerge in the same way that objective properties are thought to emerge in special sciences other than psychology,

and when naturalism is conceived of mainly as an ontological thesis, there's no reason to credit prescriptive norms with a special type of emergence. If prescriptive norms are real, there are fundamentally two kinds of things in the world, objective entities and subjective ones, since the latter wouldn't be determined by the former. No prescribed relation *need* be instantiated under any objective conditions, which is to say that the instantiation has no *objective sufficient conditions*. Moreover, any prescribed relation can be instantiated despite the difference between the actual and the ideal situation, or despite the lack of actual interaction between the *relata*; that is, the instantiation has no *objective necessary conditions*.[14] But any objective nomic relation is biconditionally related to the meeting of certain objective conditions: if certain conditions are met, the relation must be instantiated, and the relation is instantiated only under these conditions.[15] This is why the question of prescribing an objective nomic relation is empty. For example, assuming the orbits of planets in a solar system are set objectively and nomically, saying that the planets *ought* to move as they do has no added explanatory value. In so far as the orbits are objective phenomena, they are nothing more than the meeting of certain objective necessary and sufficient conditions. Naturalistic ontology is monistic rather than dualistic in the Cartesian sense, since even though there may be levels of irreducible natural

---

[14] For example, there are no objective sufficient or necessary conditions of someone's having a moral obligation to perform a certain action. This is a consequence of the logical gap between descriptive and prescriptive statements. Were there such conditions, having an obligation would be like a physical relation that is nothing more than the interaction of certain things that happens when certain conditions are met. On the surface, at least, to say that people *should* perform a certain action is to assume not just that this would be preferable, but that the action isn't simply necessitated, that there is a level of explanation according to which a person is free to perform or not to perform the action regardless of the circumstances. Being subject to a prescriptive norm isn't the same as meeting the standard. Someone may meet a moral standard by saving a drowning child, and there may be necessary and sufficient objective conditions of the performance of this action. But there seem no such conditions for having the obligation in the first place.

[15] This point holds even for a multiply realized property, since all mechanisms that can realize the property must have in common the general capacity to realize the property. What makes the list of this set of mechanisms disjunctive is only the lack of an *interesting* feature shared by them, which might motivate the formulation of a theory according to which the realized property has only one type of realizing mechanism, after all. In this way, the multiplicity of realizing mechanisms is subjective. See section 2.9.

phenomena, the higher levels are assumed to be determined, or limited by, the lower ones. What ought to be done, though, as set out by a prescription, isn't determined by what is the case as set out by an objective statement.

The two questions of content may be separate, but answers to them each must refer to things that belong in basic naturalistic ontology, once naturalism is construed in these metaphysical terms and the metaphysical question of the nature of intentionality is given priority. So given the direct approach to the metaphysical question of intentionality, taken by the three theorists, I maintain that their theories are bound to have internal conflicts. This approach forces an appeal to a hidden prescriptive norm, that is, to one that only seems descriptive, since only such a norm could seem, on the surface, to belong in a materialistic ontology. With respect to each of the three theories, a determinant of content is chosen for its ability to be explained as clearly a part of the rest of nature, but this determinant is pressed into doing work that no such part could do. A thoroughly natural determinant is expected to be able merely to prescribe a limit to a natural relation. That is, the descriptive norm, which is supposed to have prescriptive force, is said to be capable of something that neither a neural particular that engages in mechanical interactions, nor an internal indicator used to cause an organism's movement, nor a naturally developed capacity for map-making can support. This is because these realizers of semantic relations are explainable in terms of objective nomic relations, which is why the so-called norms that determine the relations are said to be descriptive, not prescriptive.

If intentionality is (1) a set of nomic relations realized by a neural particular and triggered initially by a certain perception, (2) a way of using information, learned by

reinforcement, or (3) a naturally developed relation needed for realistic knowledge, a semantic relation can't be prescriptively determined. Asymmetric dependencies between nomic relations, indicators with causal powers, and historically determined isomorphisms are just the sorts of things that are biconditionally related to the meeting of certain objective conditions. But metaphysically robust relations, functional conditions with structuring causes, and functional traits developed by a mindless tinkerer are all the sorts of things that can't be so biconditionally related, for them to determine content by a mere prescriptive norm. Thus, in spite of what the theorists may say, a symbol with metaphysically robust content needn't be caused by something in its extension under any conditions (there is no Normal situation for such a symbol); an internal condition, given a purposive function by a structuring cause, lacks causal power in its own right, since this condition is made into an artifact with subjective properties; a naturally developed capacity for mapping isn't likely to generate symbols that are isomorphic to changes in the organism's ecological niche. The hidden prescriptive element in each case, robustness or a purposive function, makes the semantic relation capable of being only partially realized, given what would count as full realization on each of the three views of the nature of intentionality. Specifically, a mechanism, or system of causal relations, is realized only under certain conditions, like a Rube Goldberg contraption in which every part is in place; something is naturally real only if the thing has its own objective causal power; and isomorphic structures must fully mirror each other. But there could be no such necessities were the limits of the mechanism, internal condition's use, or map determined by a prescriptive norm, or by a descriptive norm capable of determining what

merely *should* be so, as opposed to what would have to be so under specifiable conditions.

I summarize this account of the internal conflicts in the table below.

	Nature of Intentionality	Explicit Determinant	Implicit Determinant	Internal Conflict
Fodor	A set of nomic relations between a mechanically triggered, neural particular and some external causes	Asymmetric dependency of some of the nomic relations on another of them	Unspecified prescriptive norm, implicit in notion of metaphysical robustness	Neural mechanism subject to a Normal situation, precluding robustness of semantic relation
Dretske	Internally detected information an organism learns to use by reinforcement	Purposive function, "recruitment" as nothing more than reinforcement	The function as a prescriptive norm, generated by *C*'s initial receptivity to *R*	Info has intrinsic power as structuring cause, but functional indicator has only subjective causal power as means of satisfying *C*'s receptivity
Millikan	Isomorphism that figures in explaining a certain reproductive process	Purposive function, typical effect explaining the existence of what is reproduced	The function as a prescriptive norm, generated by value of explaining the owning of traits by a mindless tinkerer	Intentional icons must be isomorphic to environment for realistic knowledge, but won't enter into surjective relation, as products of blind tinkering

Table 1. The internal conflicts in each of the three theories of content

5.4 Naturalistic Methodology and the Substitutionary Aspects of Symbols

So the first approach, taken by Fodor, Dretske, and Millikan is to explain the nature of intentionality directly, in ontological terms, unifying semantic relations with a broader kind of relation, and then adding some account of content determinacy. This added account is problematic, because semantic relations seem governed by prescriptive norms, and yet there is no obvious way of adding an appeal to such norms to an account of intentionality that makes use of basic naturalistic ontology. But without something setting the limits of a semantic relation, there is no such relation as distinct from some other kind of relation. An alternative approach is needed. One other approach, besides eliminativism with regard to mental properties, would be to give up on explaining content in naturalistic terms, simply affirming that semantic relations are determined by prescriptive norms which have no natural basis.

My own view, however, is that a kind of naturalist can afford to posit the prescriptive norms, namely one who appeals to naturalistic methodology rather than basic ontology, in explaining the nature of intentionality. A theorist can address the metaphysical question of content indirectly, by drawing inspiration not from basic naturalistic ontology, but from scientific methods. Instead of focusing on what scientists say there is, the focus can be on what scientists do. What makes this account naturalistic is that the account can still be informative, taking care to explain the content of symbols without appealing to the unexplained content of other symbols. Moreover, the account is consistent with the naturalistic worldview and it shows how symbols fit into the scientifically explained world. However, the account isn't entirely reductive, in that I

don't try to explain the prescriptive norms in nonprescriptive terms.[16] By emphasizing methodology in this way, I think there are at least three advantages to be had. These are (1) a unified answer to the questions of the nature of intentionality and of content determinacy, that is, at least an internally consistent answer, (2) a naturalistic explanation of the substitutionary aspect of intentionality, and (3) a naturalistic explanation of the prescriptive norms that determine semantic relations. I return to (1) in section 5.5, and to the other two advantages in the next chapter.

If the first approach to explaining content begins with unifying intentionality with something posited by metaphysical assumptions of scientific explanations, and the second approach begins with unifying intentionality with the form of scientific explanation, I need to say something about this form. What is the relevant part of scientific practice? Scientists are engaged in a form of explanation, and most recently there have been two accounts of scientific explanation, the syntactic and the semantic accounts. The syntactic account, put forward by the logical positivists some decades ago, is that the explanation requires a theory, where a theory interprets theoretical terms, by linking these terms, by means of correspondence rules, to descriptions of observables. In other words, a theory is a kind of system for deductive arguments in which there are rules for deducing general statements about observables from general statements about unobservables, so that the latter become shorthand for the former. In this way, the scientific practice of appealing to unobservables can be reconciled with empiricist principles. Regardless of whether this account of scientific explanation is satisfactory,

[16] To this extent, the alternative approach to explaining content doesn't follow scientific methods, assuming scientific explanation is reductive. A naturalist taking up the alternative approach explains content in some terms that are familiar from an account of scientific practice, but is led to assume that there are irreducible normative determinants of semantic relations, given the failure of the reductive theories I've criticized.

this account is unsuitable to my purpose, which is to find the resources to explain content in naturalistic terms. The syntactic account doesn't address the question of how a theory relates to what the theory is about. Carnap (1950), for example, said that the meaning of terms in a theory is given by rules of interpretation, and these rules, which make for the language in which the theory is expressed, are chosen for various social and otherwise pragmatic reasons. The problem of the theory-to-world relation is either dissolved or pushed back, since the choice of a language is done already with symbols that bear a relation to the world. This may not be a difficulty for the syntactic account of theories, but this does pose a difficulty for someone who wants to explain content using terms derived from an account of naturalistic methods.

The more recent, semantic view of theories, however, poses no such difficulty, since the theory-to-world relation falls within this view's scope. On this view, scientific explanation is done with *models*. There are numerous kinds of scientific models. For example, there are physical representations of target phenomena, such as maps, diagrams, and scale models. There are also mathematical statements referring to abstract models, such as an equation referring, by definition, to an idealized relationship. The idealized, abstract model is then applied to the real world, with an applied mathematical model of a specific situation (Giere, 1999).[17] Giere argues against the view that the model-to-modeled relation is one of isomorphism. In the case of an equation, for example, a margin of error can be added so that the equation can be taken to refer not to an abstract object, but directly to a relation between concrete objects. According to Giere, however, a margin of error is added only relatively late in an instance of scientific practice, when an

[17] In highlighting the different kinds of scientific models, Giere wishes to argue against Suppes (1960). According to Giere, empirical models are not all mathematical ones, given a logical view of mathematical models.

explanation is tested against actual measurements. Assuming the explanation is meaningful prior to the test, the meaning must lie in a relation to an idealized situation rather than to an observed one. Giere (2004) says that at best models are *similar* to what they represent, not isomorphic, although he grants that similarity isn't sufficient for the model-to-modeled relation. A cloud formation may be similar in appearance to a train, but neither is a model of the other. Indeed, even in a genuine case of a scientific model, the notion of similarity doesn't capture the model-to-modeled relation, since similarity is symmetric, whereas the model-to-modeled relation is asymmetric. For example, a sculpture of the double-helix structure of DNA models this structure, not the other way around.

The model-to-modeled relation isn't one just of similarity, says Giere (2004), but scientists exploit the similarity and use a model as a representation. This raises the question of what sort of use is sufficient for a representation. At any rate, the uses of scientific models should range from the enterprise of learning about the environment in the attempt to control it with the model's technological applications, to the minimal use when the model is formulated out of pure curiosity at how the modeled system works. Also, Giere (1999) maintains that it's not vacuous to say a model is similar to what is modeled, since the respect and degree of similarity can be specified. If a model is similar only in some rather than all respects, and only to some degree, the model must be a *simplification* of some sort. These claims about a model's limited similarity to the modeled seem to apply to all scientific models. An idealized, mathematical model, for example, is just a highly simplified, or narrowly focused, one of a concrete situation. I happen to agree with Giere that modeling is central to scientific explanation. Indeed, the

use of so-called *ceteris paribus* laws seems another case of modeling. The point of a special science isn't to discover laws that are often violated and thus, strictly speaking, that can't be laws at all; instead, the point is to model complex patterns, simplifying them in a way that explains their cause or that predicts their consequence. But regardless of whether the semantic account of explanation is complete, this account does provide a starting point for explaining content in general. The claims that a model becomes a representation only when used in some way, such as in an attempt to control the modeled, and that a model simplifies and is thus asymmetrically related to the modeled, point to the substitutionary aspects of the symbol-to-symbolized relation. These ways a symbol stands in for something else are distinguishing features of intentionality that are left out of the three naturalistic theories I considered in earlier chapters.

So I want to elaborate on the substitutionary aspects of scientific models. One such aspect is apparent in the way a model simplifies what is modeled. This goes beyond saying that the two are similar to each other, since there's a normative aspect to the simplification. Some of the simplifications, at least, surely limit the model's uses, which makes the model a *schematic, much diminished version of the modeled.* For example, one model of a system, given by a set of CP laws, will be useful only when the special conditions are met, and so another model of the same system might have to be used on another occasion, when other conditions are met. Given that the same modeled system can be found to operate under different conditions, each model is only a partial account of the system, and there may be difficulty unifying the models. Ideally, one model might predict how some set of elements operate under any conditions. But such a model would be closer to a simulation than a representation of the modeled system, precisely because

the model wouldn't stand in for the system in a limited capacity. Instead, scientists often resort to a piecemeal approach, since partial understanding is better than none at all, and the scientist's interests guide the modeling process, focusing attention on how a system works under one rather than another special set of circumstances. For a model to predict how certain elements work under any set of conditions, the model would have to be as complicated as the possible interactions of the modeled elements. Instead, models are favoured for their simplicity: the simpler model with the greater explanatory power is considered the better one. But this is a matter of making due with the limited resources available to a human scientist; the simpler and thus limited model is better because there is no practical way of encompassing all possible interactions in a useful explanation.

Besides being not just similar to the modeled, but necessarily a limited, schematic and thus much diminished version of the latter, a model is often used to *control* what is modeled. An operation performed on the model is done as a means of operating on what is modeled, by way of preparing to deal with the modeled system. To take a simple, nonscientific example, parts of a highway map can be highlighted in preparation for driving on the highway. The instrumentalistic aspect of scientific models is also a substitutionary one, in that the model is needed to manage the modeled system in some way, and the model helps by standing in for the system as the modeler develops the means to accomplishing this task. To take the extreme example, scientists are often interested in modeling parts of the environment that, left uncontrolled, prove disastrous to humans, such as diseases or certain weather systems. In its instrumental role, a model is like a canary in a coal mine, a sort of test case that can be monitored or manipulated as a means of solving a problem with something else. Although the use of the model is

limited, compared to the possible uses of the modeled system, given the many conditions under which the system can be found, the model is used as a test subject in preparing to solve a problem with what is modeled. These two substitutionary aspects, of being a diminished version of, and a type of instrument in solving a problem with, the substituted, are what I think make the scientific model such a useful starting point in explaining intentionality, because symbols in general, too, substitute in both ways.

Taking the methodological, as opposed to the ontological approach, there are at least two ways of explaining this substitutionary aspect. The first is to say that symbols such as mental representations *are* scientific models of some sort. This claim might be supported by the so-called theory theory of cognition, proposed by some developmental psychologists, according to which cognitive processes in general, including those of children, are forms of scientific reasoning.[18] Glennan (2005), for example, applies the semantic account of scientific theories to the theory theory, arguing that what is common to the cognitive processes of children and scientists is that they are forms of modeling. Glennan agrees with Giere (1995), Cartwright (1999), and others, that a model stands as an intermediary between a theory and the modeled phenomenon, where the theory is a set, not of theoretical laws in the sense of unqualified assertions, but of principles and rules for devising models.[19] The first way, though, is problematic. Such an account would likely presuppose some symbols instead of explaining content in general. Saying that mental representations, for example, are scientific models requires an account of what makes them scientific, but science is, among other things, a social institution that already depends on the cooperation of people with their mental representations. Any prescriptive

[18] See Gopnik (1996) and Gopnik and Meltzoff (1997).
[19] As Cartwright puts it, the relation between a model and a theory is similar to that between a fable and a moral.

norm governing the use of scientific models would likely be socially determined, and thus dependent already on symbols, such as those used in scientists' beliefs about how to practice science. Moreover, in so far as scientific models are intermediaries between rules and modeled phenomena, and the rules are explicitly represented, once again the first way would have to presuppose some symbols, namely those that represent the rules used to construct the model. Perhaps these latter rules need not be internally represented, so that when a child, for example, forms a mental model of a dog, the child doesn't first have to think of a rule for doing so. In any case, I'm not going to defend the first way here.[20]

Instead, I want to explore a second way, which is to say that one *similarity* between symbols in general and scientific models is the substitutionary aspect of their relation to the symbolized or to the modeled thing. There need then be nothing specifically scientific about a child's mental representations, and yet the intentionality of those representations may best be characterized in terms that are already understood to characterize the intentionality of scientific models. My second, methodological approach to the problem of explaining content, then, is to turn to scientific practice, not directly to basic naturalistic ontology, in answering the question about the nature of intentionality. What I find is a striking case of substitution, of one thing standing in for another, in the

[20] Another way of supporting the stronger position would be to appeal to mental model theory, proposed by the psychologist, Johnson-Laird. According to this theory, human reasoning can be computational even without propositional representations or formal logic, if this reasoning uses mental models instead. Johnson-Laird (1980) distinguishes between a propositional representation and a model: while the former is descriptive, and thus either true or false, the latter "*represents* a state of affairs and accordingly its structure is not arbitrary like that of a propositional representation, but plays a direct representational or analogical role. Its structure mirrors the relevant aspects of the corresponding state of affairs in the world" (98). Whichever theory of reasoning has more advantages, the mental model theory or the orthodox theory of formal logic, I don't think Johnson-Laird's way of characterizing mental models, as being similar to states of affairs in the world, accounts for how they meaningfully relate to these states of affairs. A symbol's standing in for something as a substitute is more central to the symbol's intentionality than the symbol's similarity to something.

case of scientific models. The task then is to give a naturalistic and unified, as opposed to internally conflicted, explanation of the prescriptive norms that determine content and of the substitutionary aspect of symbols.

5.5 Some Objections

I'd like to respond now to two objections to what I've argued in this chapter. It might seem that there can't be two different strategies for explaining content, since the question of the ontological status of semantic relations is logically prior to any methodological question of how symbols are used. Thus, the methodological strategy either presupposes the metaphysical one or just changes the topic, avoiding the deeper question. There seem to me two, mutually exclusive responses to this objection. A naturalist may have to assume that there are certain superior methods for acquiring knowledge of nature, and this assumption may be *independent* of any assumption about what there is at an ontological level. This is because a naturalist's answer to metaphysical questions should appeal, directly or otherwise, to naturalistic, including scientific, methods. Only were a naturalistic account of these methods possible without *presupposing* them would the methods lack a primitive status in the naturalistic worldview.

In any case, suppose an account of naturalistic methods does indeed depend on one of basic naturalistic ontology, and thus that someone taking up the methodological strategy in explaining content is obliged to answer straightforwardly the metaphysical question, "What is intentionality?" I think there are metaphysical implications of the

substitutionary account of content, one of which is that content is determined by prescriptive norms, which raises the question of the ontological status of these norms. Another implication is that symbols must be used in a substitutionary way, which raises the question of what counts as a user. Other implications will become apparent in the next chapter, when I discuss the connection between substitution and prescriptive norms in more detail. My point here, though, is that any metaphysical question about semantic relations can be addressed *indirectly*, guided by an understanding of naturalistic methods. So no metaphysical question need be avoided, once a naturalist chooses to explain content using resources found in an account of naturalistic practices. The difference between the metaphysical accounts of content is that the one not guided by naturalistic methodology seems to have trouble accounting for the norms that determine the content of symbols, since a theorist offering this sort of account tends to rush headlong into naturalistic ontology, finding relations that are too objective to be normatively determined.

A second objection, following up on the first, is that, regardless of whether an explanation of semantic relations begins with a consideration of naturalistic methods, as soon as the metaphysical question is broached, the explanation is bound to be internally conflicted since the normative determinants are themselves highly problematic. In short, the methodological strategy is no more unified than the metaphysical one. Now, I believe I've shown why theorists who construe the problem of semantic relations in mainly metaphysical terms tend to put forward internally inconsistent theories. The cost of turning to basic naturalistic ontology *at the outset* is the inability to account naturalistically for the normative determinants of semantic relations. If a semantic

relation is just a certain relation found throughout the natural world, and this natural relation is objectively determined, the limits of a semantic relation can't be prescribed. Since these limits nevertheless seem to be merely prescribed, the theorist taking the ontological approach posits hidden normative determinants of content. Any explanation of the normativity of these norms will be inconsistent with the metaphysical explanation of the nature of intentionality, assuming prescriptive norms aren't likewise found throughout nature.

The methodological approach leads to no such necessary internal inconsistency. In appealing to scientific practice, not to metaphysical assumptions of scientific theories, I find that the substitutionary aspects of the semantic relation might be revealed by an account of the relation between the explanatory model and what is modeled. The substitutionary relation depends on the instrumentalistic use of a symbol, and this use, unlike, say, the mechanical implementation of a nomic relation, is the sort of thing that could be prescriptively normative. It's hard to see how a relation from basic naturalistic ontology might have individuating limits even though there's no set of objective conditions under which the relation must be instantiated. But, as I argue in section 6.4, if anything can enter into a prescribed relation, without being necessitated by such a set of conditions, it's the semantically relevant use of substitutes, since this use is, arguably, the primary act of self-determination.

I think the normative determinants can be accounted for within a comparatively nonreductive naturalistic framework. It's easier to take prescriptive norms as given in the naturalistic worldview when naturalism is seen as depending on a method of using explanatory models as substitutes. That is, from a methodological view of naturalism,

prescriptive norms have at least a special status, since the method in question is a substitutionary activity governed, at least, by social and rational norms. The problem with the metaphysical approach is that it defines what is natural in certain philosophical terms that make sense of the content of scientific theories, and scientists don't explain prescriptive norms. Indeed, the scientific practice of objectively testing hypotheses seems to preclude a scientific explanation of norms or of certain other constituents of subjectivity, such as qualia, the private facts of consciousness.[21] I think there can be a broadly, that is, a philosophically, naturalistic explanation of these constituents. But any explanation that begins with what, within the natural world, is made familiar by scientific methods will have to deny that there are certain aspects of subjectivity or else try to explain the subjective in terms of the objective. Neither option is promising, and the failures of the three theories I've criticized show, in particular, that the reduction isn't likely to succeed.

By contrast, when a broadly naturalistic explanation begins with a view of a method used in scientific explanations of nature, there's no longer the need for a reduction of the semantic to the nonsemantic, or of the prescriptive to the descriptive. Prescriptive norms, and perhaps qualia as well, can be credited as real and can even be accounted for in naturalistic terms obtained from a consideration of the naturalistic method. A naturalist takes for granted not just the content of scientific theories, but the methods scientists use in formulating and in justifying these theories. If an account of the nature of intentionality follows from an account of a certain naturalistic method, illuminating the substitutionary aspects of symbols, I maintain that the former account is

[21] What is often called the hard problem of consciousness is to account for how that which is most subjective, namely what it is like to have a conscious experience, could arise from the public, objective, scientifically explainable world. See Chalmers (1995) and Nagel (1974).

naturalistic and noncircular even though intentionality isn't thereby reductively explained. And if an account of the norms that determine semantic relations follows from the above substitutionary account of intentionality, the account of the norms is naturalistic even though, again, the norms aren't thereby reductively explained. There is no need for those reductions, given that naturalism takes for granted the substitutionary use of explanatory models, not just metaphysical notions of causal relations and the like.

Another way of putting the point is to say that there are two kinds of reductive explanation. The theorists who take up the metaphysical, rather than the methodological, strategy identify intentionality with a relation found throughout nature, which thus has a secure place in naturalistic ontology. Talk of intentionality is thereby reduced to a completely different kind of talk, moving from a special domain to a much more general one, from semantic relations, say, to isomorphism or to information. Another kind of reductive explanation has the more modest task of providing an illuminating account, by avoiding vicious circularity. Thus, intentionality might be explained as a kind of substitution, and this explanation could be considered reductive as long as the relevant kind of substitution weren't itself dependent on symbols that have intentionality.

So the main advantage of my approach to explaining content, over that of the three theorists, is that I don't begin with an undiscerning use of resources from naturalistic ontology. That is, I don't identify intentionality with something that can't be normatively determined, thus ensuring that the theory is internally contradictory, given that the determinants turn out to be prescriptive norms. On the contrary, I begin with the special domain of scientific practice, which is equally-well taken for granted by naturalism, and I identify intentionality with a substitutionary relation found in a part of

nature that is more likely to be normatively determined. This isn't to say I put forward the tautology that intentionality is substitution, where "substitution" is just another name for intentionality. As I point out in section 6.1, not all substitutes have intentional properties, so one of my tasks in the next chapter will be to explain which substitutes are semantically relevant and to do this in a noncircular way. Still, the relation of one thing standing in for another is an intermediate sort of relation, not as widespread or as naturally fundamental as causation or information, but more widespread than semantic relations. So although all symbols may have substitutionary aspects, it may still be informative to say what these aspects are and how they differ from semantically irrelevant kinds of substitution.

All noncircular, genuine explanations are reductive in this limited sense of explaining one thing in terms of something else. The other three theorists, though, try to ground intentionality in too fundamental a level of nature, answering too directly the metaphysical question of the nature of intentionality. They reduce talk of intentionality not just to talk of something else, but to talk of something that can't be normatively determined, even though semantic relations seem to be so determined. I identify the semantic relation with something that can be normatively determined, namely a kind of substitution, and so I avoid the necessity of putting forward an internally conflicted theory of content. And I do this without giving up on naturalism, by finding a different naturalistic starting point, in a type of methodology. While I don't claim to reductively explain prescriptive norms, I try to clear a path for a naturalistic explanation of semantic relations, given their normative determinants.

5.6 Conclusion

In this chapter, I've argued that the prescriptive norms that seem to determine semantic relations can't be naturalistically explained, given the metaphysical approach to this sort of explanation, and that the three theories I considered in earlier chapters take that approach, which accounts for their failure. But a methodological approach seems to be free of this particular obstacle. In the next chapter I adopt the methodological viewpoint, and account for the norms that determine the semantic relations between mental substitutes and what the substitutes stand in for.

Chapter 6

The Prescriptive Norms of Mental Substitution

6.1 Introduction

In the last chapter, I diagnosed a common problem with the three theories of content I critiqued in earlier chapters. The problem is that their metaphysical account of the nature of intentionality contradicts their account of how the content of symbols is determined. The diagnosis is that the theorists construe the naturalist's job in this case as one of answering directly the metaphysical question about content, by identifying the semantic relation with a relation that is part of a naturalist's ontology, and thus that is unproblematic from a naturalist's viewpoint. It becomes difficult then to address just as directly the question about content determinacy, since while content seems normatively determined, norms have only a problematic metaphysical status. Thus, this second question about content is addressed indirectly, by positing so-called descriptive norms or else something non-normative, such as Fodor's asymmetric dependency which is nevertheless supposed to account for the arbitrary robustness of symbols, for a feature that amounts to their normativity. I then argued that there may be a way to avoid this problem, while giving a naturalistic philosophical explanation of semantic relations, and even one that affirms the normativity of their determinants, namely by taking as given

naturalistic methods rather than ontology. This leads to an emphasis on the substitutionary aspects of symbols.

In this final chapter, I want to follow through with a sketch of how semantic relations are normatively determined, given their substitutionary aspects. The substitutionary aspects are taken for granted by naturalistic methodology, as shown in the last chapter, and now I want to show how an account of the prescriptive determinants of semantic relations is derivable just from a consideration of those substitutionary aspects. Thus, I mean to show how an account of the determination of content by prescriptive norms can be just as naturalistic as an account of the substitutionary nature of intentionality that is based on naturalistic methodology, by showing how the former account follows from the latter one. I argue first, in sections 6.2 and 6.3, negatively, that the determinants of semantic relations must be prescriptive norms, because they can't be descriptively, or objectively determined. The semantically relevant kind of substitute is digital, meaning the substitute's intrinsic properties are irrelevant to its ability to serve as a stand-in for something else, and anything with such arbitrary properties doesn't, as such, enter into an objectively determined relation. I then argue, positively, that the determinants are prescriptive norms, because the substitutionary relation is partly a matter of use, given the instrumentalistic aspect, and the user is just the type of creature whose actions are subject to norms (6.4). Finally, I consider an objection to what I'm calling the instrumentalistic aspect of substitutionary, and thus of semantic, relations (6.5).

In what follows, I'll focus on mental representations, not on words or sentences. I'll assume that these representations are identical to networks of neural activity or at

least to some combination of neural elements.[1] Also, I'll take the inclusive view that a mental representation of a dog isn't just a list of statements encoded in mentalese, nor just a mental image; instead, the representation is the whole web of neural elements that is semantically related to dogs, that is, everything a brain does when the brain enters into a substitutionary relation to dogs. The mental representation can include memories, perceptions, imaginings, emotions, and so on, and so the representation can change and may be partly learned and partly innate or acquired. For example, someone with an instinctive fear of spiders has different neural activity when thinking of spiders than does someone without this fear. Any neural activity associated with the emotion of fear, including memories and hormone secretions, may contribute to the realizer of the mental symbol.[2]

Now, to say that a symbol stands in for the symbolized thing isn't itself controversial.[3] However, it's just as clear that not all substitutes have semantic content. For example, a substitute teacher isn't *about*, say, the ill primary teacher stood in for by the substitute, and neither does the wiping of a windshield with a cloth that substitutes for

---

[1] And I'll speak vaguely of "combinations of neural elements" to capture the idea that the neural instantiation of a representation is highly complex, involving not just a number of distinct cells, but hierarchies of processing, synchronous or overlapping firings of neurons, and so on.

[2] My reference here to the neurological basis of mental representations skips over a number of enormously complicated issues in cognitive science. Different theorists point to different parts of the brain as the basis of different kinds of representations. I mean for my substitutionary account of intentionality to encompass many of these neurological accounts. Whatever the brain does internally, as I say, to prepare to deal with something in the outer environment, counts as substitutionary, and specifically instrumentalistic, activity. And so a theory of intentionality that focuses on the substitutionary aspects of mental symbols seems most compatible with an inclusive view of the neurological basis of these symbols. Another inclusive view is found in Recanati (2007), which emphasizes the way the circumstances under which a thought is had determine the thought's content.

[3] Haugeland (1991) gives the ability to stand in for something else as a criterion of intentionality: "plants that track the sun with their leaves needn't represent it or its position, because the tracking can be guided directly by the sun itself. But if the relevant features are not always present (manifest), then they can, at least in some cases, be represented; that is, something else can stand in for them, with the power to guide behavior in their stead. That which stands in for something else in this way is a representation; that which it stands in for is its content; and its standing in for that content is representing it" (172).

a broken windshield wiper *refer* to the broken wiper. Nevertheless, a substitute teacher or a substitute wiper substitutes in both ways discussed in section 5.4: (1) each is a limited version of an original, since circumstances are such that there's a reason why these substitutes are kept only as back-ups, and (2) each is used to solve a problem with the original.[4] Moreover, there are prescriptive norms governing the use of these substitutes. A substitute teacher ought to do his or her job while in class, and a person ought to manually wipe the dirty parts of the windshield. But just as the norms governing the use of scientific models derive from the mental representations in the beliefs that establish the institutions of science, the norms governing the behaviour of substitute teachers and windshield wipers derive from the beliefs and desires that establish educational systems and conventions of traveling by car. So assuming the two kinds of substitution discussed above are crucial to the semantic properties of mental representations, there must be something distinctive about the prescriptive norms that determine the content of these representations, and thus about the way the representations are used as substitutes.

Of course, I'm assuming the norms that determine the content of mental representations aren't derivative from the beliefs and interests that define some social institution or convention. A naturalistic explanation of the content of mental

[4] In what follows I'll speak of the second, instrumentalistic aspect of substitution in terms of a symbol's "solving a problem" or "dealing with" some other object, referring back to what I say about this aspect in section 5.4. I think these locutions are general enough to include the notion of control, or of regulating and directing one thing with something else. Still, there are problems with these ways of summarizing the point about the instrumental use of symbols. Some uses of substitutes may not be solutions to a problem or ways of dealing with something else, in any strict sense. Moreover, the paradigmatic kind of instrumental use of something may be the intentional choice of a means to achieve a *goal*. But any theory of intentionality that appeals to this symbol-guided kind of use would explain the content of some symbols in terms of the content of other symbols, which is not what a naturalistic theory of content is supposed to do. To the extent that "solving a problem," "dealing with something," or "using one thing to control something else" presupposes the use of explicit representations of a goal, an instrumentalistic theory of content that depends on any of these locutions fails as a naturalistic theory. I return to this problem in section 6.5. One advantage of a substitutionary theory of content, I think, is that it shows how to naturalize the instrumental use of symbols. What I call the use of a substitute to solve a problem or to deal with the substituted object is done by most animal species, and in many cases reflexively, without explicit representations of a goal.

representations shouldn't explain the content of some symbols in terms of the unexplained content of other symbols. But I think there are two other differences between the way mental representations, on the one hand, and substitute teachers, windshield wipers and so forth, on the other, serve as substitutes. Thus, there are two other reasons to favour the methodological approach to explaining content, taking intentionality to be a substitutionary relation determined by prescriptive norms, despite the apparent counterexamples. One of these reasons is negative, the other positive.

6.2 Digital and Analogue Substitutes

The negative reason is based on the distinction between *digital* and *analogue computers*. Often, this distinction is made by saying that the representations in a digital computer are carried by vehicles that can vary to some extent without affecting the content, in which case these representations are *discrete*, whereas the vehicles in an analogue computer can't vary without affecting the content, making these other representations *continuous*. Thus, the moving hands on an analogue watch represent time in a continuous fashion, since each position of the hands in their smooth sweeps across the watch's face correspond to a different time, including the minutest fraction of movement. A digital watch uses a sequence of countable digits to represent different times, and these digits don't merge continuously into each other, but have gaps between them which are treated as semantically irrelevant. There are problems, though, with this way of drawing the distinction. Sometimes it's unclear whether a computer has discrete

or continuous vehicles of representation. For example, some analogue watches have hands that move in discrete steps as opposed to continuous sweeps.

Searching for a more useful way of drawing the discussion, Pylyshyn (1984) says that the properties that make something an analogue system are just the system's own intrinsic, physical properties, which are therefore projectible properties, or properties referred to in a physical law. Some properties that make something a digital system, however, are multiply realizable and extrinsic to the system. These are the semantic properties possessed by the symbols, such as the digits that flash on the face of a digital watch, that refer to things outside the system according to a code, or to a set of syntactic rules making up the computer program. Thus, while the movement of hands in an analogue watch may be interpreted as referring to the time, no semantic interpretation is needed to explain the movement or any other internal working of the watch. The movement of the hands is due not to any irreducible semantic property of the hands, but solely to the physical relations between the parts of the watch. A digital system, however, requires a semantic explanation, because such a system uses some of its parts in a way that isn't explainable solely in terms of the internal workings of the system. These internal parts are the symbols that must be assumed to bear an external relation to things outside the system.[5]

As it stands, I don't think Pylyshyn's way of stating the difference quite works either. It's just not true that only the intrinsic properties of an analogue computer need be

[5] Pylyshyn's distinction between analogue and digital computers builds on Lewis (1971) and on the reply to Lewis, by Fodor and Block (1973). Lewis says that an analogue computer represents numbers, for example, "by physical magnitudes that are either primitive or almost primitive," given the language of physics, whereas a digital computer represents them "by differentiated multidigital magnitudes," which is to say, by emergent patterns of physical magnitudes. For Fodor and Block, the behaviour of an analogue computer instantiates a physical law, whereas the best way of generalizing about the behaviour of a digital one makes use of a special law that doesn't refer to physical types as such.

referred to in an explanation of the machine's behaviour. An analogue watch is an artifact whose parts serve a purposive function that depends on the intentions of designers and users, and the semantic relation between these intentions and the components of the watch account for why the components are positioned where they are and why they work as they do. So the point can't be just that the workings of an analogue computer are explainable solely in terms of the computer's intrinsic properties. Also, it's unlikely that an analogue computer's behaviour can be explained strictly in terms of physical, as opposed to special or CP laws. Again, the way an analogue watch's parts are organized has a teleological explanation, not just a physical one. Pylyshyn seems to appreciate this point, which is why he speaks vaguely of physical magnitudes as "magnitudes that can appear in reasonably general laws" (201).

In any case, I think the distinction should be drawn as follows. The properties of an analogue computer that are relevant to its *similarity* to some other system are possessed by the *vehicles* of the representations and are relatively *nonarbitrary* when compared to the properties of the similar system. Thus, there is an analogy between, on the one hand, the directionality of time and the subjective appearance of time's flow from one moment to the next, and on the other, the *motion* of the arms on an analogue watch's face. Indeed, there is the added analogy between the merging of one moment of time into the next and the *circular* motion of the arms. Suppose that instead of sweeping arms, an analogue watch were to have as a marker of each moment a steady light that moves not along the circumference of a circle, but along the edges of a square, on the watch's face. What would be strange about this watch is the dissimilarity between the ways in which time and the watch would work. There would be a change in the light's motion as it

moves along an edge of the square to one of its corners, but no corresponding change in the flow of time. So an analogue watch must have parts capable of motion, and more specifically of circular motion, to make for the watch's similarity to the continuous transitions between times; that is, the materials out of which the watch is made must be chosen for their capacity to stand in this relation of similarity to the other system.

There's no such restriction on the workings of a digital watch, and so there's much greater variety in what could constitute such a watch. What gives a digital representation its content isn't some intrinsic property of the vehicle that carries the content, but the ability of the vehicle to follow a code or program that isn't reducible to a physical law. Indeed, the algorithms that make up a computer program are written in an artificial language in which symbols can be introduced at the programmer's whim, as long as they're rigorously defined in the program. A computational symbol just has to follow the instructions in this code to have digital content. Of course, an analogue watch tends to have digital representations as well, such as the numbers or lines inscribed on its face. But an analogue watch represents time not just with these frozen inscriptions, but with the arms' motion. Indeed, some analogue watches have no helpful markers to judge the exact position of the arms at a given time, but just the moving arms. The main point, though, is that even though an analogue and a digital watch are each used as a device for telling the time, with their ways of representing different times, the capacity of analogue representations to be interpreted as having content depends on their similarity to what they represent. This similarity, in turn, depends on the intrinsic properties of the vehicles of the representations.

Now, to return to the difference between those substitutes that have semantic properties and those that don't, a substitute teacher or a substitute windshield wiper can be explained as an analogue system in my sense. These analogue systems are also substitutes, and so their relation to some other system isn't one just of similarity, but of standing in for the other system and of being used as a helpful alternative to it. Still, there are much greater restrictions on what could count as a substitute teacher than on what could count as a digital representation of a teacher. A substitute teacher has to perform the same tasks as the primary teacher, whereas a digital representation, such as the word "teacher," obviously doesn't have to teach and so doesn't have to possess the capacity to do so. The same is true with respect to a substitute windshield wiper or to anything that occupies something else's position, as compared to a digital representation of what is substituted.[6] Another way to put this point is to say that a natural law about the workings of an analogue substitute is derivable from a law about the workings of what is substituted, because this sort of stand-in and what is stood in for must be sufficiently similar to each other. So even though the hand-wiping of a windshield with a cloth is a comparatively limited version of an automatic windshield wiper—the latter works while the car is moving, doesn't get tired, and so on—and is used to solve a problem with what is substituted, in this case as a backup when the automatic wiper fails, the use of the analogue substitute can be explained without reference to these substitutionary aspects. What's crucial to the use of the hand-wiping of a windshield is just that this technique is sufficiently similar to the technique used by the automatic wiper. By applying a soft

[6] Terminologically speaking, I'll refer to the *substitute* and the *substituted*, and this distinction should be understood as formally similar to that between the *model* and the *modeled*. So instead of speaking of the original or primary thing whose position comes to be occupied by a substitute, as the *substituted-for*, I'll speak of it as the substituted or as what is substituted.

material to the windshield surface, and using a back-and-forth motion and perhaps water to prepare for the removal of dirt, the hand-wiping of the windshield is an attempt to perform the same work performed by the automatic wiper.

In other words, besides being a substitute, the hand-wiping is what I'll call a *near-replacement.* A near-replacement must be sufficiently similar to what is replaced, and thus the properties that are relevant to something's ability to serve as a near-replacement can't be arbitrarily different from those of what is replaced.[7] The same can't be said of a digital substitute. The words "automatic wiper" stand in for, without even nearly replacing an automatic wiper. Laws about the use of the word aren't derivable from laws about the use of the wiper, because the digital substitute and the substituted thing need have no significant similarities. The properties that are relevant to a word's ability to serve as a limited version of something else can be arbitrarily different from the properties of what the word stands in for. This is simply because a digital substitute isn't expected to do the same job as the substituted object. The relation between a digital substitute and the substituted object isn't explainable just in terms of the former's being an approximation of the latter. But, in the limit case, an analogue substitute shares *all* of the relevant properties with the substituted, as in the case of a spare tire, for example. In this case, the asymmetry of the substitutionary relation disappears. This is why some substitutes, namely what I'm calling analogue ones, or substitutes that are also near-replacements, lack semantic properties of their own. Reference to a near-replacement's

[7] Regarding the terminology I'm using here, a near-replacement is still a substitute in the above sense, of being a *limited* version of, and an instrument in solving a problem with, what is substituted. A substitute teacher and a cloth for wiping a windshield are near-replacements, in that they have only *some* of the relevant properties of what they stand in for. A spare tire is, rather, a replacement, assuming it has *all* of the relevant properties of what it replaces. For this reason, a replacement isn't a substitute in my sense, since the replacement isn't a comparatively limited version of what it replaces.

substitutionary aspects isn't needed to explain the use of the near-replacement. A near-replacement has to be just sufficiently similar to something else, and this is a symmetric relation, whereas a semantic relation is asymmetric: a symbol is about something else, not the other way around.

There might seem to be some counterexamples to my claim that analogue substitutes aren't about anything. After all, scale models and paintings are representations, and they're significantly similar to what they represent. But these analogue substitutes aren't sufficiently similar to their referents to be used as near-replacements for them. These analogue substitutes stand between digital ones and near-replacements, in that *some* of the properties that are relevant to their ability to stand as limited versions of other things are arbitrarily different from the properties of these substituted things. Such analogue substitutes are interpretable as representations, with semantic content, just to the extent that they are *partly* digital and that they aren't near-replacements for what they stand in for. Representational artworks, for example, are only superficially similar to what they are about, and this preservation of their substitutionary aspect is what allows them to have representational content. A painting of a door only looks like a door, because of tricks in the way paint can be applied to a two-dimensional surface, but the painting lacks a door's physical, three-dimensional properties. Again, a highway map has points that refer to points on the highway, because, despite its similarity to the highway, the map can't be used as a surface on which to drive.

However, a computer simulation of a highway might be sufficiently similar to the highway, as in the case of so-called virtual reality, that the simulation nearly replaces the highway instead of referring to it. Other simulations may be imperfect, as in a flight

simulator that simulates only some of what is involved in piloting an aircraft. Indeed, any simulation of only part of what there is to simulate makes for an imperfect replacement. For example, a flight simulation may have boundaries in the simulated sky through which the simulated plane can't pass. These limitations may be arbitrary if they are the result of the programmer's decision about what to simulate. But when the similarities in a representation vastly outnumber the dissimilarities to what is represented, the representational relation between the two systems is explainable in terms of the symmetric relation of similarity. Only when this is not the case, when the representation is far from being a near-replacement for what is represented, is there a need to explain the representational relation in the asymmetric terms of one thing's being a limited version of, and an instrument in dealing with, another thing.[8]

This substitutionary relation's asymmetry makes it a candidate for the semantic relation of one thing's being *about* something else and not the other way around. What I mean is that a near-replacement, such as a substitute teacher, and what is nearly replaced can each be used to perform the same sort of tasks, since their sets of intrinsic properties are sufficiently similar to each other. Granted, a token primary teacher, nearly replaced by a token substitute teacher, can't be used to teach while the primary one is being substituted, since the primary teacher may then be sick or away, which is what calls for the substitute in the first place. But, in general, there are conditions under which a type of near-replacement and the nearly replaced type perform the same type of work. The same

[8] To return to the example of the analogue watch, I think its representations are near-replacements for what they represent, in at least one respect. The arms on the watch's face move, and so they themselves pass through time, as does everything else in the physical world. But the arms' motion is also circular, which mimics the *feeling* of passing continuously through time. Thus, the use of the watch's representations is a near-replacement for the phenomenological aspect of what they represent, in that staring at the circular motion of the watch's hands brings to the forefront of consciousness the feeling of passing through time, submerging less focused subjective impressions of the passage of time. The representations are only near-replacements, though, since the watch can break or need winding or a change of battery.

can't be said of a type of digital substitute and its substituted type, since the digital substitute lacks the requisite intrinsic properties. At least, there is no need for a digital substitute to be similar to its substituted type, for the substitution to occur, whereas an analogue substitute, or a near-replacement, is useable as a stand-in only because of its similarity to the substituted type. Although there is some asymmetry between a *near*-replacement and what is replaced, making for a possible substitutionary relation between them, a substitutionary explanation of their relationship is derivable from the analogy between them. Whenever two things are sufficiently similar to each other, either is a limited version of, and can be used as an instrument in dealing with, the other, as in the cases of a flight simulation and of flying a real plane. But the analogy does the work in this explanation of the relation between them. By contrast, the substitutionary relation between a digital substitute and its substituted type is fundamentally asymmetric, since there is no underlying analogy between them.

6.3 The Normativity of Digital Mental Representations

The relevance of this distinction, between the digital and the analogue, to an account of mental content is that a neurally instantiated mental representation is a digital substitute in the foregoing sense. The neural structure standing in for an automatic windshield wiper, when someone thinks of this wiper, doesn't itself wipe the windshield. The representation is a substitute without being a near-replacement, and so there's no necessity of deriving even a partial explanation of the representation's usability from an explanation of the substituted object's usability. The mental representation might be

instantiated in something other than part of an organic brain, such as in a set of silicon chips in an artificially intelligent machine. In any case, there's much potential variety in the token vehicles of this type of representation, since different brains can think differently about the same object, bringing to bear different memories, moods, emotions, skills, disorders, and various other cognitive dispositions that provide different raw materials for the vehicles. Just as the words "windshield wiper" can be expressed in many different languages, making the properties of the vehicles arbitrarily different from the properties of their referent, the mental representation has many different neural instantiations—indeed, billions upon billions of them, given the number of unique connections the billions of mutable neurons in each person's brain can form.[9]

It's worth pointing out that this way of arriving at the kind of substitution that accounts for the content of mental symbols, by pointing to their digital aspect, is derivable from the above point about the first of the two substitutionary aspects of symbols. That there can be analogue and digital substitutes in my sense follows from the claim that a substitute is a comparatively limited version of something else. This is because the substitute can be *more or less limited.* A digital substitute is *maximally*

---

[9] There's debate about whether the activation patterns among nodes in a connectionist network should be interpreted as symbols and the processing regularities as rules, comparable to the symbols and rules in a classical, digital computer. Smolensky (1988) called the activation patterns "subsymbolic," and dynamic systems theorists, such as van Gelder and Port (1995) argue that the patterns are better interpreted in nonrepresentational terms. Others, such as Clark (1997), define "symbol" broadly enough to support a representationalist interpretation of the patterns.

Taking a connectionist network to model an organic brain, I think the representationalist interpretation is supported just to the extent that the neural patterns or states are digital and substitutionary. Still, it might be more accurate to say that such patterns stand with such things as paintings, sculptures, and scale models, somewhere between digital substitutes, which clearly have representational content, and near-replacements, such as substitute teachers, which clearly do not. If there's doubt about whether the patterns support a semantic interpretation, it's because the vehicles that generate the patterns are digital in some ways and analogue in others. Clearly, the activation pattern isn't a near-replacement for the external object, but some aspect of the pattern may be an abstract analogue of some aspect of the object. In any case, my account of the content of mental representations applies to neural networks in so far as they instantiate digital substitutes.

limited, compared to the capacities of what is substituted; in other words, the vehicle acting as a digital substitute need have no similarities to the substituted object that are relevant to the vehicle's ability to serve as a limited version of the object. An analogue substitute is *minimally* so limited, which means that this substitute's ability to serve as a limited version of something else depends on the object's sufficient similarity to the substituted object. This similarity, though, makes the analogue substitute a near-replacement that lacks semantic properties, whereas the maximally limited substitute bears a semantic relation to the substituted. This is because the substitutionary relation between a digital substitute and what is substituted stands out as an asymmetric relation, without being based on a relation of similarity or isomorphism.[10]

Assuming the human brain is a digital computer in my limited sense, in that the brain uses digital substitutes as mental representations, a negative reason can be given for saying that the content of these representations is *normatively determined.*[11] There's no descriptive law about what a system would have to do with this symbol under any physical or otherwise objective set of conditions. Again, almost any combination of neural structures could serve as the vehicle of a digital mental symbol, because the

---

[10] The distinction between digital and analogue substitutes is consistent also with Giere's point that the respect and degree of similarity between a scientific model and what is modeled can be specified. See section 5.4.

[11] There's a large literature on arguments for and against the claim that ascriptions of content to linguistic or to mental symbols are normative. Some arguments for the claim are found in Kripke (1982), Brandom (1994), and in Gibbard (2003). Contrary arguments can be found in Glüer and Wikforss (2009), Hattiangadi (2006), and in Rey (2007). Many of the arguments for the claim, especially the Wittgensteinian ones, refer to *social* norms, and take these norms to pose a problem for naturalistic theories of content. This sort of argument must assume that social norms and symbols are somehow interdependent, since social practices surely depend on the use of symbols by the society's members. I assume, on the contrary, that an explanation of the content of mental symbols is deeper than an explanation of social practices, since the individuals that make up social groups already have these symbols. Therefore, the norms I take to determine the content of mental symbols aren't those that govern social practices. Other arguments point to norms of rationality, but these arguments tend to take the sentence as the primary semantic unit, and truth as the primary semantic relation. I'm interested in the substitutionary aspects of each symbol, and I think there are separate, and indeed primary norms that govern the use of mental symbols as substitutes.

symbol is used merely as a digital substitute, not as a near-replacement. A digital symbol is such a limited version of what is substituted, that the vehicle's own properties are irrelevant to this substitutionary aspect of the symbol.[12] Anything with such arbitrary objective properties can't enter into a descriptive, scientifically discovered nomic relation. Nevertheless, there are patterns of the use of digital symbols as substitutes, and these patterns are explainable in general terms. For example, a person often thinks about how to modify part of her environment, and the thinking may produce an attempt to bring about the modification. There are patterns of some sort that hold between logical, emotional, or other uses of mental substitutes and bodily efforts to affect what is substituted. It's just that these patterns aren't explainable in terms of objective nomic relations.

So assuming there are only descriptive and prescriptive types of generalizations, the patterns that emerge from the use of a digital substitute must be only prescribed. And so the digital aspect of mental substitutes accounts for their difference from substitutes that clearly have no semantic properties, and points also to the normative determination of the content of these substitutes.

When looking at this negative reason for thinking mental content is normatively determined, as opposed to being determined by asymmetric dependence or by a so-called descriptive norm, it's important to keep in mind the difference between the descriptive and the prescriptive. I think the crux of the difference is what I say in section 5.2.1. In the case of patterns, regularities, or other explainable phenomena, the explanation is descriptive if it says what always must and thus does happen or what would happen were

[12] I grant later in this section that the vehicle's properties do affect the symbol's other substitutionary aspect, that is, the symbol's usefulness as a means of dealing with the substituted object.

certain objective conditions in place. The conditions are objective in that they are what they are, and they are either met or not, regardless of any subjective evaluation of them. For example, regardless of anyone's feelings or thoughts about gravity, stepping off of a building's roof causes the body to fall.

What's posited by a prescriptive explanation, though, is in part a normative connection between *relata* that isn't necessitated by the meeting of any set of objective conditions, and that can therefore be instantiated even without all of its *relata*. For example, someone can have an obligation to perform an action and yet fail to perform it, because circumstances are such that the person is inclined to choose a different course. Even though the action doesn't occur, the person is actually related to this action, by way of having an obligation to perform it, and this is so even under the very conditions that are unfavourable to the person's fulfilling the obligation. Contrast this with the instantiation of an objective nomic relation. For example, there's a relationship between certain variables that determines the behaviour of a gas, as stated by the Ideal Gas Law. These variables are a gas's volume, amount, and absolute pressure and temperature. When the ambient conditions are abnormal, the actual behaviour of a gas will only approximate that of the ideal gas. The natural law applies only counterfactually to situations in which the special conditions aren't met, since the law says that *were* the conditions met, the nomic relation *would* be instantiated. But this means that the nomic relation itself isn't instantiated when the conditions aren't met. The described nomic relation is the product of the coming together of a set of conditions, and so is found only

when they do come together. When only some of the conditions are met, what is instantiated is only an approximate version of the nomic relation.[13]

When a theory becomes complicated and ever more objective conditions must supposedly be in place for a nomic relation's instantiation, to explain empirical evidence that appears to conflict with what the theory predicts, this is a reason to think there's no such nomic relation. Either the nomic relation is usually instantiated in most observable situations, in which case there should be only rare exceptions to the law, or the relation should usually be instantiated under a set of rarer, experimentally controlled conditions. Either way, the best reason to believe a descriptive natural law is true is that the law is confirmed by empirical evidence. This isn't so with regard to a prescriptive law. It might be the case that people ought always to tell the truth even were most people to lie most of the time, and this might be so even were there no way of ensuring that anyone always tells the truth. In the case of a prescribed relation, the regularity may be just that some pattern is usually found to be somehow incomplete. Still, despite the rarity with which a prescription may be actually fulfilled, the ideal, norm, or standard applies even when conditions are such that the prescription can't be fulfilled. There is no algorithm for fulfilling a prescriptive norm, nor is there a scientific model of the objective conditions under which the prescribed event would occur, since there is no combination of material forces that necessitates the fulfillment of what's prescribed. (Were there such a combination, the prescription would be empty and the event would be explainable in purely descriptive terms.) But there is nevertheless some prescribed connection between

[13] A more striking example of a context-sensitive nomic relation can be found in the experience of unrequited love. Just because one person loves another doesn't mean the feeling is mutual. There are special nomic relations that hold between two people who love each other, but these relations are instantiated only when both *relata* meet certain conditions.

the actual and the possible *relata*, between the person and the action—not just an approximation of the ideal state of affairs, but the norm itself that governs the situation regardless of what actually happens. At least, the realist intuition is that any objective change in human history, for example, would not have brought a change in the prescriptive norms that govern people's actions.[14]

So a normative connection, such as an obligation to perform an action, is found even where conditions are unsuitable to the instantiation of the prescribed relation, whereas an objective nomic relation isn't found where conditions are unsuitable to its instantiation. A nomic relation between properties isn't like a puppet master who is present even when one of the strings breaks and the puppet doesn't perform according to the script: when the variables aren't actually related according to the natural law, there is no nomic connection between the token objects. But there is an actual prescribed connection between certain tokens even when some of these tokens don't actually exist or when objective conditions are such that the prescribed relation can't be instantiated: the norm somehow applies even when the norm isn't fulfilled.

I've said that there are no objective constraints on a digital symbol's ability to stand in for something else as a limited version of it, and relatively few such constraints on the symbol's usefulness as an instrument. The properties of the vehicle that are relevant to its ability to substitute digitally for what the symbol is about are arbitrary, compared to the properties of what is substituted. This doesn't amount, though, to saying

---

[14] See Stern (2004) for a Kantian principle to the contrary, that a prescription applies only to what has the *capacity* to fulfill it. This principle has the consequence that there should be no negative evaluation of something that can't help but violate the principle. A tornado that kills innocent people, for example, isn't bad. But the same would have to be said of a psychopathic killer. If evil is the necessity of violating prescriptive norms, the Kantian principle implies that there's no evil. If, instead, a psychopathic killer deserves condemnation, a prescription doesn't depend on the objective capacity to live up to the norm.

that the property of being a mental representation is *multiply realized*; rather, the point is that this property can't enter into a scientifically discovered nomic relation, because there's not even a single *mechanism*, as such, that realizes the property. After all, the mechanism would have to be a complex system of causal relations that works under some objective conditions, and these conditions would have to determine, in a bottom-up, synchronic way, the realized nomic relation.[15] When the conditions are not met, the relation isn't instantiated, because the nomic relation is supported by the mechanism's work. A nomic relation between multiply realized properties is supported by different mechanisms, but each instantiation of the relation is supported by *some* mechanism and thus determined by the conditions under which the mechanism works according to its own CP law.

But there can be no such mechanism underlying the use of a type of digital symbol. That is, there can be no mechanism accounting for why one configuration of neural elements is used as a token of this symbol type when, given the arbitrary nature of a token of any digital symbol, an indefinite number of other configurations might have served just as well, and indeed do so serve for other users. Someone's accumulation of a largely accidental set of life experiences accounts etiologically for why some combination of neural elements serves as her token vehicle of a mental symbol type. And there's no mechanism for forming a digital mental symbol type that makes that sort of token accumulation an instantiation of the mechanism. A mechanism that implements a multiply realized property operates in the same way for all instances of a kind, all things being equal. Thus, there might be one mechanism for implementing *pain* in human brains, and different mechanisms for implementing *pain* in Martian and in artificially

[15] See my earlier discussion of multiple realization, in section 2.9, especially footnote 32.

intelligent control centers. There is no such mechanism, though, for implementing a mental symbol type in human brains, given the symbol's digital substitutionary aspect, since there are no relevant descriptive generalizations that could account for the heterogeneity of the neural vehicles. Even were there such generalizations, a digital mental symbol wouldn't be determined by the properties of its vehicle, since its ability to stand as a limited version of something else is independent of the vehicle's distinguishing features. But that which realizes a property does determine the property. So the property of being a digital, substitutionary mental symbol wouldn't be multiply *realized* by the vehicle's properties.

No theory that posits only objective types can account for a type with such arbitrary conditions of instantiation as those of a digital mental substitute. For a similar reason, Fodor (1990) says that a symbol's arbitrarily robust content can't be determined by any one nomic relation. Arbitrariness may be found in either the obtaining of a semantic relation despite the lack of objective conditions that necessitate its instantiation, or in the ability of something to be used as a digital substitute despite the lack of objective constraints on its instantiation. In each case, the arbitrariness points to a prescriptive determinant of content. A nondigital multiply realized property can also have a variety of ways of being instantiated, but the mechanisms must have in common some ability to realize the property. Moreover, this ability can't be defined just as the ability to realize the property; rather, the ability must have some independently specifiable limits such that some mechanisms can't realize the property. A property that could be realized by any mechanism at all isn't a scientifically posited property.[16] But there's no such ability shared by the possible instantiations of a digital substitute, because this substitute

[16] Again, see my earlier discussion of multiple realization in section 2.9.

is indeed instantiated under objectively arbitrary conditions. That's why the semantic relations into which this substitute enters can be only normatively, not objectively determined.

I hasten to add that this isn't to deny that digital mental symbols are token-identical with combinations of neural elements or that these combinations work in mechanistic ways at the neurological level. My point is that the use of these vehicles as digital substitutes isn't realized by these mechanisms in the way that a relation between multiply realized properties is realized by mechanisms. The neural mechanism that realizes *pain* sets the limits of the higher-level property in a way that no neural mechanism can determine the property of being a digital mental substitute. This is because the vehicle of this substitute has arbitrary intrinsic properties, with respect to the substitute's ability to stand as a limited version of its substituted type.

Having said this, I should grant that there is probably one objective constraint on tokens of a digital *mental* symbol, that is relevant to the symbol's other substitutionary aspect, to its usefulness as an instrument. This objective constraint is that the symbol tokens have to be part of a *brain* or at least of some material control center. Some of the differences between the neural instantiations of a type of mental representation aren't like the difference between a blue cloth and a red cloth, which can each be used equally well to wipe a windshield, under most circumstances. Rather, some ways of thinking about something affect the way the neural state is used as an instrument for solving a problem with what is substituted. Each type of mental representation is instantiated in any of a wide variety of neural nets, depending on the accumulated and processed mental associations. These associations affect a representation's usefulness, by causing different

types of outward behaviour, and some of these behaviours may indicate that the symbol stands in for one type rather than another. This isn't to say that the intrinsic properties of the neural vehicle determine the mental symbol's extension. The point, rather, it that the neural vehicle is more arbitrary with regard to its first substitutionary aspect, of serving as a limited version of something, than to its second such aspect, of being used as an instrument in dealing with something. This is because a symbol deals with what is substituted by causing outward behaviour, and some—but not all—of the otherwise arbitrary differences between the neural vehicles cause different types of outward behaviour.

Again, the neural properties are irrelevant to a brain's ability to carry relatively limited versions of objects in the environment, since these versions are arbitrarily different from those objects. That's mainly what makes the substitutes digital. Moreover, even the one constraint allows for billions of distinct tokens of each mental symbol type, so even if something which isn't a material control center could be objectively excluded as a user of digital mental symbols, there could be no ruling-out of any combination of relevant neural elements as such a token. A person's brain is perhaps the most plastic, or mutable, sort of object there is, allowing for unique connections between billions of working parts, formed in large part from the person's accidental experiences. One person may associate dogs with the orange dog house that housed the person's second dog, Skipper, whereas another might associate dogs with a peculiar noise this other person once heard a dog make yawning in a limousine. Given the extent of the variability at the level of neural vehicles, none of the combinations of neural elements realizes, in the sense of supporting and objectively determining, a digital symbol type's standing in for

something else. The digital symbol's ability to stand as a limited version of something else is entirely unaffected by the distinguishing features of the symbol's various neural vehicles. This is just what it is to be a digital symbol with a substitutionary aspect.

Without a material control center there would be no substitutionary use of a mental symbol, since there would be no mental symbol and thus no content. Likewise, without persons there would be no prescribed actions and no failures or successes relative to an ideal. Moreover, changing a particular brain surely changes the mental representations instantiated in that brain. But there's no type of neural mechanism found in all human brains such that were this mechanism changed, there would be a corresponding change in a mental symbol type found in all of these brains. This is because of the sheer variety of neural instantiations of the same type of representation. In short, the difference between a multiply realized property, such as *pain*, and the property of being a digital symbol is that the former has ***multiple*** ways of being instantiated, as determined by a set of nonarbitrary objective conditions of a certain mechanism's operation, whereas the latter has—at least in certain respects—*objectively arbitrary* instantiations.

The problem here is similar to the problem faced by a causal role theory of mental content.[17] According to this theory, a symbol's content is determined by the descriptively normative purposive function the symbol fulfils within a system of symbols. The function at issue is supposed to be not only objective but internal to a system of symbols, and so the theory must identify a symbol as some sort of neural structure, as such, not as anything necessarily bearing an external relation to a referent. As I argued in Chapters 3 and 4, there is no such thing as a descriptively normative function. But even if there were,

[17] See, for example, Block (1986).

presumably the function at issue here could change depending on the shifting of background symbols in the system. Different systems, and indeed different moments in any one system, with different background symbols determining the role and thus the content of any symbol in the system, would have to result in a special content for that symbol, and so content would be holistic rather than atomistic. That is, the content of any one symbol would depend on there being other symbols with content of their own.

Fodor and Lepore (1992) object that such a holistic theory of content can't account for the evident use to which symbols are put in communication. The theory can't account for the ability of very different vehicles of content to serve as tokens of the same symbol type, when these vehicles are taken to include the whole system of symbols that determines a symbol's causal role. Of course, the empirical reason to think there's a symbol *type* at issue, given the various combinations of neural elements that would have to serve in different brains as vehicles of a symbol type, is that communication does happen and patterns of behaviour do emerge.[18] But there would be no accounting for this empirical evidence were content determined holistically by causal role, and thus there would be no basis for positing any symbol type instantiated in different networks defined by the context-sensitive causal role of their nodes. That is, there would be no basis for generalizations about symbol use. I would put this point by saying that there's a contradiction between the claims that the content of mental symbol types is determined objectively, by a *causal* role, and that these symbols are *digital* in that there's an arbitrary variety of objective properties their vehicles can have. Given the arbitrary variety and the

[18] Communication is aided, of course, by the use of external digital symbols in natural language. These symbols focus attention on conventional definitions of meaning, leaving aside the many idiosyncratic associations taken to be significant by individual speakers. These external symbols are also digital in their substitutionary aspects, since they're arbitrarily dissimilar to their referents.

facts of communication, of behavioural commonalities, and thus of the instantiation of the same symbol type in such a variety of vehicles, content can't be determined objectively, by causal role.

I've argued in this section negatively that, given the digital aspect, the determinants of the content of mental representations can't be descriptive, or objective, and that therefore they are prescriptive, or normative. In the next section, I lay out another difference between mental substitutes and substitute teachers, and the like. This difference will provide for a positive reason to think the content of mental representations is normatively determined, and for a nonmechanistic account of the production of mental representations.

6.4 The Substitutionary Use of Mental Representations

In this chapter, I've assumed that a mental symbol is identical to some combination of neural elements, but I've argued that the symbol as such isn't realized or implemented, and thus determined, by such a combination or by any mechanism or objective process. A mental symbol is digital in the sense that the intrinsic properties of its vehicle are irrelevant to the symbol's standing as a limited version of something else. As a combination of neural elements, of course, a mental symbol follows the electrochemical laws of neuroscience, but as something that stands in for something in the environment, the symbol's objective features are objectively arbitrary. What then does determine, or set the limits of, the substitutionary relation between a digital mental symbol and the external object (or type)? I gave a negative reason to think the

determinant is a prescriptive norm, which is that the determinant isn't the opposite, such as a mechanism or some objective process or nomic relation.

I think enough has been said, though, to suggest there's a positive reason as well to think mental content is normatively determined. There is, after all, a certain convergence of themes here regarding arbitrariness. In the same way that the asymmetric dependency posited by Fodor would have to be a prescriptive norm to account for arbitrary semantic robustness, that is, for a semantic relation's being underdetermined by any causal relation, the arbitrariness of a digital symbol indicates the work of a prescriptive norm. A norm is arbitrary, given what happens in the physical world, in that the norm isn't objectively determined; the norm could apply regardless of whether it's ever fulfilled and of whether there's a means of necessitating its fulfillment. This is the basis of the intuition behind the naturalistic fallacy. The difference between prescriptive and descriptive properties isn't just the difference between higher- and lower-level properties, given token, or nonreductive, physicalism. The idea isn't just that prescribed relations are multiply realized, so that the prescriptive laws are irreducible, say, to some lower-level ones. An objective property realized by a lower-level one supervenes on the latter, in that a change in the lower-level property would necessarily bring a change in the higher-level one. But the arbitrariness of prescribed relations, from an objective viewpoint of what actually is the case or of what would have to be so under certain conditions, means that these relations aren't so dependent on objective relations. The rightness of some objective state of affairs, according to an ideal or a standard, doesn't depend on any objective property.

In the same way that objective facts are independent of their discovery by creatures, prescriptive norms are independent of both objective facts and the discovery of the norms. By definition, then, a prescriptive norm isn't realized or implemented, and thus determined, by a physical system. Were this not so, a prescriptive norm would be a descriptive one and the prescription would be equivalent to a description. So prescriptive norms are arbitrary in that they're independent of objective facts. Meanwhile, digital mental symbols are arbitrary in that the intrinsic properties of their vehicles are mostly irrelevant to their substitutionary, and thus to their semantic properties. Assuming a substitutionary theory of the intentionality of mental symbols, an argument by analogy suggests that a semantic relation is also prescriptively determined.

In any case, the positive reason I want to focus on in this section stems not from this point about arbitrariness, and thus not from the first of the two substitutionary aspects, but from the second, instrumentalistic aspect. Recall that the methodological approach to explaining content in naturalistic terms finds its resource in scientific practice and thus in the use of scientific explanatory models. This sort of model (1) is a simplified version of what it stands in for, to some degree or other, depending on the type of model, and (2) is often used as an instrument to control the modeled. I take these to be two substitutionary aspects of symbols in general. Now, the point about the arbitrariness of the intrinsic properties of some substitutes follows from the claim that a substitute is a simplified, limited version of something else. This arbitrariness indicates that digital symbols are *somehow* prescriptively determined. The other claim, about the instrumental *use* of substitutes, points to the source of the prescriptive norms that determine semantic relations.

The use I'm speaking of is the stimulus-independent processing of mental symbols, which Bickerton (1995) calls offline, as opposed to online or stimulus-dependent, thinking. In Bickerton's words, "on-line thinking involves computations carried out only in terms of neural responses elicited by the presence of external objects, while off-line thinking involves computations carried out on more lasting internal representations of those objects." These latter computations "need not be initiated by external causes, nor need they initiate an immediate motor response" (90). Bickerton points out that "off-line thinking was impossible until there existed areas of the brain where new information could be processed without needing to be triggered by environmental input and without invoking immediate behavioral consequences" (59). So the use of mental substitutes is relatively detached from perception and from motor response; a mental representation is a stand-in for something else, giving it a semantic as opposed to a causal, perceptual relation to the external object, when the representation isn't just an extension of the object, as it were, processed online, but an alternative to this object.

Dennett (1995) draws the distinction in a more fine-grained way, in terms of the evolutionary value of different design options for brains. He distinguishes between what he calls Darwinian, Skinnerian, Popperian, and Gregorian creatures. Darwinian creatures aren't relevant to the distinction at issue, but the point about them is that they don't think for themselves: their behaviour is hardwired and the direct result of trial and error in the process of natural selection itself. Skinnerian creatures are capable of learning, but only by trial and error, blindly trying out different behavioural responses to the environment until one is selected by reinforcement, causing the creature to repeat the response under

similar conditions. By contrast, Popperian creatures have "an inner selective environment that previews candidate acts." This inner environment "is structured in such a way that the surrogate actions it favors are more often than not the very actions the real world would also bless, if they were actually performed." And so the creature acts with foresight as opposed to blindly responding in different ways to the outer environment itself and suffering from these direct tests of its behaviour when they fail to get what the creature wants. Most mammals aren't pure Skinnerian, online creatures, since they have some capacity to plan and to think in other ways offline. The mental states of a pure Skinnerian creature should lack semantic content, in so far as this content must have the two substitutionary aspects, since the mental states would be more like extensions of stimuli than stand-ins for them. However, most mammals also aren't as sophisticated in using their internal environment as are humans. And so Dennett posits the Gregorian creature, named after the psychologist, Richard Gregory, who speaks of the role of stored, potential intelligence in artifacts in the creation of behavioural intelligence. The idea is that, compared to most Popperian creatures, humans have better mind-tools, such as language, which Dennett thinks are imported from the *cultural* environment and which give us more access to our minds, allowing for more sophisticated planning (374-8).

I think Dennett is right to suggest that offline operations on mental substitutes are "surrogate actions" in an internal, mental environment that, in some ways, reflects the outer one. The internal environment is populated, as it were, by representations that stand in for outer objects. But these stand-ins would provide no advantage to a Popperian or to a Gregorian creature were the creature incapable of using them as alternatives to their complements in the outer, real environment. Mental representations model what would

happen were the creature to act in some bodily way. But useful models require stand-ins also for the forces that determine the outer environment's responses to the creature's actions. So, for example, instead of imagining a dog in isolation, an offline thinker applies rules of inference to the representation, as in a thought experiment, or else relives a memory or examines a feeling about dogs, and these ways of using the representation themselves stand in for what the outer environment can do, as it were, with dogs or with a creature that interacts with dogs. Specifically, the offline processes have the first but not the second substitutionary aspect, since these processes *are* the uses of the mental substitutes that have content, instead of being used themselves by still more offline processes, leading to an infinite regress. Some digital mental substitutes have content, others don't, depending on whether they have one or both of the substitutionary aspects.[19]

Besides the explanatory gap, then, between a digital substitute and the substituted, there's another such gap between an external process and the offline processing of a symbol. As a near-replacement, an analogue substitute has objective similarities to the substituted, and to this extent there's no semantic relation between the two. There's no such objective relation between a digital substitute and the substituted, since this substitute has arbitrary intrinsic, that is, internal objective properties. Given that there's nevertheless at least the substitutionary connection between the combination of neural elements and some external object, some prescriptive norm seems needed to fill this

[19] This point about offline thinking being itself a substitution for outward action raises the question of how the sort of insanity that results in hallucinations can be explained on a substitutionary account of content. The distinction between the symbol and the referent is built into this account, since the symbol is regarded as a more or less limited version of the referent. In an hallucination, this distinction is lost and a representation is mistaken for an external object. Whatever the cause of this, it's not that offline cognitive processes are themselves ordinarily treated *as if* they were bodily actions, with the absurd result that all mental representations are hallucinations, on my view of content. Sane offline thinkers know implicitly that moving in the mind from one thought to the next, for example, isn't the same as moving the body from one place to the next, even though the mental process may stand in for the bodily action, giving the representations their content by the substitutionary use of them.

explanatory gap. Likewise, cognitive processes are digital versions of other processes, so there's another explanatory gap, from an objective viewpoint.

Compare this latter gap to that between sets of nomic relations posited by different special sciences. What happens is that physical processes evolve and become more complex, but in a variety of ways depending on the conditions in different times and places. Thus, different sets of special, separately evolved natural processes are like the peaks of separate mountains or like the behaviour of species that evolve on different continents under very different environmental pressures. Different sets of specialized processes can become so complex that no sense of each, in its own terms, can be made using the terms that account for another of the sets; common explanatory ground between them is lost, and the one set of processes has nothing to do with the other. In the same way, the inner, mental environment, with its digital population of substitutes and offline processes, develops in its own way, separate from the outer environment. There is a crucial difference, though, which is that there is no need to fill the explanatory gap in the case of the relation between the domains of most special sciences. In the case of psychology, however, there is a need to say how mental entities relate to external ones, such as to the nonmental entities posited by other special sciences. This is because there's a pattern of inner and outer substitutionary activity linking a mind to things in the outer world. A mind has inner substitutes that are used to affect what they stand in for by causing the body to act. These mental entities are directed towards other things, which is just to say that there are semantic relations between the two sets. It's as if one mountain peak were to have signs on it that point to parts of a different mountain peak. Indeed, from the perspective of, say, metallurgy or meteorology, a Gregorian creature's mental

world, in which symbols are processed offline, is *doubly* removed and arbitrary. Not only are the substitutionary neural nets and processes digital, and thus lacking in resemblance to what they stand in for, but the mental development happens in a cultural cocoon, aided by mind-tools that derive from, or at least are impacted by, a technologically-transformed natural environment.

If there's no objective link between the offline mental world and the outer world, again there seems at best a normative relation between them. But a *positive* reason to think such norms are needed to explain the link in specific cases, and thus to explain the determinacy of mental content, is that the offline cognitive processes are arguably—from a certain influential Kantian perspective, at least—what makes a mind an autonomous, relatively self-controlling being, subject to prescriptions in the first place. The more isolated thought processes are from perception and from the motor system in a brain, the more a mind can control itself and its body, as opposed to having its behaviours be immediate or conditioned responses to stimuli. It's precisely the actions of a Gregorian and, to a lesser extent, of a Popperian creature that call for a prescriptive explanation. It's precisely such a relatively autonomous creature that might perform an action without being compelled to do so under any external circumstance. An autonomous creature is necessarily detached in certain ways from things in its outer environment, since the two operate independently, and so it's just such a creature that might also lack a way of ensuring that its self-controlled actions are successfully performed.

Here, then, is another difference between semantic and nonsemantic substitutes. Even though both may be normative, as in the uses of a mental representation and of a substitute teacher, the norms that govern the use of the former are *primary* in that this use

is the reason there are any other norms. There would be no offline thought processes without mental symbols which they process, and these symbols must have substitutionary aspects, standing in for other things instead of being dependent on stimuli and instead of causing reflexive, as opposed to more open-ended instrumental behaviour. Moreover, there would be no mental symbols, having both substitutionary aspects, without some offline thought processes. Thus, the norms that govern the substitutionary use of mental symbols are those needed to account for the *subjective* precondition of all norms except for those that arise with the first act of self-determination, which is the inner activity of offline thinking, or the use of mental substitutes.[20] Substituting mentally for things in the outer world is how creatures seem to establish their relative autonomy in the first place, and the more autonomous a creature is, the more fitting it is to explain its behaviour as being guided by rules it can follow or violate regardless of the outer conditions. The basic rules are those that give content to the mental substitutes, since without these substitutes, there would be no offline thinking, and thus no self-control and no basis for positing other norms, such as those of outward, bodily actions.

Instead of being mechanically implemented or dependent on a Normal situation, a mental symbol is used by a mind, by a Popperian or a Gregorian creature. And instead of being objectively determined, the symbol's substitutionary relation is determined by a norm that governs the symbol's use in offline thinking, which shapes the neural vehicle, thus determining indirectly the symbol's use in outward behaviour. If mental content is

[20] I'm distinguishing here between two sets of prescriptive norms, those that govern the inner activity that is the precondition of all other norm-governed activities, and the norms that govern these other activities. For example, ethical norms are posited on the assumption that there are self-determining creatures, and thus, I think, on the assumption that the creatures are already engaged in offline thinking. But this inner activity of offline thinking is already governed by norms that determine the substitutionary relation between mental symbols and what they stand in for. The norms that determine mental content have to be posited to account for the initial, primary act of self-determination, which is the use of mental substitutes that separates the Skinnerian from the Popperian and the Gregorian creatures.

normative, there must be a standard substitutionary, and thus semantic, relation between each mental symbol type and its referent. The offline processing of a mental substitute isn't mechanistic, or objectively determined, but this processing helps form the substitute's neural vehicle, counting some mental associations (memories, perceptions, feelings, skills, and so on) as relevant, others as not. Assuming offline thinking makes a creature relatively autonomous, whose activity is thus governed by prescriptive norms, and the offline processing of a digital substitute counts as an inner activity which is thus governed by a norm, the substitutionary relation is normatively determined. Assuming this substitutionary relation is the semantic relation between the mental symbol and its referent, the semantic relation is also normatively determined.

A prescriptive norm holds regardless of whether the norm is ever fulfilled, or of whether there's a way of necessarily fulfilling the norm. That is, the standard substitutionary relation between a symbol and its referent is constant, regardless of how the symbol is actually used. There's a norm for each type of mental representation, one for DOG and another for FOX, for example, and while this norm may have fuzzy boundaries, it includes some uses in offline thinking and in outward behaviour as right, and excludes others as wrong. The use of any neural vehicle as a digital substitute is governed by the prescription that Popperian and especially Gregorian (human) creatures are supposed to model the outer environment in their heads, and to act in that environment as dictated by the internal processing of those mental substitutes. Some uses count as poor from a substitutionary perspective, depending on whether the uses run afoul of the norm. Evidence that a mistake is made in the symbol's use is found in the symbol-user's outward behaviour, that is, in the way the symbol is used as an instrument to

control or otherwise to deal with what is substituted. Someone with a mental representation of tigers doesn't deal well with tigers by treating them as harmless animals; that is, some neural vehicles can succeed or fail to serve as instruments in dealing with what is substituted, as discovered by the outward behaviour caused by the offline processing of these vehicles.

Here, then, is how a case of misrepresentation can arise on my account. Suppose someone has a mental symbol referring to dogs, and another referring to foxes. The vehicles of these symbols consist of some overlapping neural material, including memories, feelings, images, and so forth. For this reason, the one neural vehicle might be tokened by perception of a fox; in any case, either symbol token might be caused in various ways, by semantically relevant or irrelevant objects. But each assortment of neural materials retains its semantic relation to some distal type in virtue of the assortment's standing in a substitutionary relation to the type. This substitutionary relation is normatively determined. So *some neural material makes for a better substitute for dogs than for foxes*, and for this reason the material counts as a symbol bearing a semantic relation to something else. The neural vehicles for DOG and for FOX each have arbitrary intrinsic properties, with respect to the ability of these vehicles to stand as limited versions of their referents. So in this respect, either vehicle could serve just as well as the other as a symbol referring to the other's referent. But the different neural associations making up each of the neural vehicles do affect each vehicle's usefulness as an instrument that deals with the referent, by entering into different offline thoughts and by causing different outward behaviours. For example, one of the neural vehicles may include the memory of petting a certain domesticated animal, and this memory works

better as part of a means of dealing with dogs than as one of dealing with foxes. Content is determined not by the neural differences themselves, nor by their internal causal roles, nor by a vehicle's tendency to cause certain outward behaviours, but by these things' fulfillment of a norm governing the substitutionary process. Some set of neural associations makes for a better mental substitute for dogs than for foxes, by making the set a better means of dealing with a certain distal type, despite the randomness and otherwise arbitrariness of the associations. So a fox may cause a perceiver's brain to process neural material that nonetheless refers to dogs, not to foxes, because the material is better suited as a substitute for dogs.[21]

I want to distinguish what I'm saying here from some claims I don't wish to affirm. While offline thinking and outward behaviour are needed for a substitutionary, and thus for a semantic, relation between a mental symbol and its referent, the symbol's content isn't determined by these activities nor by objective conditions under which these activities tend to occur. Instead, the activities are determined by a prescriptive norm. For example, ways of thinking about, and behaving towards, dogs are right or wrong, depending on the normative connection between some set of neural vehicles and dogs. The semantic relation between the mental symbol and dogs is made up of (1) the standing in of some token neural vehicles as digital substitutes for dogs, and (2) the use of any of these vehicles, in offline processing and outward behaviour, as an instrument in dealing with dogs. So part of what directs something in the head towards something out in the

[21] A symbol's content is *epistemically* determined externally by reasonable judgments about which mental substitutes a creature has, given its outward behaviour. The content may also be so determined internally by the symbol-user's own offline processing, which analyzes a concept and arrives at a reasonable judgment about which mental substitutes the user possesses, given the available mental associations. But content is *metaphysically* determined by a prescriptive norm that has to be posited to account for the semantic relation between two things that lack an objective, physical connection. Even though neural vehicles are arbitrarily different from certain distal types, the former semantically relate to the latter by standing in for them, and this mental substitution is subject to an instrumental standard.

world is the bodily action taken towards the latter as caused by mental processes. But I'm not proposing a use theory of content nor a behaviourist one. On my view, mental content isn't determined by use or by behaviour, but by a prescriptive norm governing not just instrumental uses but the other part of the substitutionary relation, which is one thing's standing in as a limited version of something else. An analogue substitute is a limited version of something else if the two are objectively similar to each other. A digital substitute is a limited version of something else if the substitute is a form of the other thing despite the substitute's having objectively arbitrary properties compared to those of what is substituted.

Also, although there is one content-determining prescriptive norm for each mental symbol type, the norm isn't an *ideal* in the sense of one, narrowly defined substitutionary relation that makes for just one correct use, with all other uses being better or worse compared to the ideal one. Just as there are a variety of correct ways to use "dog," the norm governing the correct uses of a type of mental substitute includes a variety of uses as correct. Moreover, the billions of differences between neural vehicles of a symbol type affect the symbol's usefulness as an instrument, but not what counts as the symbol's extension. The norm determining the content isn't itself determined by the intrinsic properties of the neural vehicle. On the contrary, the vehicle is shaped in part by offline cognitive processes that are themselves governed by a prescriptive norm.[22]

[22] I've assumed throughout this dissertation that if norms aren't descriptive and objective, they're prescriptive and subjective. I've said more in this chapter about what I think prescriptivity and subjectivity involve. But I've also said in previous chapters that prescriptive norms are evaluative, that a norm isn't value-neutral. I won't say much here about what I think values are, but I would be inclined to derive an account of them from an account of self-determination.

6.5 An Instinct for Using Substitutes

I want to respond now to an objection to what I've said about the instrumentalistic aspect of semantic relations. What I've said might seem to lead to an infinite regress, since the use of a symbol is goal-directed and thus might require an explicit representation of the goal, which I then would have to explain in similar substitutionary terms. In this case, I will have explained, at best, the content of some symbols in terms of the unexplained content of others. This is indeed a problem with pragmatic theories of content, especially with ones that take content to be determined by social norms, given that these norms already depend on the use of symbols.

However, I do have a response. Clearly, the use of something, in general, as a means of dealing with something else might proceed with a representation of the goal. However, some such use may instead be instinctive, requiring no such explicit representation. It so happens that there is an instinct for substitutionary behaviours in most animal species. Following Lorenz, ethologists such as Théry and Heeb (2008) call these behaviours "ritualized" in that the behaviours are, in some sense, empty gestures that nevertheless serve as compromises dealing with stress or frustration, or as resolutions of a conflict between opposing motives (607). There are many examples of these behaviours in intraspecies fighting, displays, and other means of redirecting aggression. When members of an animal species fight each other, they usually engage in escalating forms of competitive wrestling rather than fighting for the sake of killing. This protects each of the members and the gene pool, and so the behaviours are naturally selected. For example, sierra dome spiders enter into a tournament with three phases, from

noncombative displays, to cooperative wrestling, to potentially fatal fighting; only if the loser doesn't withdraw in an earlier stage do they proceed to the next stage. Opposing rattlesnakes don't inject each other with venom, but press against each other, trying to force the other to the ground. Gazelles lock horns in a shoving match, trying to prove which animal is stronger without injuring the other. Another example is the gorilla's displays of aggression when it beats its chest, bears its teeth, or shakes a tree. In each case, the gorilla selects a middle option between fleeing and fighting or killing. Yet another example is a person's slamming of her fist into a car horn, by way of rechanneling aggression.

So my response is that the use of mental symbols in offline thinking isn't deliberately substitutionary, although a sane offline thinker must have an implicit understanding of the difference between a representation and what is represented. It's just that while most members of animal species reflexively use their physical bodies in substitutionary ways, Gregorian creatures such as ourselves have mind-tools in addition to the tools provided by our bodies and by culture. Thus, we instinctively use our mental states, in addition to our bodies and our external artifacts, as stand-ins. There's no need to assume an offline thinker has the explicitly represented goal of trying to substitute for direct bodily action towards an external object. Most animal species engage in this same sort of behaviour without such awareness of what they are doing, and this can be true also of Gregorian creatures.

6.6 Conclusion

In this chapter and in the fifth one, I've sketched a naturalistic strategy for explaining content, that serves as an alternative to the one I criticized in the earlier three chapters. I've argued that intentionality, or at least the semantic relations between mental representations and what they represent, is fundamentally a substitutionary relation with the same sort of substitutionary aspects as those of a scientific explanatory model. This argument's conclusion *is* naturalistic; indeed, the conclusion is as naturalistic as the scientist's use of this sort of model is fundamental to the naturalistic worldview. The point is that a naturalist assumes not just an ontology, but a methodology, and the claim that mental representations are substitutes is just a generalization based on an analysis of the methodology. Indeed, the substitutionary aspects of symbols are left out of the three influential naturalistic theories I've criticized. The methodological strategy highlights these aspects. Another advantage of this strategy is that it avoids the obstacle in the way of carrying out the metaphysical strategy. That is, the nature of semantic relations is identified in such a way that it becomes at least possible for them to be normatively determined.

As for the prescriptive norms, I've tried to make sense of them in terms of digital substitutes, offline thinking, and relatively autonomous users of symbols. I haven't offered a reductionistic account of these norms, but I don't think a naturalist has to claim that all of the true generalizations about the natural world are about objective facts. I think prescribed patterns emerge when offline thinkers become sufficiently detached from their direct perceptual and behavioural links to the external world, and thus when

they become relatively self-controlling. Some of these patterns are found in the substitutionary, semantic links between what is used in offline processing and in outward behaviour, and what is substituted outside the offline thinker. A combination of neural elements stands in for an external object, the cognitive process for an outward bodily behaviour towards this object. These substitutionary patterns seem prescribed and normative, for the reasons I've given, which means the semantic relation between a mental representation and the represented is determined by whether the relation between a neural vehicle and an external object fulfills the standard governing the relation. Instead of potentially uninstantiated nomic relations as determinants, the determinants are prescriptive norms, or standard substitutionary relations. The difference is that instead of being realized by an objective mechanism operating in a Normal situation, a symbol entering into the prescribed substitutionary relation is necessarily used by a relatively autonomous creature. I don't think a naturalist needs to reduce talk of the prescriptive norms that govern the inner uses of mental representations, and thus the substitutionary relations, to talk of so-called objective purposive functions, as though offline cognitive processes were understandable just in terms of causal relations between types of objects.

Still, I don't claim to have fully accounted for prescriptive norms or to have shown that they pose no challenge to naturalism. But the main point of this dissertation isn't to show that substitutionary symbols are in fact normatively determined. Instead, the point is that, *given* that they are, there's a naturalistic strategy for explaining this fact. I think the assumption can be made that the determinants are prescriptive norms, if only because the three theories I've critiqued imply that the determinants are these sort of norms, even while the theorists overlook these implications. I've tried to show that the

normative aspect of mental content poses difficulty not so much for any naturalistic view of symbols, but specifically for one naturalistic strategy for explaining them, namely for the metaphysical one that has perhaps most influenced recent naturalistic discussion of intentionality. Shifting the focus to naturalistic methods has the prospect of providing for a naturalistic explanation of the substitutionary aspects of symbols and of the norms that determine mental content.

Bibliography

Antony, L. and Levine, J. (1991) "The Nomic and the Robust," *Meaning in Mind: Fodor and His Critics*. B. Loewer and G. Rey (eds.). Cambridge, MA: Blackwell, 1991. 1-16.

Baker, L.R. (1991) "Has Content Been Naturalized?" *Meaning in Mind: Fodor and His Critics*. B. Loewer and G. Rey (eds.). Cambridge, MA: Blackwell, 1991. 17-32.

Bedau, M. (1991) "Can Biological Teleology be Naturalized?" *The Journal of Philosophy*. 88.11: 647-655.

Begon, M., Townsend, C., and Harper, J. (2006) *Ecology: From Individuals to Ecosystems*. Fourth ed. Malden, MA: Blackwell.

Behe, M. (1996) *Darwin's Black Box*. New York: Free Press.

Bickerton, D. (1995) *Language and Human Behavior*. Seattle: U. of Washington.

Block, N. (1986) "Advertisement for a Semantics for Psychology," *Midwest Studies in Philosophy*. 10: 615-678.

Boorse, C. (1976) "Wright on Functions," *The Philosophical Review*. 85: 70-86.

Brandom, R. (1994) *Making it Explicit*. Cambridge, MA: Harvard.

Bratman, M. (1990) "Dretske's Desires," *Philosophy and Phenomenological Research*. 50.4: 795-800.

Brentano, F (1874/1973) *Psychology from an Empirical Standpoint*. A. Rancurello, D.B. Terrell and L. McAlister (trans.). New York: Routledge.

Buller, D. (1999) "Natural Teleology," *Function, Selection, and Design*. D. Buller (ed.). Albany: SUNY, 1999. 1-27.

Carnap, R. (1950) "Empiricism, Semantics, and Ontology," *The Linguistic Turn*. R. Rorty (ed.). Chicago: U. of Chicago, 1967/1992. 72-84.

Cartwright, N. (1983) *How the Laws of Physics Lie*. New York: Oxford U.

Cartwright, N. (1999) *The Dappled World: A Study of the Boundaries of Science*. New York: Cambridge U.

Chalmers, D. (1995) "Facing up to the problem of consciousness," *Journal of Consciousness Studies*. 2.3: 200-219.

Chisholm, R. (1957) *Perceiving: A Philosophical Study*. Ithaca: Cornell.

Chomsky, N. (1995) "Language and Nature," *Mind*. 104.413: 1-61.

Churchland, P.M. (1981) "Eliminative Materialism and the Propositional Attitudes," *Journal of Philosophy*. 78.2: 67-90.

Clark, A. (1997) *Being There: Putting Brain, Body, and World Together Again*. Cambridge, MA: MIT.

Cummins, R. (1991) "The Role of Mental Meaning in Psychological Explanation," *Dretske and his Critics*. B. McLaughlin (ed.). Cambridge, MA: Blackwell, 1991. 102-117.

Davidson, D. (1980) *Essays on Actions and Events*. New York: Oxford U.

Davidson, D. (1987) "Knowing One's Own Mind," *Proceedings and Addresses of the American Philosophical Association*. 61: 441-58.
Dennett, D. (1971) "Intentional Systems," *Journal of Philosophy*. 8: 87-106.
Dennett, D. (1987) "Evolution, Error and Intentionality," *The Intentional Stance*. Cambridge, MA: MIT. 287-321.
Dennett, D. (1991a) "Real Patterns," *Brainchildren*. Cambridge, MA: MIT. 95-120.
Dennett, D. (1991b) "Ways of Establishing Harmony," *Dretske and his Critics*. B. McLaughlin (ed.). Cambridge, MA: Blackwell, 1991. 119-130.
Dennett, D. (1995) *Darwin's Dangerous Idea*. New York: Touchstone.
Dennett, D. (2003) *Freedom Evolves*. New York: Penguin.
Dretske, F. (1977) "Laws of Nature," *Philosophy of Science*. 44: 248-268.
Dretske, F. (1981) *Knowledge and the Flow of Information*. Cambridge, MA: MIT.
Dretske, F. (1986) "Misrepresentation," *Belief: Form, Content and Function*. R. Bogdan (ed.). New York: Oxford U, 1991. 17-36.
Dretske, F. (1988) *Explaining Behavior*. Cambridge, MA: MIT.
Dretske, F. (1990) "Reply to Reviewers," *Philosophy and Phenomenological Research*. 50.4: 819-839.
Dretske, F. (1991) "Dretske's Replies," *Dretske and his Critics*. B. McLaughlin (ed.). Cambridge, MA: Blackwell. 180-221.
Dretske, F. (2000) "Norms, History, and the Constitution of the Mental", *Perception, Knowledge and Belief*. New York: Cambridge U, 2000. 242-258.
Eliasmith, C. (2002) "The Myth of the Turing Machine," *Journal of Experimental & Theoretical Artificial Intelligence*. 14:1-8.
Fodor, J. (1974) "Special Sciences," *Synthese*. 28: 97-115.
Fodor, J. (1975) *The Language of Thought*. New York: Crowell.
Fodor, J. (1980) "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology," *Behavioral and Brain Sciences*. 3.1: 63-110.
Fodor, J. (1987) *Psychosemantics*. Cambridge, MA: MIT.
Fodor, J. (1989) "Making Mind Matter More," *Philosophical Topics*. 17.1: 59-79. Republished in *A Theory of Content and Other Essays*. Cambridge, MA: MIT, 1990. 137-159.
Fodor, J. (1990) "A Theory of Content II: The Theory," *A Theory of Content and Other Essays*. Cambridge, MA: MIT, 1990. 89-136.
Fodor, J. (1998) *Concepts*. New York: Oxford U.
Fodor, J. (2008) "Against Darwinism," *Mind and Language*. 23.1: 1-24.
Fodor, J. and Block, N. (1971) "Cognitivism and the Analog/Digital Distinction." Unpublished manuscript.
Fodor, J. and Lepore, E. (1992) *Holism: A Shopper's Guide*. Cambridge, MA: Blackwell.
Fodor, J. and Pylyshyn, Z. (1988) "Connectionism and Cognitive Architecture: a Critical Analysis," *Cognition*. 28: 3-71.
Gibbard, A. (2003) "Thoughts and Norms," *Philosophical Issues*. 13: 83-98.
Giere, R. (1995) "The Skeptical Perspective: Science without the Laws of Nature," *Laws of Nature: Essays on the Philosophical, Scientific, and Historical Dimensions*. F. Weinert (ed.). Berlin: Walter de Gruyter & Co, 1995. 120-138.
Giere, R. (1999) "Using Models to Represent Reality," *Model-based Reasoning in Scientific Discovery*. L. Magnani, N.J. Nersessian, and P. Thagard (eds.). New

York: Plenum, 1999. 41-57.
Giere, R. (2004) "How Models are Used to Represent Reality," *Philosophy of Science*. 71: 742-752.
Gillett, C. (2003) "The Metaphysics of Realization, Multiple Realizability, and the Special Sciences," *The Journal of Philosophy*. 100.11: 591-603.
Glennan, S. (2005) "The Modeler in the Crib," *Philosophical Explorations*. 8.3: 217-227.
Glüer, K. and Wikforss, Å. (2009) "Against Content Normativity," *Mind*. 118.469: 31 -70.
Godfrey-Smith, P. (1994) "A Modern History Theory of Functions," *Function, Selection, and Design*. D. Buller (ed.). Albany: SUNY, 1999. 199-220.
Gopnik, A. (1996) "The Scientist as Child," *Philosophy of Science*. 63.4: 485–514.
Gopnik, A. and Meltzoff, A. (1997) *Words, thoughts, and theories*. Cambridge, MA: MIT.
Grice, H.P. (1957) "Meaning," *Philosophical Review*. 66.3: 377-388.
Hattiangadi, A. (2006) "Is Meaning Normative?" *Mind & Language*. 21.2: 220–240.
Haugeland, J. (1991) "Representational Genera," *Philosophy and Connectionist Theory*. W. Ramsey, D. Rumelhart, S. Stich (eds.). Mahwah, NJ: Erlbaum, 1991. 61-89. Republished in Haugeland, J. (1998) *Having Thought: Essays in the Metaphysics of Mind*. Cambridge, MA: Harvard. 171-206.
Hempel, C.G. (1959) "The logic of functional analysis," *Symposium on Sociological Theory*. L. Gross (ed.). New York: Harper and Row, 1959. 271-307. Republished in Hempel, C.G. (1965) *Aspects of Scientific Explanation*. New York: Free Press. 297-330.
Horgan, T. (1991) "Actions, Reasons, and the Explanatory Role of Content," *Dretske and his Critics*. B. McLaughlin (ed.). Cambridge, MA: Blackwell, 1991. 73-101.
Hubbell, S. (2001) *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton, NJ: Princeton.
Hutchinson, G.E. (1957) "Concluding Remarks," *Cold Spring Harbor Symposium on Quantitative Biology*. 22: 415-27.
Johnson-Laird, P. (1980) "Mental Models in Cognitive Science," *Cognitive Science*. 4: 71-115.
Kim, J. (1989) "The Myth of Nonreductive Materialism," *American Philosophical Association Proceedings*. 63: 31-47.
Kim, J. (1991) "Dretske on How Reasons Explain Behavior," B. McLaughlin (ed.) *Dretske and his Critics*. Cambridge, MA: Blackwell, 1991. 52-72.
Kripke, S. (1982) *Wittgenstein on Rules and Private Language*. Cambridge, MA: Harvard.
Kuhn, T. (1962/1970) *The Structure of Scientific Revolutions*. Second ed. Chicago: U. of Chicago.
Lewens, T. (2004) *Organisms and Artifacts*. Cambridge, MA: MIT.
Lewis, D. (1971) "Analog and Digital," *Nous*. 321-327.
Livingston, K. (1993) "What Fodor means: Some thoughts on reading Jerry Fodor's *A Theory of Content and Other Essays*," *Philosophical Psychology*. 6.3: 289-301.
Loewer, B. and Rey, G. (1991) "Editors' Introduction," *Meaning in Mind: Fodor and His Critics*. B. Loewer and G. Rey (eds.). Cambridge, MA: Blackwell, 1991. xi -xxxvii.

McLaughlin, B. (1993) "On Davidson's Response to the Charge of Epiphenomenalism," *Mental Causation*. J. Heil and A. Mele (eds.). New York: Oxford U., 1993. 27-40.

Millikan, R.G. (1984) Language, Thought, and Other Biological Categories. Cambridge, MA: MIT.

Millikan, R.G. (1989) "An Ambiguity in the Notion 'Function,' " *Biology and Philosophy*. 4: 172-176.

Millikan, R.G. (2000) *On Clear and Confused Ideas: An Essay about Substance Concepts*. New York: Cambridge U.

Millikan, R.G. (2002) "Biofunctions: Two Paradigms," *Functions: New Essays in the Philosophy of Psychology and Biology*. A. Ariew (ed.). New York: Oxford U. 113 -143. Online: www.philosophy.uconn.edu/department/millikan/biofunct.pdf

Millikan, R.G. (2004) *Varieties of Meaning*. Cambridge, MA: MIT.

Millikan, R.G. (2005) *Language: A Biological Model*. New York: Oxford U.

Nagel, E. (1961) *The Structure of Science*. New York: Harcourt, Brace & World.

Nagel, T. (1974) "What is it like to be a bat?" *Philosophical Review*. 83: 435-450.

Neander, K. (1991) "The Teleological Notion of 'Function,' " *Australasian Journal of Philosophy*. 69.4: 454-468.

O'Dea, J. (2000) "Why do Honeybees Dance?" *Natural Science*. Article 13. Victoria, BC: Heron Publishing. Online: http://naturalscience.com/ns/articles/01-13/ns_jdo.html

Pietroski, P. and Rey, G. (1995) "When Other Things Aren't Equal: Saving *Ceteris Paribus* Laws," *British Journal for the Philosophy of Science*. 46: 81–110.

Polger, T. (2007) "Realization and the Metaphysics of Mind," *Australasian Journal of Philosophy*. 85.2: 233-259.

Putnam, H. (1975) "The Meaning of 'Meaning,' " *Mind, Language and Reality: Philosophical Papers Vol. 2*. New York: Cambridge U. 215-271.

Pylyshyn, Z. (1984) *Computation and Cognition*. Cambridge, MA: MIT.

Recanati, F. (2007) *Perspectival Thought: A Plea for (Moderate) Relativism*. New York: Oxford U.

Rey, G. (2007) "Resisting Normativism in Psychology," *Contemporary Debates in Philosophy of Mind*. B. McLaughlin and J. Cohen (eds.). Malden, MA: Blackwell, 2007. 69-84.

Seager, W. (1993) "Fodor's Theory of Content: Problems and Objections," *Philosophy of Science*. 60.2: 262-77.

Sellars, W. (1962) "Philosophy and the Scientific Image of Man," *Frontiers of Science and Philosophy*. R. Colodny (ed.). Pittsburgh: Pittsburgh U, 1962. 35-78. Republished in Sellars, W. (1963) *Science, Perception and Reality*. New York: Humanities Press. 1-40.

Shannon, C. (1948) "A Mathematical Theory of Communication," *The Bell Systems Technical Journal*. 27. Online: http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf.

Smolensky, P. (1988) "On the Proper Treatment of Connectionism," *Behavioral and Brain Sciences*. 11:1-74.

Stern, R. (2004) "Does 'Ought' Imply 'Can'? And Did Kant Think It Does?" *Utilitas*. 16.1: 42-61.

Stich, S. (1983) *From Folk Psychology to Cognitive Science*. Cambridge, MA: MIT.

Stich, S. and Warfield, T. (eds.) (1994) *Mental Representation: A Reader.* Cambridge, MA: Blackwell.

Suppes, P. (1960) "A Comparison of the Meaning and Uses of Models in Mathematics and the Empirical Sciences," *Synthese.* 12: 287–300.

Théry, M. and Heeb, P. (2008) "Communication, Sensory Ecology, and Signal Evolution," *Behavioral Ecology.* E. Danchin, L.A. Giraldeau, and F. Cezilly (eds.). New York: Oxford U, 2008. 577-614.

van Fraassen. B. (1980) *The Scientific Image.* New York: Oxford U.

van Gelder, T. and Port, R. (1995) "It's About Time: An Overview of the Dynamical Approach to Cognition," *Mind as Motion.* T. van Gelder and R. Port (eds.). Cambridge, MA: MIT. 1-43.

Viger, C. (2001) "Locking On to the Language of Thought," *Philosophical Psychology.* 14.2: 203-215.

Wright, L. (1973) "Functions," *Philosophical Review.* 82: 139-168. Republished in *Function, Selection, and Design.* D. Buller (ed.). Albany: SUNY, 1999. 29-55.

Wright, L. (1976) *Teleological Explanation.* Berkeley: U. of California.