Chapter 3
A Theory of Content, I: The Problem

Introduction

It counts as conventional wisdom in philosophy that (*i*) the intentional/semantical predicates form a closed circle and (*ii*) intentional states are intrinsically holistic. (*i*) unpacks as: 'It may be possible to formulate sufficient conditions for the satisfaction of some of the intentional/semantic predicates in a vocabulary that includes other of the intentional/semantic predicates; but it is not possible to formulate such conditions in a vocabulary that is exclusively *non*semantic/intentional.' (*ii*) unpacks as: 'Nothing can exhibit any intentional properties unless it exhibits many intentional properties; the metaphysically necessary conditions for a thing's being in any intentional state include its being in many other intentional states.' (*i*) is supposed to rule out the possibility of framing physicalistically sufficient conditions for the truth of intentional ascriptions; (*ii*) is supposed to rule out the possibility of punctate minds.

Working severally and together, (*i*) and (*ii*) have served to ground quite a lot of philosophical skepticism about intentional explanation. For example, (*i*) appears to preclude a physicalistic ontology for psychology since if psychological states were physical then there would surely be physicalistically specifiable sufficient conditions for their instantiation.[1] But it's arguable that if the ontology of psychology is not physicalistic, then there is no such science.

By contrast, (*ii*) could be true consonant with physicalism; why, after all, shouldn't there be properties that are both physicalistic and holistic? But it's nevertheless plausible that (*ii*) would preclude an intentional psychology with scientific status. One important way that psychological laws achieve generality is *by quantifying over all the organisms that are in a specified mental state* (all the organisms that believe that *P*, or intend that *Q*, or whatever). But holism implies that very many mental states must be shared if any of them are. So the more holistic the mind is, the more similar the mental lives of two organisms (or of two time slices of the same organism) have to be in order that the same psychological laws should subsume them both. At the limit of holism, two minds share any of their intentional states only if they share all of them. And since, of course, no two minds ever do share all of their intentional states, the more (*ii*) is true the more the putative generalizations of intentional psychology fail, de facto, to generalize.[2] (It's a question of some interest whether, having once embraced a holistic view of intentional content, there is anywhere to stop short of going the limit. I'm inclined to think that anyone who takes it seriously that there is no analytic/synthetic distinction is obliged to answer this question in the negative. I shan't, however, argue the point here.)

The moral, in short, is that the price of an Intentional Realism that's worth having—at least for scientific purposes—is a physicalist and atomistic account of intentional states. And, as I say, it's the conventional wisdom in philosophy that no such account can be given.

There is, however, an increasingly vociferous minority in dissent from this consensus. In particular, recent developments in "informational" semantics suggest the possibility of a naturalistic and atomistic theory of the relation that holds between a predicate and the property that it expresses. Such as theory would, of course, amount to a good deal less than a complete understanding of intentionality. But it would serve to draw the skeptic's fangs since his line is that irreducibility and holism are *intrinsic* to intentionality and semantic evaluability. Given any suitably atomistic, suitably naturalistic break in the intentional circle, it would be reasonable to claim that the main *philosophical* problem about intentionality had been solved. What remained to do would then be a job of more or less empirical theory construction or a more or less familiar kind.

What follows is in part a review paper; things have recently been moving so fast in work on "naturalized semantics"[3] that it seemed to me that an overview might be useful. Here is how I propose to proceed. In chapter 3, I'll give a sketch of how approaches to the naturalization problem have evolved over the last couple of decades. (Since what I primarily want to do is make clear the current appreciation of the structure of the naturalization problem, my treatment will be dialectical and polemical, and I'll settle for my usual C− in historical accuracy.) In chapter 4, I'll offer what seems to me to be a promising version of an information-based semantic theory: this will have the form of a physicalist, atomistic, and putatively sufficient condition for a predicate to express a property. I will then go through all the proposed counterexamples and counterarguments to this condition that my friends and relations and I have thus far succeeded in dreaming up. I will try to convince you (and me, and Greycat) that none of these counterexamples and counterarguments works. Or, anyhow, that none of them *certainly* works.

Even, however, if I am right that none of them works, someone will surely find one that works tomorrow. So, the proposed moral of the paper isn't really that there is no longer a philosophical problem about intentionality. Rather, the moral I'm inclined to draw—and that I hope I can convince you to take seriously—is that a number of the problems that once made the construction of a naturalistic semantics seem absolutely hopeless now appear rather less utterly intractable than they used to. It might therefore be wise, when one goes about one's business in the philosophy of language and the philosophy of mind, to become cautious about taking intentional irrealism for granted; more cautious, at a minimum, than has been the philosophical fashion for the last forty years or so.

1. *The Background*

Skinner

Our story starts with, of all things, Chomsky's (1959) review of Skinner's *Verbal Behavior.*[4] Skinner, you'll remember, had a theory about meaning. A slightly cleaned-up version of Skinner's theory might go like this:

The English word "dog" expresses the property of *being a dog* (and hence applies to all, and only, dogs). This semantical fact about English reduces to a certain fact about the behavioral dispositions of English speaker; viz., that their verbal response "dog" is 'under the control of' a certain type of discriminative stimuli; viz., that it's under the control of dogs. Roughly, a response is under the control of a certain type of discriminative stimulus if it is counterfactual supporting that the probability of an emission of the response increases 'in the presence of' a stimulus of that type.

There is also a Skinnerian story about how English speakers come to have these sorts of behavioral dispositions. Roughly, an operant response (including an operant linguistic response) comes under the control of a type of discriminative stimulus as a function of the frequency with which the response elicits reinforcement when produced in the presence of stimuli of that type. So tokens of "dog" express the property *dog* because speakers have been reinforced for uttering "dog" when there are dogs around.

Notice that—prima facie—this theory is naturalistic by the present criteria: The condition in virtue of the satisfaction of which "dog" means *dog* is specified in the prima facie *non* intentional/semantic vocabulary of response frequency and stimulus control; and the theory is atomistic since there is, in general, no internal connection between having any one response disposition and having any other. It is, for example, conceptually possible that there should be a speaker whose response "dog" is under the control of dogs but who has no verbal response (including, a fortiori, "cat") that is under the control of cats. Indeed, Skinner's semantics allows the possibility of a speaker who has *no* discriminated verbal operants other than the disposition to respond "dog" to dogs. That could be, as Wittgenstein (1953) says in a related context, "the *whole* language . . .; even the whole language of a tribe."

As everybody knows, Chomsky rolled all over this theory; no term was left unstoned. Nor, I think, could anyone reasonable deny that his having done so was a Very Good Thing. Behaviorism had become an incubus; Chomsky's critique effected a liberation of theoretical imagination in psychology and was a critical episode on the way to developing a serious cognitive science. But for all that—as people like MacCorquadale (1970) correctly pointed out—the theory of language we were left with when Chomsky got finished with Skinner was embarrassingly lacking in answers to questions about meaning. It still is, and something needs to be done about it.

Now that the dust has settled, it's worth trying to get clear on exactly what Chomsky showed that Skinner was wrong about. I want to suggest that there is an only somewhat quixotic sense in which Chomsky's criticism, though devastatingly effective against Skinner's behaviorism and against his attempt to apply learning theory to explain language acquisition, nevertheless left the *semantical* proposal per se pretty much untouched. It is, I think, the implicit recognition of this that grounds the recent interest in informational semantics.

For example, one of Chomsky's best lines of attack is directed against the idea—required by Skinner's learning theory—that the characteristic effect of linguistic apprenticeship is to alter the strength of an operant response. (Before you learn English, the probability of your uttering "dog" when there is a dog around is presumed to be very small; after you learn English it is presumed to be appreciably bigger). Chomsky argues, to begin with, that the technical sense of response strength, according to which it is measured by, for example, frequency, intensity, and resistance to extinction, doesn't have any serious application to the use of language. One does not, qua English speaker in the presence of a dog, utter "dog" repeatedly, tirelessly, and in a loud voice. Unless, perhaps, one is bonkers.

More important, Chomsky points out, in the usual case utterances aren't *responses* at all; they're *actions*. This is to say, at a minimum, that the character of one's verbal behavior is sensitive to the content of one's beliefs and utilities. Verbal behavior is 'cognitively penetrable', as one says these days: whether one utters "dog" in the presence of a dog depends on things like whether one thinks one's auditors would be interested to hear that there's a dog about, and whether one is desirous of telling them what one thinks they would be interested to hear, and so forth. To say nothing of its depending on whether one happens to notice the dog. To put the same point slightly differently: as Skinner uses it (at least when he's outside the laboratory) "response" is really a crypto-intentional term. So the idea that Skinner has achieved the naturalization of a semantical concept by the (putative) reduction of linguistic meaning to verbal responding turns out to be a sham.

And finally, Chomsky remarks, it appears just not to be true that language learning depends on the application of carefully scheduled socially mediated reinforcement. Language seems to be learned without being taught, and Skinner's story doesn't explain how this could be so.

This is, I think, all perfectly correct and brilliantly observed. But just how much damage does it do, and just which doctrines does it do the damage to? Notice, in the first place, that in principle Skinner's semantics can perfectly well dispense with his learning theory. Skinner could—though, of course, he wouldn't want to—tell the story that goes '"dog" expresses the property *dog* because tokenings of the former are under the control of instantiations of the latter' without saying *anything* about how discriminated responses *come to be* under the control of discriminative stimuli. He could therefore simply jettison the stuff about language learning reducing to social reinforcements mediating alterations in the strength of verbal operants; which would be a very good thing for him to do since it's hopeless.

The objection that notions like 'response' are crypto-intentional when applied to the use of language is fatal to Skinner's behaviorism but, once again, not to his semantics. For, although *talking* is a form of voluntary behavior, and hence a kind of acting, *thinking* presumably isn't. Someone who is an Intentional Realist but not a behaviorist could thus embrace a Skinnerian semantics *for thoughts* while entirely rejecting Skinner's account of language. Here's how the revised story might go: There is a mental state—of entertaining the concept DOG, say—of which the intentional object is the property *dog*. (Or, as I shall sometimes say for brevity, there is a mental state that *expresses* the property *dog*). The fact that this state expresses this property reduces to the fact that tokenings of the state are, in the relevant sense, discriminated responses to instances of the property; i.e., instancings of the state covary with (they are 'under the control of') instancings of the property, and this covariation is lawful, hence counterfactual supporting.

This account isn't behavioristic since it's unabashed about the postulation of intentional mental states. And it isn't learning-theoretic since it doesn't care about the ontogeny of the covariance in terms of which the semantic relation between dog-thoughts and dogs is explicated. But it is atomistic since it is presumably conceptually possible for dog-thoughts to covary with dog instances even in a mind none of whose other states are intentional; the conditions for meaning can thus be satisfied by symbols that don't belong to symbol *systems*.

It's also atomistic in a further sense; one that I want to emphasize for later reference. The basic idea of Skinnerian semantics is that *all* that matters for meaning is "functional" relations (relations of nomic covariance) between symbols and their denotations. In particular, it doesn't matter *how that covariation is mediated*; it doesn't matter what mechanisms (neurological, intentional, spiritual, psychological, or whatever) sustain the covariation. This makes Skinnerian semantics atomistic in a way that Quineian semantics, for instance, isn't. It's a typically Quineian move to argue that since the semantical relations between, as it might be, 'proton's and protons is *theory mediated* (since, in particular, theoretical inferences mediate our applications of 'proton' to protons), it must be that *what one means by 'proton' is partly determined by the theories about protons that one endorses*. And since, for Quine, the observation vocabulary/theory vocabulary distinction isn't principled, it comes out that what one means by *any* 'X' is partly determined by what one believes about Xs.[5]

But Quine is not a good Skinnerian in holding this. A good Skinnerian says that what 'proton' means is determined *just by its functional relation to* (its causal covariance with) *protons*; given that this covariation holds, the theoretical inferences by which it's mediated are *semantically irrelevant*. In particular, two individuals whose 'proton' tokens exhibit the *same* functional relation to protons ipso facto mean the same thing by 'proton', *whatever theories of protons they may happen to hold*. The conditions for meaning constrain the functional relation between a symbol and its referent, but they *quantify over* the mechanisms (theoretical commitments, as it might be) that sustain these functional relations.[6] For Skinner, then, though not for Quine, content is radically detached from ideology. Quine's affection for

Skinner is merely sentimental after all; given his semantic holism, Quine *can't* be a Skinnerian.

Well, finally, this updated Skinnerian semantics is physicalistic on the assumption that token states of entertaining a concept can be picked out by reference to their nonsemantical properties (e.g., by reference to their neurological, or functional, or 'syntactic' properties). Which perhaps they can; who knows?[7] The point is that this highly reconstructed Skinnerianism—from which, to be sure, practically everything that Skinner cares about has been removed—would satisfy the naturalism requirement; and, as far as I can tell, it is not touched by the arguments that Chomsky mounted against *Verbal Behavior*.

In fact, if you take the behaviorism and the learning theory away from the theory of meaning in *Verbal Behavior*, what you're left with is a doctrine that looks quite a lot like the informational semantics of Dretske's *Knowledge And The Flow of Information*. Which brings us to the next stage of our story.

Dretske

F1 gives what I take to be the basic idea of Dretske's theory.

> F1. *S-events (e.g., tokenings of symbols) express the property P if the generalization 'Ps cause Ss' is counterfactual supporting.*

For example, tokenings of "dog" express the property *dog* because the generalization, 'Dogs cause "dog"-tokens' is counterfactual supporting.

I like this way of putting Dretske's proposal because it makes clear the continuity of his program with Skinner's. In Dretske's own formulation, however, the fundamental semantic relation is 'carrying information' (rather than 'expressing a property'). A first-blush account of carrying information is given by F2.

> F2. *S-events carry information about P-events if 'Ps cause Ss' is a law.*[8]

However, F2 would also not be acceptable to Dretske. For example, according to his theory, Ss carry information about Ps only if the probability that an arbitrary *S* is *P*-caused is always one; in effect, Dretske requires that '*Ps* and only *Ps* cause *Ss*' be a law.

His main argument for this very strong condition is this:[9] suppose we allow that *Ss* carry information about *Ps* even when the probability that *Ss* are *P*-caused is some *p* less than one. Then we could get a situation where *Ss* carry information about *Ps*, *Rs* carry information about *Qs*, but *S&Rs* don't carry information about *P&Qs* (viz., because the probability that *P&Q* given *S&R* is less than *p*).

But I think this argument is ill advised. There is no reason why a semantical theory should assign informational content *independently* to each expression in a symbol system. It will do if contents are assigned only to the *atomic* expressions, the semantics for molecular symbols being built up recursively by the sorts of techniques that are familiar from the construction of truth definitions. In what follows, I will in fact assume that the problem of naturalizing representation reduces to the problem of naturalizing it for atomic symbols (mutatis mutandis, atomic *mental states* if it is mental representation that is being naturalized).[10]

F1 and F2 are more closely related than may appear since we can assume that '*Ps* cause *Ss*' is counterfactual supporting only if it's a law. The connection between information and nomologicity that is explicit in F2 is therefore implicit in F1. Because the notions of law and counterfactual support are so close to the heart of both Skinner's and Dretske's views of semantics, the theories share a feature that will be important to us much later in the discussion: both imply that what your words (thoughts) mean is dependent entirely on your *dispositions* to token them, the *actual history* of their tokenings being semantically irrelevant.

This principle—that actual histories are semantically irrelevant—follows from the basic idea of informational semantics, which is that the content of a symbol is determined solely by its nomic relations. To put it roughly but intuitively, what laws subsume a thing is a matter of its *subjunctive* career; of *what it would do* (or would have done) *if* the circumstances were (or had been) thus and so. By contrast, a thing's actual history depends not just on the laws it falls under, but also on the circumstances that it happens to encounter. Whether Skinner and Dretske are right to suppose that a naturalized semantics can ignore actual histories in favor of purely subjunctive contingencies is a question we'll return to late in chapter 4. Till then, we will cleave rigorously to the principle that only nomic connections and the subjunctives they license count for meaning.

For the present, then, I propose to take F2 as my stalking horse. It formulates a doctrine that is within hailing distance of both Skinner's version of naturalized semantics and Dretske's, and it makes clear the intimate connection between the information that's generated by a causal transaction and the existence of a causal law that "covers" the transaction.[11] And as far as I can tell, the problems we're about to raise for F2 will have to be faced by any version of information-based semantics that can claim to be remotely plausible.

2. *Error and the Disjunction Problem*

You have to get error in somewhere, and so far we've made no room for it. In fact, there looks to be a dilemma about this. Suppose, to put it crudely, that "dog" means *dog* (and thus has dogs and only dogs in its extension) because it's a law that dogs cause "dogs." Then there are two possibilities:

First Possibility
Only dogs cause "dog"s. If this is so, then only things in the extension of "dog" cause it to be tokened; so it looks as though all the tokens of "dog" must be true.

Second Possibility
Some non-dogs cause "dog"s. Suppose, for example, that either being a dog or being (the right sort of) cat-on-a-dark-night is sufficient to cause a "dog" token. F2 says, in effect, that symbols express the properties whose instantiations are nomically sufficient for their tokening. So "dog" expresses the property of being *either a dog or a cat-on-a-dark-night*. So the extension of "dog" is the union of the dogs and the cats-on-dark-nights. So tokens of "dog" that are caused by cats on dark nights are *true,* and we still don't have a story about falsehood and error.

If F2 is the best that a causal theory of content can do, it looks as though such theories can't distinguish between a true token of a symbol that means something that's disjunctive and a false token of a symbol that means something that's not. The literature on informational semantics has come to call this the "disjunction problem."

What, exactly, is going on here? Well, it seems plausible that the least you'd want of a false token of a symbol is that it be caused by something that is not in the symbol's extension. But this is a condition that F2 has trouble meeting. Because:

> (*i*) it's a truism that *every* token of a symbol (including the false ones) is caused by something that has some property that is sufficient to cause a tokening of the symbol

and

> (*ii*) according to F2, any property whose instantiation is sufficient to cause the tokening of a symbol is thereby expressed by that symbol.

Since the extension of a symbol is just the set of things that have the property that the symbol expresses, it appears to follow from (*i*) and (*ii*) that *every* token of a symbol is caused by something that belongs to its extension; hence that no token of a symbol can be false. This is, to put the case mildly, not satisfactory.

Indeed, it is *so* not satisfactory that the question whether a naturalistic semantics is possible has recently come to be viewed as identical in practice to the question whether the disjunction problem can be solved within a naturalistic framework. Accordingly, most of the rest of this paper will be about the vicissitudes of recent attempts to find such a solution.

With an exception that I will retail later, all the standard attempts to solve the disjunction problem exhibit a certain family resemblance. The basic idea is to distinguish between two types of situations, such that lawful covariation determines meaning in one type of situation but not in the other. The revised theory says, in effect, that a symbol expresses a property if instantiations of the property are nomically sufficient for instantiations of the symbol *in situations of type one*. Since the tokens of a symbol that occur in type one situations are ipso facto caused by things that are in its extension, it follows that all such tokens are true. However, properties whose instantiations cause tokens of a symbol (only) *in situations of the second type* are *not* thereby expressed by the symbol; so tokens of a symbol that occur in type two situations are *not* ipso facto caused by things in its extension; so it is left open that such tokens may be false.

The strategy of the revised theory is thus to solve the disjunction problem by localizing it. It's accepted that symbol tokens in type one situations are ipso facto true;[12] and it's thereby conceded that if tokenings of a symbol are caused by more than one sort of thing *in type one situations* then it follows that the meaning of the symbol is disjunctive. But, according to the new story, not *all* sorts of situations enjoy this privilege of conveying infallibility; for example, type two situations don't. So the new story does make room for the possibility of error, which, as we've seen, the old story failed to do.

Here's a slightly different, though convergent, way to think about this distinction between type one and type two situations. It might reasonably occur to a philosopher to wonder, "Why is it that our canonical specifications of thoughts, beliefs and the like operate by employing phrases—embedded 'that' clauses—that (apparently) express actual or possible states of affairs? Why, for example, do we pick out the thought that it's raining by using the expression 'it's raining'? What is it about thoughts, and about states of affairs, that makes this practice possible?" (Papineau, 1988, wonders this sort of thing, circa p. 88, as does Loar, 1981). This is closely related to a revealing question that I believe was first raised by Donald Davison:

how are we to understand the fact that the expressions that can appear as freestanding declarative sentences can also appear as the complements of verbs of propositional attitude?

All informational accounts tell essentially the same story about this; what's going on, they say, is a species of *etiological* identification. When we use "it's raining" to specify the intentional object of the thought that it's raining, we are picking the thought out by reference to the state of affairs that would, in certain circumstances, cause it to be entertained. It's rather like an alcoholic stupor; you specify the state by reference to the sort of thing that brings it on.

All right so far; but since, in general, the tokening of an intentional state can have any of a variety of different kinds of causes (unlike, by the way, tokenings of alcoholic stupors) the problem arises, under *which* circumstances the cause of a thought is ipso facto identical to its intentional object. Answer: By definition, this coincidence obtains in situations of type one. The moral is that the disjunction problem is a, but not the only, consideration that might motivate an informational semanticist to try to draw a type one/type two distinction. Other philosophical interests point to the same desideratum.

So everything is fine; all we need is a convincing—and, of course, naturalistic—explication of the type one/type two distinction and we will understand, within the framework of an informational account of content, both how error is possible and how it is possible to individuate intentional states in the ways that we do. As it turns out, however, convincing naturalistic explications of this distinction have proved to be a little thin on the ground.

3. *Dretske's Story about Error*

The first attempt was owing to Dretske (1981). In a nutshell, Dretske's idea was to identify the type one (i.e., meaning-bestowing) situations with the ones *in which a symbol is learned*:

> In the learning situation special care is taken to see that incoming signals have an intensity, a strength, sufficient unto delivering the required piece of information *to* the learning subject. . . . Such precautions are taken in the learning situation . . . in order to ensure that an internal structure is developed with the information that s is F. . . . But once we have meaning, once the subject has articulated a structure that is selectively sensitive to information about the F-ness of things, instances of this structure, tokens of this type, can be triggered by signals that *lack* the appropriate piece of information. . . . We (thus) have a case of

> misrepresentation—a token of a structure with a false content. We have, in a word, meaning without truth. (emphasis Dretske's).

See chapter 2 for an extended discussion of this proposal; the heart of the matter is as follows.

F2 implies that *S* expresses the property that, as a consequence of the training, came to be nomically sufficient for causing *S*-tokens. It therefore matters a lot which property this is, and the crucial point is that its identity is *not* determined by the actual *S*-tokenings that the trainee produces during the learning period. For example, even a learner all of whose "dog" tokens are caused by dogs throughout the course of his training may nevertheless be using "dog" to mean not *dog* but *dog or cat-on-a-dark-night*. Whether he is doing so won't show in his overt behavior (in his tokenings of "dog") unless he happened to run into a cat-on-a-dark-night; which, by assumption, he didn't. But remember, in informational semantics, it's the subjunctives, *counterfactuals included,* that count. That is, it's the actual *and counterfactual* *S*-tokenings in training situations that fix the identity of the property that *S* expresses. Since it goes without saying that there must always be indefinitely many properties whose instantiations are *not* encountered in any finite linguistic apprenticeship, there are always indefinitely many disjunctive properties that the trainee's use of "dog" could express, *consonant with all of his actual tokenings of "dog" being dog-occasioned.* This creates a dilemma for Dretske's proposal that is itself just a version of the disjunction problem.

Case one. If a cat-on-a-dark-night had been encountered during the learning period, it would have caused a "dog" token. But then the consequence of training has been that "dog" means *dog or cat-on-a-dark-night,* and tokens of "dog" caused by cats on dark nights outside the training situation are true. So there is still no room for false tokens.

Case two. If a cat-on-a-dark-night had been encountered during the learning period, it would *not* have caused a "dog" token. Then, the consequence of the training has been that cats-on-dark-nights don't cause "dog" tokens after all; presumably, only dogs do. (If a cat-on-a-dark-night encountered *during* the training period wouldn't have caused a "dog" token, why on Earth should a cat-on-a-dark-night encountered *after* the training period cause one?) But if only dogs cause "dog" tokens, all such tokens are true and again there's no room for errors.

The moral seems to be that when you take the counterfactuals into the reckoning, the story about the training doesn't help with the disjunction problem.

I once heard Dretske make what I took to be the following suggestion: What determines the identity of the concept the student has learned is *not* the actual and counterfactual distribution of his tokenings (as per the preceding), but rather the distribution of actual and counterfactual *punishments and rewards* that prevails in the training situation. So, for example, imagine a student who has been reinforced for positive responses to *apples*, and suppose that no *wax apples* have been encountered. Then what determines that the student has learned the concept APPLE rather than the disjunctive concept APPLE OR WAX APPLE is that, *were he* to respond positive to a wax apple, the teacher (or some other environmental mechanism) *would contrive to punish the response.*

But I don't think Dretske really wants to hold this (and it's entirely possible that I have misconstrued him in thinking that he thinks that he does). For, on this account, *it would be impossible to mistakenly learn a disjunctive concept when a nondisjunctive one is being taught.* Suppose you are trying to teach me APPLE; i.e., suppose that you would punish me for positive responses to wax apples. And suppose that it somehow nevertheless gets into my head that the concept you are trying to teach me is the disjunctive APPLE OR WAX APPLE. On the current view, however explicitly I think that that *is* the concept that you are trying to teach me, and however much it is the case that I *would* respond positive to instances of WAX APPLE were any such to be presented, still the concept that I have in fact acquired is not APPLE OR WAX APPLE but APPLE. Because: the proposal is that it's the objective distribution of (actual and counterfactual) punishments and rewards in the training situation that determines the identity of the concept that I learn; and, by hypothesis, in this training situation it's APPLEs and not APPLE OR WAX APPLEs, to which the actual and counterfactual rewards accrue. This, surely, is a reductio of the proposal. If the objective reinforcement contingencies determine which concepts we acquire we'd all be practically infallible and induction would be a snap. Alas, what constitutes my concepts is not *the objective reinforcement contingencies,* but rather *the reinforcement contingencies that I take to obtain.* Cf. a point that Chomsky made against Skinner: what's reinforced is one thing, what's learned is often quite another.

None of this shows, of course, that you can't get out of the disjunction problem by restricting the circumstances under which causation makes content. But it does suggest that the identification of type one situations with *learning* situations won't do the trick.

4. *Teleological/Functional Solutions*

The basic idea for dealing with the disjunction problem was to define a *type one situation* such that:

> (*i*) If it's a law that *P*s cause *S*-tokens in type one situations, then *S* means *P* (and if *P* is disjunctive, then so be it);

and

> (*ii*) not all situations in which *S* gets tokened qualify as type one, so that tokens of *S* that happen in *other* sorts of situations are ipso facto free to be false.

Well, it looks as though type one situations can't be learning situations; but here's an alternative proposal. *Normal* situations are just the sort of situations we require. We are now about to spend some time looking at this proposal.

Prima facie, this kind of idea is sort of attractive; it's sensitive to the plausible intuition that errors are cases where *something has gone wrong*: "Where beliefs are false . . . we also expect some explanation for the deviation from the norm: either an abnormality in the environment, as in optical illusions or other kinds of misleading evidence, or an abnormality in the internal belief-forming mechanisms, as in wishful thinking or misremembering" (Stalnaker, 1984, p. 19). Conversely, normal situations are maybe just the one's where *everything has gone right*. In which case—since it's plausible (perhaps it's tautological) that when everything has gone right what you believe is true—it's maybe OK if S-tokens are all true in normal situations.

So maybe it's OK if, in normal situations, the conditions for meaning and truth come out to be the same. *Normal*—at least when it's used this way—is a normative notion,[13] and *true* is a normative notion, so maybe it's not surprising if the former notion reconstructs the latter. So, at least, one might be inclined to argue at first blush.

Of course, if the intentional circle is to be broken by appeal to *Normal* situations for symbol tokenings, we had better have some naturalistic story to tell about what it is for a situation to be *Normal* in the relevant respect. What might such a story look like? Roughly, the suggestion is that *Normality* should somehow be cashed by appeal to (natural) teleology; e.g., to some more-or-less Darwinian/historical notion of biological mechanisms *doing what they were selected for*.

So, then, here's a sketch of the story: an organism's mental-state tokens get caused by, for example, events that transpire in the organism's local environment. There are, of course, mechanisms—typically neuronal ones—that mediate these causal transactions. And these mechanisms have presumably got an evolutionary history. They are presumably the products of processes of selection, and it's not implausible that what they were selected *for* is precisely their role in mediating the tokening of mental states. So there are these cognitive mechanisms, and there are these cognitive states; and the function of the former is to produce instances of the latter upon environmentally appropriate occasions.

Strictly speaking, it doesn't, of course, follow, that *the cognitive states themselves*—states like believing that *P* or desiring that *Q* or doubting that the Dodgers will ever move back to Brooklyn—have a Normal function; in fact, it doesn't follow that they have any function at all. (You could perfectly well have a machine whose function is to produce things that are themselves functionless. In a consumer society you might have quite a lot of these.) Since the assumption that there is a teleological story to be told about the mechanisms of belief *fixation* does not imply that there is a teleological story to be told about *beliefs*, it a fortiori does not imply that beliefs (or, mutatis mutandis, other intentional states) can be *individuated by reference to their functions*. This is important because it's more intuitive that belief-fixing mechanisms (nervous systems, for example) have functions than that beliefs do; and the implausibility of the latter idea ought not to prejudice the plausibility of the former.

Nor would a teleological solution of the disjunction problem require that intentional states can be functionally individuated. All solving the disjunction problem requires is a distinction between Normal and abNormal circumstances for *having* a belief (hence between type one circumstances for having a belief and others). There would be such a distinction even if there were no such things as Normally functioning beliefs, so long as there are such things as Normally functioning mechanisms of belief fixation. Per se, teleological solutions to the disjunction problem do not therefore require that there be Darwinian (or, indeed, any) answers to questions like, "What is the belief that seven is prime for?"

There seems to be a certain amount of confusion about this point in papers like Millikan (1986). Millikan thinks that beliefs, desires and the like must have "proper functions," and she thinks this because she thinks that "there must, after all, be a finite number of general principles that govern the activities of our various cognitive-state-making and cognitive-state-using mechanisms and there must be explanations of why these principles have historically worked to aid our survival" (p. 55).

But the assumption that the mechanisms that make/use cognitive states have functions does not entail that cognitive states themselves do. And the assumption that it's useful to have cognitive states does not entail that you can distinguish among cognitive states by reference to their uses. It's a sort of distributive fallacy to argue that, if having beliefs is functional, then there must be something that is the distinguishing function of each belief. The function of the human sperm cell is to fertilize the human ovum; what, then, is the distinguishing function of *this* sperm cell? The hair on your head functions to prevent the radiation of your body heat; what, then, is the distinguishing function of *this* hair (or, for that matter, of *red* hair)?

Conversely—and contrary to Millikan—if there is nothing that the belief that seven is prime is *for* (and that the belief that four is even is not for), it wouldn't follow that "our cognitive life is an accidental epiphenomenal cloud hovering over mechanisms that evolution devised with other things in mind." Having toes is a good idea; I suppose there's even a selectional story about why we have them. It does not follow that each toe has its distinguishing function, or that this toe has any function that one hasn't. Nor, for all that, are my toes at all like epiphenomenal clouds hovering over something.

Millikan's idea is that, on the one hand, cognitive states are distinguished by their functions and, on the other, it's the function of a cognitive state that determines its intentional object. ". . . the descriptions we give of desires [and the like] are descriptions of their most obvious proper functions [so that the fact that] desire(s) are . . . individuated . . . in accordance with content is as ordinary a fact as . . . that the categories 'heart', 'kidney', and 'eye' are carved out by reference to *their* most obvious proper functions" (pp. 63–64). The idea that content reduces to Normal function is one of the two main threads in the story we're examining (the other being the idea that function reduces to selectional history, of which a lot more presently).

Now there is, right at the beginning, something fishy about the idea that the content of a mental state is to be understood by reference to its function since this sort of account leaves it mysterious why the identification of content with function works *only* for intentional states; why beliefs have intentional content in virtue of their functions but hearts, eyes, and kidneys don't. In any event, the disanalogy between the functional individuation of propositional attitudes and the functional individuation of hearts, eyes, and kidneys would seem to be glaring. Functions are, I suppose, species of Normal effects. We find out that the function of the heart is to pump the blood when we find out that, among the Normal effects of heart beat, blood circulation (and not, say, heart noise) is the effect that hearts are designed to produce. But how would the corresponding analysis go in the case of intentional states like desires? What is it that the desire to be rich and famous can Normally be relied upon to effect in the way that hearts can Normally be relied upon to effect the circulation of blood? *Trying to become rich and famous* is perhaps a candidate since, I suppose, people who want to become that do Normally try to become it. But trying is no good for the job at hand since it is itself an intentional state. *Actually becoming rich and famous* would do, except that it's so wildly implausible that it is, in any nonquestion-begging sense, a Normal effect of *wanting* to become it.

Contrary to what Millikan claims, it's just not on the cards that "the proper function of every desire . . . is to help cause its own fulfillment." (p. 63) For, on the one hand, nothing is the proper function of Xs except what Xs Normally help to cause; and, on the other, if Xs Normally help to cause Ys, then presumably *when the situation is Normal Ys can be relied upon to happen when(ever) it's the case that* X. Thus the activity of the heart helps to cause a state of affairs—viz., that the blood circulates—that can Normally be relied upon to happen when the heart beats (i.e., that can be relied upon to happen when the heart beats and the situation is Normal). But does Millikan really believe that wanting to become rich and famous helps to cause a state of affairs—viz., that one becomes rich and famous—which can Normally be relied upon to happen if one wants that it should? And, if she really does believe this, isn't that because she's sort of sneaked a look at the intentional object of the want?[14]

Millikan remarks—in one breath, as it were—that "a proper function of the desire to eat is to bring it about that one eats; [and] a proper function of the desire to win the local Democratic nomination for first selectman is to bring it about that one wins the local Democratic nomination for first selectman" (p. 63). But while there is arguably a *law* that connects desires to eat with eatings (ceteris paribus) and a law that connects functioning hearts with blood pumpings (ceteris paribus), what's the chance that there is any Normally reliable, nonintentional connection between desires to win elections and election winnings? Stevenson wanted to win just as much as Eisenhower did, and the circumstances were equally Normal for both. But Eisenhower won and Stevenson didn't. In Normal circumstances, not more than one of them could have, what with elections being zero-sum games. So how could it be that, in virtue of a law or other reliable mechanism, in Normal circumstances everybody wins whatever elections he wants to? When the situation is Normal, the lion wants to eat and the lamb wants not to be eaten. But. . . .

The proposal is that the proper function of a desire is to bring about the state of affairs that it Normally helps to cause, and that the state of affairs that a desire would bring about were it performing its proper function is its intentional object. Thus far I've been running the discussion of this proposal on the reading of 'Normally helps to cause' that examples like hearts, eyes, kidneys, and the like most obviously suggest: 'if *X Normally helps to cause Y*, then *"if X then Y" is true if the situation is Normal.*' But, as Tim Maudlin has pointed out to me, it's entirely possible that Millikan has a less robust notion of 'Normally helping to cause' in mind; perhaps it's enough for *X* Normally helping to cause *Y* that the probability of *Y* given *X* is Normally greater than the probability of *Y* given not-*X*.[15] This would cope with the kinds of counterexamples I've been offering since it wouldn't require that when the situation is Normal you actually get *Y*s whenever you get *X*s.

This revised proposal is, however, clearly too weak. For example: the recording that I want to buy is the Callas *Tosca*, but I'm prepared to "suboptimize": I'll settle for the Milanov if Milanov is all they've got. So my wanting to buy the one recording increases the probability that I'll actually buy the other; "all ships float on a rising tide," as Granny is always saying. Nor is there the slightest reason to doubt that this sort of suboptimizing has survival value; probably if we didn't do it, we'd all go mad. (Perhaps if we didn't do it we'd already *be* mad since our willingness to suboptimize is arguably a constituent of our practical rationality.) In short, *helping me to get the Milanov Tosca* satisfies the revised condition for being the proper function of my wanting the Callas *Tosca*. (As does, of course, help me get the Callas *Tosca*. One consequence of this construal of 'proper function' being too weak is that it fails to yield unique proper functions.) But it is, for all that, the Callas *Tosca* and not the Milanov *Tosca* that is the intentional object of my want.

Other sorts of cases point the same moral. Normally, my desiring to win the lottery increases at most very slightly the likelihood that I will do so. It increases considerably more the likelihood that I shall presently be five dollars poorer, five dollars being the price of a ticket. For all that, what I want is to win the lottery, not to get poorer; getting poorer comes in not as the intentional object of my want but merely as a calculated risk.

So, for one reason and another, the revised construal of 'Normally helping to cause' is too weak; but like the original construal it is also too strong, and this is the more serious fault. It is simply intrinsic to the logic of wants that they can be causally isolated from the states of affairs whose occurrence would satisfy them, even when things are perfectly Normal. So, I can want like stink that it will rain tomorrow and spoil Ivan's picnic. Not only is it not the case that my wanting this is Normally sufficient to bring it about; my wanting it doesn't alter in the slightest scintilla the likelihood that it will happen. That it is possible to have wants that are arbitrarily causally inert with respect to their own satisfaction is, indeed, one of the respects in which wants are intentional; it's what makes wanting so frightfully nonfactive. "If wishes weren't causally isolated from horses, beggars would ride ceteris paribus," as Granny is also always saying.

As we've seen, however, the teleological solution to the disjunction problem doesn't have to go Millikan's way; in particular, it doesn't require either that intentional states (as opposed to cognitive mechanisms) should have proper functions, or that the putative proper functions of intentional states should determine their contents. Let us therefore leave Millikan and return to the main line of argument.

There are—let's assume—these cognitive mechanisms whose function is to mediate the causal relations between environmental states on the one hand and mental states on the other. Of course, they don't mediate those relations in just *any old* circumstances. Organisms don't hear well when they have carrots in their ears, and they don't see well when they have dust in their eyes . . . etc. But if there is an evolutionary story about a cognitive mechanism, then presumably there must be naturalistically specifiable circumstances C such that

> (*i*) ceteris paribus, the mechanism in question mediates the relations in question whenever circumstances C obtain;

and

> (*ii*) ceteris paribus, possession of the mechanism bestows selectional advantage because it does mediate the relation whenever circumstances C obtain.

Let's suppose that all of this is so. Then we identify 'Normal' (hence, type one) situations as the ones in which it's the case that C; and we say that if mental state tokens of type *S* are caused by *P*-instantiations in such situations, then tokens of mental state *S* mean (express the property) *P*. Since situations where it isn't the case that C are ipso facto not Normal for the tokening of *S*, and since it's only in Normal circumstances that causation is supposed to be constitutive of content, *S*-tokens that transpire when it isn't the case that C are free to be caused by anything they like. In particular, they are free to be false.

So, then, Darwinian teleology underwrites the appeal to Normal functioning and the appeal to Normal functioning solves the disjunction problem and naturalizes content. In consequence, if you say to an informational semantical "Please, how does meaning work?" you are likely to get a song and dance about what happens when frogs stick their tongues out at flies. "There is," so the song goes, "a state *S* of the frog's nervous system such that:

(*i*) *S* is reliably caused by flies in Normal circumstances;
(*ii*) *S* is the Normal cause of an ecologically appropriate, fly-directed response;
(*iii*) Evolution bestowed *S* on frogs because (*i*) and (*ii*) are true of it."

S, one might say, Normally resonates to flies. And it is only because it Normally does so that Mother Nature has bestowed it on the frog. And it is only because Mother Nature bestowed it on the frog only because it Normally resonates to flies that tokens of this state *mean* fly *even in those (abNormal) circumstances in which it is not flies but something else to which the S-tokens are resonating.*[16]

So that, at last, is the full-blown causal/teleological/historical-Darwinian story about how to solve the disjunction problem and naturalize content.[17]

Now, anybody who takes the picture of evolutionary selection that this teleological story about Normal circumstances presupposes to be other than pretty credulous should look at Gould and Lewontin's splendid paper, "The Spandrels of San Marco" (1979). It is, I think, most unlikely, even on empirical grounds, that Darwin is going to pull Brentano's chestnuts out of the fire. For present purposes, however, I'm goinp to bypass the empirical issues since there are internal reasons for doubting that the evolutionary version of the teleological account of intentionality can do the work for which it has been promoted.

In the first place—contrary to advertisements that you may have seen—the teleological story about intentionality does *not* solve the disjunction problem. The reason it doesn't is that teleological notions, insofar as they are themselves naturalistic, always have a problem about indeterminacy just where intentionality has its problem about disjunction. To put it slightly more precisely, there's a kind of dilemma that arises when you appeal to the function of a psychological mechanism to settle questions about the intentional content of a psychological state. If you specify the function of the mechanism *by* *reference to* the content of the state (for example, you describe the mechanism as mediating the initiation of actions under certain maxims or the fixation of beliefs *de dicto*) then you find, unsurprisingly, that you get indeterminacy about the function of the mechanism wherever there is ambiguity about the content of the state. And if, on the other hand, you describe the function in some way that is intentionally neutral (e.g., as mediating the integration of movements or the fixation of beliefs *de re*) you may get univocal functional ascriptions but you find, still unsurprisingly, that they don't choose between competing ascriptions of content. Either way then, the appeal to teleology doesn't help you with your disjunction problem.

We can see this dilemma play itself out in the case of the frog and the flies. Here is David Israel (1987) expounding a teleological solution to the frog's disjunction problem:

> We've talked of [a certain neural state of the frog's as] . . . *meaning* that there's a fly in the vicinity. Others have said that what 'fly' means to the frog is just [a] characteristic pattern of occular irradiation—i.e., as of a small black moving dot. This is just backwards. The facts are that, in a wide range of environments, flies are what actually cause that pattern on the frog's eyes and that *flies on the fly are what the frog is after*. This convergence of the 'backward looking' (environment-caused) and 'forward looking' (behavior-causing) aspects of the state is a good thing (from the frog's parochial point of view of course) (pp. 6–7). . . . Talk of belief is essentially functional talk: the crucial function . . . of belief states is that they represent the world as being a certain way and, together with desire states, cause bodily movements. What movements? . . . . If things go well, they cause those movements which, if the world is as it is represented, will constitute the performance of an action that satisfies the agent's desires. If the world is not the way it is represented as being, the bodily movement is considerably less likely to succeed. (p. 15)

The trouble is, however, that this doesn't *solve* the disjunction problem; it just begs it. For, though you *can* describe the teleology of the frog's snap-guidance mechanism the way that Israel wants you to—in Normal circumstances, it resonates to flies; so its function is to resonate to flies; so its intentional content is *about* flies—there is precisely nothing to stop you from telling the story in quite a different way. On the alternative account, what the neural mechanism in question is designed to respond to is little ambient black things. It's little ambient black things which, "in a wide range of environments

. . . are what actually cause that pattern on the frog's eyes" and little ambient black things are "what the frog is after." Hence, a frog is responding *Normally* when, for example, it snaps at a little ambient black thing that is in fact *not* a fly but a bee-bee that happens to be passing through.

Notice that, just as there is a teleological explanation of why frogs should have fly detectors—assuming that that is the right intentional description of what they have—so too there is a teleological explanation of why frogs should have little-ambient-black-thing detectors—assuming that *that* is the right intentional description of what they have. The explanation is that *in the environment in which the mechanism Normally operates* all (or most, or anyhow enough) of the little ambient black dots are flies. So, in this environment, what ambient-black-dot detectors Normally detect (de re, as it were) is just what fly detectors Normally detect (de dicto, as it were); viz., flies.

It bears emphasis that *Darwin doesn't care which of these ways you tell the teleological story*. You can have it that the neural mechanism Normally mediates fly snaps, in which case snaps at bee-bees are ipso facto errors. Or you can have it that the mechanism Normally mediates black dot snaps that are, as one says at Stanford, "situated" in an environment in which the black dots are Normally flies. (On the latter reading, it's not the frog but the world that has gone wrong when a frog snaps at a bee-bee; what you've got is a Normal snap in an abNormal situation.) It is, in particular, true *on either description* of the intentional object of the frog's snaps that, if the situation is Normal, then "if the world is as it is represented [snapping] will constitute the performance of an action that satisfies the agent's desires."

Correspondingly, both ways of describing the intentional objects of the snaps satisfy what Millikan (1986) apparently takes to be the crucial condition on content ascription: Both make the success of the frog's feeding behavior not ". . . an accident [but] . . . the result of the elegant self-programming of his well designed nervous system. More explicitly [they both make it a] result of his nervous system's operating in accordance with general principles that also explained how his ancestors' nervous systems programmed themselves and used these programs so as to help them to proliferate" (p. 68). Huffing and puffing and piling on the teleology just doesn't help with the disjunction problem; it doesn't lead to univocal assignments of intentional content.[18]

The Moral, to repeat, is that (within certain broad limits, presently to be defined) Darwin doesn't care how you describe the intentional objects of frog snaps. All that matters for selection is how many flies the frog manages to ingest in consequence of its snapping, and this number comes out exactly the same whether one describes the function of the snap-guidance mechanisms with respect to a world that is populated by flies that are, de facto, ambient black dots, or with respect to a world that is populated by ambient black dots that are, de facto, flies.[19] "Erst kommt das Fressen, denn kommt die Morale." *Darwin cares how many flies you eat, but not what description you eat them under*. (Similarly, by the way, flies may be assumed to be indifferent to the descriptions under which frogs eat them.) So it's no use looking to Darwin to get you out of the disjunction problem.

I've been arguing that a teleologically based theory of content will have to put up with a lot of intentional indeterminacy. In defiance, probably, of prudence, I propose to push this line of argument further. Let's ask *how much* intentional indeterminacy one would have to put up with on the teleological story.

I think that the right answer is that appeals to mechanism of selection won't decide between *reliably equivalent* content ascriptions; i.e., they won't decide between any pair of equivalent content ascriptions where the equivalence is counterfactual supporting. To put this in the formal mode, the context: *was selected for representing things as F* is transparent to the substitution of predicates reliably coextensive with *F*. A fortiori, it is transparent to the substitution of predicates *necessarily* (including *nomologically* necessarily) coextensive with *F*. In consequence, evolutionary theory offers us no contexts that are as intensional as 'believes that. . . .' If this is right, then it's a conclusive reason to doubt that appeals to evolutionary teleology can reconstruct the intentionality of mental states. Let's look at the frog case again with this in mind.

It might be argued that there is a real indeterminacy about whether frogs snap at flies or at little black dots. But, surely, if there are any matters of fact about content, it's one of them that frogs don't snap at flies under the description *fly or bee-bee*. Yet, as far as I can see, it's equally OK with Darwin which way you describe the intentional objects of fly snaps, so long as it's reliable (say, nomologically necessary; anyhow, counterfactual supporting) that all the local flies-or-bee-bees are flies. The point is, of course, that if all the local flies-or-bee-bees are flies, then it is reliable that the frog that snaps at one does neither better nor worse selection-wise than the frog that snaps at the other. So evolutionary teleology *cannot tell these frogs apart*.

Here one has to be a little careful to avoid red herrings. It might be argued that you can't have a fly-or-bee-bee concept unless you have a bee-bee concept, and, since having a bee-bee concept would do the frog no good, we do, after all, have Darwinian reason to suppose that it's flies, and not flies-or-bee-bees that frogs snap at. This argument is in jeopardy of proving that *we* don't have the concept UNICORN. And, anyhow, its major premise is false. In principle, the frog could perfectly well have a *primitive* concept whose *extension* is disjunctive (from our point of view, as it were). In particular, it could perfectly well have the concept *fleebee,* whose extension embraces the flies and the bee-bees but which has neither the concept *bee-bee* nor the concept *fly* as constituents. The present question, then, is whether considerations of evolutionary (or other) utility can distinguish the hypothesis that the intentional object of the frog's snap is a fleebee from the hypothesis that it's a fly. And I claim that the line of argument I've been running strongly suggest that they cannot. Selectional advantage cares how many flies you get to eat in Normal circumstances; and, in Normal circumstances you get to eat the same number of flies whether it's flies or fleebees that you snap at.

Notice, by the way, how exactly analogous considerations show that, if "*F* iff *G*" is reliable, then just as *evolutionary theory* cannot appeal to a difference in probable utility to distinguish organisms that respond to Fness from organisms that respond to Gness, so too *reinforcement theory* cannot distinguish between such organisms by appealing to a difference in probable reward. This is what generated the traditional problem about "what is learned" over which Skinnerians used to agonize; it's precisely what one should expect given the very close similarity between Darwinian accounts of how environments select genotypes and Skinnerian accounts of how environments select behavioral phenotypes.

Suppose, in an operant conditioning paradigm, I train an organism to prefer green triangles to some negative stimulus. Is it then the greenness or the triangularity or both that the animal is responding to? I can tell only if I can "split" the greenness from the triangularity (e.g., by providing a red triangle or a green nontriangle as a stimulus) and see which way the animal generalizes. Similarly, I can teach a preference for greenness *as opposed to* a preference for triangularity only if *greens are triangles and vice versa* is not counterfactual supporting in the training situation, since that's the only circumstance in which responses to greenness and responses to triangularity can be differentially reinforced.[20] Since, however, *responding to Fness* and *responding to Gness* can be distinct intentional states even when 'F iff G' is reliable, I take this to be a sort of proof that there could not be a conditioning-theoretic solution of the disjunction problem. Contexts like "whether the stimulus is . . . determines the probability of reinforcement" slice specifications of the stimulus thicker than typical intentional contexts do; if 'F' makes this context true, so too does

'G,' so long as 'Fs are Gs' is reliable. So, the same reasoning that shows that Darwin is no use to Brentano shows that Skinner is no use to him either.[21]

Perhaps you are now yourself prepared to bite the bee-bee; perhaps you are now prepared to say that it's OK after all if there's no fact of the matter about whether the intentional objects of the frog's snaps are fleebees rather than flies. But notice that that isn't *solving* the disjunction problem; it's just deciding to live with it. Specifically, it's deciding to live with the massive intentional indeterminacy that the disjunction problem implies. But, if all you want to do is *not* solve the disjunction problem, then unvarnished, *non*teleological/*non*evolutionary versions of causal theories of content will do that quite adequately *without* appealing to the Darwin stuff. So, either way, it wouldn't appear that the Darwin stuff is buying you anything.

Let me pause a bit to rub this in. Dennett (1987) argues that Dretske and I have this disjunction problem *because* we don't take account of "utility." ". . . when we adopt the intentional stance . . . the dictated attributions are those that come out veridical *and useful* (sic). Without the latter condition . . . [one is] stuck with Fodor's and Dretske's problem of disjunctive dissipation of content . . . " (p. 311). But as far as I can see, usefulness is useless for the purposes at hand. After all, it *is* useful, in fact it's simply *super* (for a frog) to eat *flies or bee-bees* in any world in which the *flies or bee-bees* are reliably flies. It's eating flies-or-bee-bees in worlds like that that keeps frogs going.

I suppose it might be a way out of this fix to appeal to counterfactuals about what *would* happen if the locally reliable coextension between flies and flies-or-bee-bees were broken. The thought would be that snapping at flies-or-bee-bees would be bad for the frog in a world where many of the flies or bee-bees are bee-bees. But:

First, Dennett is explicit in rejecting the sort of theory that makes content rest on the causal relations that *would* hold in (merely) counterfactual circumstances (see p. 309). For Dennett (as for Millikan) it's selectional *history* that determines content.

Second (to revert to a point I made in discussing Papineau; see note 19), it's not clear how to decide which counterfactuals are the ones that count; fleebee snaps aren't advantageous in abNormal worlds where the fleebees mostly aren't flies *unless* it happens that the bee-bees in those worlds are edible.

Third (and this is the crucial point), going counterfactual to define function (and hence content) would be to give up on a Darwinian solution to the disjunction problem since utility that accrues only in *counterfactual* environments *doesn't produce actual selectional advantages*. This means that you can't reconcile appeals to counterfactual advantages with an analysis that construes content and function in terms of selection history.

That ought to be just obvious. Consider, for example, the brightly colored fish that, according to popular legend, are found in sunless ocean deeps. I don't know what the evolutionary explanation is supposed to be, but one thing is for certain: it can't be that the fish are colored because for them to be so *would be* advantageous if their environment *were* lit up. How could the selectional advantages that would accrue if you lived in an illuminated world (which, we're assuming, you don't) explain your being colored in *this* world (which, we're assuming, you are). Merely counterfactual advantages don't affect actual histories of selection. So appeals to merely counterfactual advantages can play no role in Darwinian explanations.

Well, similarly, in the present case, if it's reliable that all the flies-or-bee-bees are flies, then that's true not just of all the flies-or-bee-bees that *this* frog has encountered, but also of all the flies-or-bee-bees that its Granny encountered, and that its Granny's Granny encountered . . . and so on back to the primordial protoplasmic slime. But then, by what mechanism could selection have preferred frogs that snap at flies to frogs that snap at flies-or-bee-bees? What selection wants is that some actual frogs should actually go hungry in consequence of actually snapping at the wrong sort of things. But that won't ever happen if, in point of nomological necessity, all the frog-or-bee-bee-snaps that are prompted by bee-bees are ipso facto counterfactual.

It can't be overemphasized, in this context, that Darwinian explanations are species of *historical* explanations: they account for the geneotypical properties of organisms (or, if you prefer, for the statistical properties of gene pools) by reference to the actual—not the counterfactual—histories of predecessors. (See, for example, Millikan, 1984, p. 3: "The 'functions' of these natural devices are, roughly, the functions upon which their continued reproduction or survival has depended." Note the tense and mood.)

So far, I've followed Dennett, Millikan, et al. and assumed that it's essential to teleological semantics to be Darwinian. But, of course, one might just give up on the reduction of content to selectional history and try for a *nonhistorical* theory of content; one in which content is determined not by the selectional pressures that *actually* governed the evolution of a psychological state but by the selectional pressures that would apply if certain counterfactuals were true. E.g.: Either fly-snaps and fly-or-bee-bee snaps are equally advantageous in *this* world. But the intentional objects of frog snaps are flies and not flies-or-bee-bees because fly-snaps would be selected in nearby worlds where there are flies *whether or not there are bee-bees there* but fly-or-bee-snaps would not be selected in nearby worlds where there are bee-bees *unless there are also flies there*. In effect, there's a question about which of two locally confounded properties selection is contingent on; so one applies the method of differences across counterfactual worlds to deconfound them. Appealing to the counterfactuals licenses an intensional (with an 's') notion of selection; it distinguishes the effects that selection *really* cares about (getting flies in) from those that are merely adventitious (getting fleebees in).

But the question arises why these counterfactuals should matter for determining *content* even if, as seems quite plausible, they are exactly what matters for determining *function*. Consider the following case: I suppose that the function of the preference for sweets is to get sugars (hence calories) aboard, and I suppose that the ingestion of saccharine is nonfunctional. This works out fine on the counterfactual approach to function: A preference for sweets would be a good thing to have in a world where all the sweet things are sugar but it would lack survival value in a world where all the sweets are saccharine. But the trouble is that, in this sort of case, function and content come apart. The function of a sweet tooth is to get you to ingest sugar; but its intentional object is—not sugar but—sweets; that's why saccharine satisfies the craving. N.b., saccharine *satisfies* the craving for sweets; it doesn't just cause the craving to go away.[22]

It looks to me as though the evolutionary line on content makes two mistakes, either of which would be adequately fatal: On the one hand, it supposes that you can get a historical/selectional analysis of function (that the function of a state is what it was actually selected for) whereas what you need for function is pretty clearly some kind of counterfactual analysis (the function of a state is what it would have been selected for even if. . .). And, on the other hand, it supposes that if you're given the function of a state you are thereby given its intentional object, and the sweet tooth case strongly suggests that this isn't so.

In my view, what you've got here is a dead theory.

One last point before I stop jumping up and down on this dead theory. One way that you can really confuse yourself about the value of appeals to Darwin in grounding intentionality is to allow yourself to speak, sort-of-semi-seriously as you might say, of evolutionary teleology in terms of "what Mother Nature has in mind." The reason that this can be so confusing involves a point I called attention to above: The expressions that are deployed where we seriously and nonmetaphorically explain things by appealing to people's purposes, intentions, and the like, are far less transparent to the substitution of coextensive predicates than those that evolutionary explanations use.

As far as I can see, so long as we're dealing strictly in Darwinian (viz., historical) explanations, there's no sense to the claim that a state is selected for being F but not for being G in cases where it's necessary that F and G are coextensive.[23] In effect, Darwinian explanations treat reliably coextensive representations as synonymous; whereas, of course, psychological explanations don't. So if you're in the habit of thinking of evolutionary explanations on the model of appeals to an invisible engineer, you are likely to think that they're doing you a lot more good than they really can do when it comes to the individuation of contents.

Look, if Granny builds a mechanical frog, she may have it in mind that her frog should snap at flies, and not have it in mind that her frog should snap at things that are flies-or-bee-bees. So her mechanical frog is a fly-snapper and not a fly-or-bee-bee snapper, however reliably all the local fly-or-bee-bees are flies. (This is just like Dennett's "two-bitser," though apparently our intuitions don't agree about such cases. On my view, but not on his, if I build a machine that I intend to go into state S whenever I put a quarter in, then the machine is a quarter-accepter even if there are, in some other part of the forest, Mexican rupees which are physically very like quarters and hence *would* make the machine go into state S if it *were* to encounter any.) Attributions of (so-called) "derived intentionality," unlike specifications of "what Mother Nature has in mind" are typically opaque to the substitution of reliably coextensive expressions. In particular, they can distinguish between fly-snaps and fly-or-bee-bee snaps.

So there is no disjunction problem for derived intentionality. Where we have things whose states have derived intentionality (the intentionality of all the artifacts that Granny's made so far, by the way) we can construe very fine distinctions among the contents of their states. That's because we can construe very fine distinctions among the contents of *our* states, and derived intentionality is intentionality that's derived *from us*. Ascriptions of derived intentional objects to *Granny's frog* can distinguish between reliably coextensive contents because attributions of mental states to *Granny* can distinguish between reliably coextensive contents. There really is a difference between mechanical fly-snappers and mechanical fly-or-bee-bee snappers *because* there really is a difference between Grannies who intend their frogs to snap at the one and Grannies who intend their frogs to snap at the other.

The logic of teleological explanations that appeal to selectional advantage would appear, however, to be *very* different. As we've seen, it's quite unclear that appeals to "what Mother Nature has in mind" can rationalize distinctions between reliably equivalent content attributions. Indeed you might put Brentano's thesis like this: The difference between Mother Nature and Granny is precisely that Granny does, and Mother Nature doesn't, honor merely intentional distinctions. I don't say that Granny is *smarter* than Mother Nature; but I do say she's much more *refined*.

It is, in consequence, *very, very* misleading to say that since ". . . in the case of an organism . . . [content] . . . is not independent of the intentions and purposes of Mother Nature, [it is] just as derived as . . . the meaning in [states of an artifact]" (Dennett, p. 305).[24] The putative analogy gets it wrong about attributions of derived intentionality since it underestimates the distinctions among contents that such attributions can sustain relative to those that attributions of content to "Mother Nature" can. And—what is maybe worse—it deeply misinterprets the Darwinian program, which was precisely to *purge* biology of anything that has the logic of the kinds of explanation that are intentional with a 't.' Really (as opposed to metaphorically), Darwinian explanation isn't *anything like* ascribing goals to Mother Nature. Contrary to what Dennett says, Darwin's idea is not that " . . . we are artifacts designed by natural selection . . . " (p. 300). Darwin's idea is much deeper, much more beautiful, and appreciably scarier: We are artifacts designed by selection in exactly the sense in which the Rockies are artifacts designed by erosion; which is to say that we aren't artifacts and nothing designed us. We are, and always have been, entirely on our own.

Of course Darwin has nothing to say to Brentano; the whole point of Darwin's enterprise was to get biology out of Brentano's line of work.

And that's not all that's wrong with the evolutionary/teleological treatment of the disjunction problem. Many paragraphs back, I remarked on the naturalness of the intuition that grounds the teleological story, the intuition that error is what happens when something goes wrong. But you need more than this to license a teleological solution to the disjunction problem; you also need it that when things go right—more particularly, when things are Normal—whatever causes a symbol to be tokened is ipso facto in the extension of the symbol. It's this that ties the *teleological* story about Normalcy to the *causal* story about content. Teleology defines the class of situations in which everything is Normal; but it's the assumption that Normally caused symbols ipso facto *apply* to their causes that brings the semantics in. In particular, it's this assumption that licenses the identification of the Normal situations with the ones in which causation makes content.

As it turns out, however, this key assumption—that when the situation is teleologically Normal, symbol tokens ipso facto apply to what they are caused by—is simply no good. What's true at best is that when symbol tokens are caused by what they apply to the situation is de facto teleologically Normal. Maybe it's plausible that when everything goes right what you believe must be true. But it's certainly *not* plausible that when everything goes right what causes your belief must be the satisfaction of its truth conditions. To put it still another way, if all that the appeal to Normal functioning allows you to do is abstract from *sources of error*, then the Normal situations are *not* going to be identical with the type one situations.

The glaring counterexample is the occurrence of representation *in thought*. Suppose, having nothing better to do, I while away my time thinking about frogs. And suppose that, in the course of this meditation, by a natural process of association as it might be, my thoughts about frogs lead me to thoughts about flies. The result is a token of the mental state type *entertaining the concept FLY*, which is, surely, caused in a perfectly Normal way (the teleology of mental functioning may abstract from *error*, but surely it doesn't abstract from *thinking*). But it is *not* an instance of an intentional state that was caused by what it means. What caused me to think about flies was thinking about frogs; but the effect of this cause was a thought about flies for all that. It may be that there are causal connections to flies *somewhere* in the historical background of thoughts about flies that are prompted by thoughts about frogs. But such thoughts haven't got the sort of causal histories that Skinnerian/Dretskian accounts contemplate the reduction of content to: they aren't *occasioned* by flies, and they don't carry information about flies in any sense in which what symbols carry information about is their causes. Specifically, the "covering" law that connected my fly-thought tokening with its cause projects the relation between fly-thoughts and frog-thoughts, *not* the relation between fly-thoughts and flies.

Compare Papineau: " . . . sometimes [a belief] will be triggered by 'abnormal' circumstances, circumstances other than the one that in the learning process ensured the belief had advantageous effects and which therefore led to the selection of the disposition behind it. My suggestion is that the belief should be counted as false in these 'abnormal' circumstances—. . . the truth condition of the belief is the 'normal' circumstance in which, given the learning process, it is biologically supposed to be present" (pp. 65–66). The basic idea is that all of the following pick out the same state of affairs:

- P's truth condition,
- the 'normal' (viz., the Normal) circumstance for entertaining P,
- the situation in which P is biologically supposed to be present.

But this can't be right. Thinking is a circumstance in which beliefs are, often enough, Normally entertained; and, I suppose, it's a circumstance in which biology intended that they should occur. But the matrix of mental states in which a belief is tokened in the course of mental processing is patently not to be identified with its truth condition. (Here as elsewhere, coming down heavily on "the learning process" doesn't help much. Lots of words/concepts aren't learned ostensively.)

This is, I think, a real problem. In fact, it's the disjunction problem in still another guise. What we want is that fly-occasioned "fly"s, and bee-bee occasioned "fly"s, *and representations of flies in thought* all mean FLY. At best, teleological solutions promise to allow us to say this for the first two cases—bee-bee-occasioned tokens are somehow 'abNormal'; hence not type one; hence their causation is not relevant to the content of "fly"—though we've seen that it's a promise that they welsh on. But teleological theories don't even pretend to deal with the third case; they offer no reason not to suppose that fly-thoughts mean *fly or thought of a frog* given that both flies and thoughts of frogs normally cause fly-thought tokens.

God, by definition, doesn't make mistakes; His situation is always Normal. But even God has the disjunction problem on the assumption that the content of His thoughts is determined by their causes and that some of His thoughts are caused by some of His others. The sad moral is, we still have the disjunction problem even after we idealize to infallibility.

I think a lot of philosophers (and a lot of psychologists in the Dewey/Gibson/American Naturalist tradition) believe deep down that content starts with perceptual states that are closely implicated in the control of action. It's perception—and, specifically, such perceptions as eventuate in characteristic corresponding behaviors, as in orient and capture reflexes—that provides the aboriginal instance of intentionality. Thought and the like come later, *not just phylogenetically but also in the order of explanation*. Thus, Israel remarks that, in theorizing about naturalized semantics, "it makes sense to look first at perceptual states of living organisms before moving on to anything more sophisticated" (p. 6). Since, as we've seen, Israel holds that the content of a state is determined by its function, he must be assuming that the function of perception is, at least in principle, dissociable from its role in the fixation of belief;[25] if the connection between perception and belief fixation is internal, the advice to look at perception first doesn't noticeably simplify the theorist's problems.

But even on this dubious assumption, this is dubious advice. Presumably, perception and thought are intentional *in the same sense,* so it's likely that a semantics that works only for the former works for the wrong reason. In perception there is generally a coincidence between what a cognitive state carries information about and what it represents (viz., between its Normal cause and its intentional object). But the intentionality of thought shows that this coincidence is an artifact; it's not essential to content.

In light of all this, I'm inclined to think that the teleological story about content is just hopeless. On the one hand, the appeal to teleologically normal conditions doesn't provide for a univocal notion of intentional content; specifically it doesn't solve the disjunction problem. And, on the other hand, type one situations can't be identified with teleological Normal conditions; it's just not true that Normally caused intentional states ipso facto mean whatever caused them. So we need a nonteleological solution of the disjunction problem. So be it.

Notes

1. This would be true even if, as functionalists suppose, physicalistic formulations of *necessary and sufficient* conditions for being in psychological states are typically not lawlike.
2. Some intentional laws constrain the relations among the states of a given organism at a given time (e.g., ceteris paribus, if you believe P & Q then you believe P). These laws could generalize even over organisms that had *none* of their mental states in common; in the present case, there's no P or Q that two organisms both have to believe in order that both should fall under the law.

   But laws that quantify into opaque contexts, e.g.: *(x) (y) (if x believes that y is dangerous then ceteris paribus x tries to avoid y),* look to be in deep trouble if holism is true, since such laws purport to generalize over organisms *in virtue of the shared intentional contents of their mental states.* Similarly for laws that constrain the mental states of a given organism across time, including, notably, the laws that govern belief fixation in reasoning, learning, and perception (about 96.4% of serious psychology, at a rough estimate). Suppose, for example, that it's a law that, ceteris paribus, the more of the *x*s an organism comes to believe are *F*, the more the organism comes to believe (*x*) *Fx*. Such a law would presuppose that an organism can hold the same (quantified) belief for different reasons at different times. But it's hard to square this with an intentional holism that implies that changing any one of one's beliefs changes the content of all the rest.
3. To avoid repetition, I shall use this as a technical term for a theory of content that is both physicalistic and atomistic; i.e., a theory according to which (i) and (ii) are both false.

4. Maybe it starts earlier—with the breakdown of image theories of Ideas. The theory that Ideas refer to what they resemble is, after all, both physicalistic (on the assumption that resemblance is some sort of geometrical relation and that physics contains geometry) and atomistic (since, presumably, what one of one's Ideas resembles does not depend on what other Ideas one has). Alas, the image theory, though naturalistic, is, by general consensus, untenable.
5. Quine isn't, of course, the only one. See the first two chapters of Putnam's *Representation and Reality* (1988) where it's assumed without *any* argument that if you're holist about confirmation you've got to be holist about meaning too.
6. On this view, there's an interesting analogy between the semantical role of the theories that one espouses and the semantical role of the instruments of observation that one deploys: They both just function to sustain the head/world coordinations that constitute meaning. As I remarked in *Psychosemantics* (1987), the Operationalists were right in thinking that "star" means *star* because we have procedures that have stars on one end and "star"s on the other; they went wrong—they stumbled into holism—by supposing that such procedures are *constitutive* of meaning, so that "star" meant something different with the invention of telescopes.

   By the way, not just one's own skills, theories, and instruments, but also those of experts one relies on, may effect coordinations between, as it might be, "elms" in the head and elms in the field. That would be quite compatible with the meaning relation being both atomistic *and individualistic,* assuming, once again, the Skinnerian view that the conditions for meaning are purely functional and that they quantify over the mechanisms that sustain the semantically significant functional relations. Putnam (1988) argues that since appeals to experts mediate the coordination of one's tokens of "elm" with instances of *elm,* it follows that "reference is a social phenomenon." Prima facie, this seems about as sensible as arguing that since one uses telescopes to coordinate one's tokens of "star" with instances of *star,* it follows that reference is an optical phenomenon.

   That Putnam, of all people, should make this mistake is heavy with irony. For, it is Putnam who is always—and rightly—reminding us that ". . . 'meanings' are preserved under the usual procedures of belief fixation . . . " (1988, chapter 1, p. 14). I take this to be a formulation of anti-instrumentalist doctrine: the ways we have of telling when our concepts apply are *not,* in general, germane to their semantics. Why, I wonder, does Putnam make an exception in the case where our way of telling involves exploiting experts?
7. The nicety at issue is that my revised Skinnerian story isn't, strictly speaking, naturalistic as I've been telling it: it requires a counterfactual supporting correlation between dogs and *dog-thoughts* (token states of entertaining the concept DOG); and, 'is a dog-thought' is a nonnaturalistic predicate; it picks out a thought by reference to its intentional object. Skinner gets around the corresponding problem in the original version of his theory by (tacitly) assuming that he can specify the content-bearing expressions of natural languages "formally": e.g., phonologically or orthographically. (Thus, the regularity in virtue of which the English word "dog" expresses the property *dog* connects instances of *dog* with tokens of the expression #"d"^"o"^"g"#.) A Skinnerian semantics for mental states would have to assume analogously formal specifications for the tokens of mental states.
8. This may not strike you as sounding a lot like Dretske. That's because—at least as late as the BBS Precise (1983)—Dretske actually has two stories about content running together. There's the one I've sketched in the text, which takes the notion of nomic connectedness as basic; and there's one that's elaborated in terms of

conditional probabilities (roughly, whether an event el carries information about an event e2 is a function of the conditional probability of e2 given el). It's not clear just how these two theories fit together, or what the second one buys you that the first one doesn't. To give just one example, on the nomic-connectedness story, the transitivity of 'carries information about' (what Dretske calls the "Xerox Principle") follows from the transitivity of 'is lawfully connected to'; on the conditional probability story, by contrast, it requires special stipulation. (Specifically, it requires the stipulation that el carries information about e2 only if the conditional probability of e2 given el is one.)

I think that the conditional probability story is a dead end and that connecting content to nomic relatedness is the really interesting idea in *Knowledge and The Flow of Information*. Anyhow, I propose to read Dretske that way for purposes of this discussion.

9. A subsidiary argument is that it's required to guarantee the Xerox principle. See preceding footnote.

10. According to this view, a semantic theory provides a naturalized condition for content in terms of nomic relations among properties; roughly, the symbol S expresses the property P if it's a law that Ps cause S-tokens. This condition is perfectly general in the sense that it can be satisfied both by atomic symbols and complex ones. Correspondingly, the appeal to recursive ("Tarskian") apparatus in a semantic theory functions *not* as part of the definition of content, but rather to show how the conditions for content could be satisfied by infinitely many formulas belonging to a productive system of representations. The idea is that content emerges from lawful relation between tokenings (in the world) of the property that a symbol expresses and tokenings (in the organism) of the symbol; and the internal representation of the Tarskian apparatus is part of the computational mechanism that mediates this lawful relation.

These remarks are intended to soothe philosophers who hold that ". . . a Tarskian truth characterization . . . makes no contribution at all to a solution of the problem of intentionality for semantic notions . . . [because] even if the inquirer has a materialistically acceptable explanation of what it is about the simpler sentence A and its relation to the world that makes it true, he gets no help at all from the truth definition in his search for an explanation of the physical basis of the semantic status of the complex sentence" (Stalnaker, 1984, p. 31). Still there's something to what Stalnaker says. As we'll see in chapter 4, no nomic connection theory could account for the content of complex predicates that can't be instantiated (e.g., "is a square circle" and the like). And, for just the reason that Stalnaker points out, adding Tarskian apparatus doesn't help with the naturalization problem in these areas.

11. As F2 understands 'information carried', there is a metaphysical assumption that if *x* causes *y*, then there are properties of *x* and *y* in virtue of which it does so, and there is a law that subsumes ("covers") the causal interaction and relates the properties. See also chapter 5.

12. This approach to the disjunction problem thus exhibits a certain spiritual affinity with 'paradigm case' arguments in epistemology. Both assume that there are situations such that the fact that a sort of symbol is applied to a sort of thing in those situations is *constitutive* of the symbol meaning what it does. '"Dog" can't but be true of Rover because it's constitutive of the meaning of "dog" that Rover is a paradigm of the kind of thing that one says "dog" about. So pooh to people who think that there's a skeptical doubt about whether there are dogs!' But if this is not to beg the argument against skeptics, 'Rover is a paradigm of the kind of

thing that one says "dog" about' can't mean 'Rover is the kind of thing that "dog" is true of'; rather, it's got to mean something like 'Rover is the kind of thing that "dog" is said of'. And now there needs to be a caveat: viz., Rover has to be the kind of thing that "dog" is said of *when the conditions for dog-spotting are pretty good.* (There are *other* conditions—dark nights and such—when cats are paradigmatic of the kind of thing that "dog" is said of; a consideration that's grist for the skeptic mill.) In effect, paradigm case arguments presuppose that there is a distinction between type one situations and others; and that dark nights don't count as type one situations for saying "dog." It was not, however, in the tradition of paradigm case arguments to be explicit about much of this.

13. Cf. examples like *normal pulse rate* rather than examples like *snafu*. I shall follow the convention initiated by Ruth Millikan and write "Normal" with a cap N when I want to stress that a normative rather than a statistical notion of normalcy is intended.
14. I should emphasize that what's being denied here isn't just the statistical claim that all or most or much of the time if you want to become rich and famous you do become it. I'm claiming that a situation in which somebody wants very much to become rich and famous can be perfectly Normal in *any* reasonable sense of the term, and yet what's wanted very much may nevertheless fail to come off. This seems to me to be a truism.
15. Notice that Normalcy isn't a statistical notion even on this account. It's assumed that if *X* Normally causes *Y*, then *if the situation is Normal* then if *X* then it's relatively likely that *Y*. This is, of course, perfectly compatible with *X*s never causing what they Normally cause because the situation is never Normal. Dennett (in a 1988 manuscript called "Fear of Darwin's Optimizing Rationales") succumbs to ill temper because he thinks I have misread Millikan as proposing a statistical account of normal functioning. But she doesn't and I haven't and none—I mean *none*—of the arguments I've proposed depends upon assuming that she does. I am a little miffed about this.
16. So, to keep the record straight: whereas Millikan apparently wants to define the content of a belief state in terms of its selectional history, the alternative proposal defines belief content by reference to the teleology of the belief fixing mechanisms (roughly, a belief is about what would cause it to be tokened in the sort of circumstances in which the mechanisms of belief fixation were designed to operate). The present proposal includes both nations so as not to prejudice the case against either.
17. Though other sorts of teleological accounts are not precluded in principle, I assume in what follows that any naturalistic story about teleology is going to rest on some sort of appeal to evolutionary history. But actually, as far as I can tell, the main line of argument goes through just as well if it's assumed only that the account of teleology is consequentialist and not subjunctive; i.e., that the purpose of a biological mechanism is somehow determined by the good results it (actually) produces, whether or not *good result* is itself construed in terms of selectional advantage.
18. Millikan has this to say about the frog/fly/bee-bee example: "We say that the toad thinks the pellets are bugs merely because we take it that the toad's behavior would fulfill its proper functions (its 'purpose') Normally only if these (viz., the pellets) were bugs *and* that this behavior occurs Normally (not necessarily normally) only upon encounter with bugs" (pp. 71–72). But assume that the toad thinks that the bee-bees (and the bugs) are black spots (so the bee-bee elicited snaps are "true"). If the Normal environment for snapping at black spots is one

where black spots are predominantly bugs, it still goes through that frog snaps at bee-bees would fulfill their proper functions Normally only if the bee-bees were bugs. This is because, in the cases where the black dots that the frog snaps at *aren't* bugs, the environment, ipso facto, isn't Normal. And, for the same reason, it still goes through that frog snaps occur Normally "only upon encounter with bugs." So we still haven't got a solution to the disjunction problem even after we've satisfied the conditions that Millikan imposes; i.e., satisfying her conditions on the Normal function of frog snaps is compatible with taking the intentional objects of the snaps to be (not flies but) little black dots.

19. Millikan and Israel are by no means the only philosophers who are hoist on this petard (whatever, precisely, a petard may be). David Papineau, who runs a teleological line on content in *Reality and Representation* (1988), suggests that ". . . the biological function of any given desire type is to give rise to a certain result: the result is then the desire's satisfaction condition" (p. 64). But this assumes that a naturalistic account of the teleology of desires will specify a *unique* biological function for each desire type; in particular, it supposes that the teleology will be univocal in cases where the disjunction problem would otherwise make intentional content indeterminate. Papineau provides no argument that natural teleology is univocal in this respect, and we've just seen why, if it's grounded by appeals to selection, it pretty clearly won't be.

Correspondingly, Papineau suggests that "the truth conditions for beliefs are . . . the circumstances in which they will have effects that will satisfy the desires they are working in concert with." Well, suppose that what the frog desires is food; suppose, even, that what it desires is that it should ingest flies. It's still true that (given Normal circumstances), *either* the belief that there are flies *or* the belief that there are black dots will have effects that will satisfy the frog's desire.

It's also true, of course, that snapping at black dots won't satisfy the frog's desire for flies in the *abNormal* circumstance where the black dots are bee-bees; and some of the things that Papineau says (p. 72) suggest that he wants to rest on this. But that won't do since there are other, also abNormal, circumstances in which *snapping at flies* won't satisfy the desire to ingest flies either (the frog's tongue is covered with silicon, and the flies slip off; the flies are of a new high-tech variety and can fly faster than frogs can snap, etc.). The moral is that you can rely on the frog's fly-beliefs leading to fly-ingestions (and thus bestowing selectional advantage when entertained in the presence of flies) only if you are taking it for granted that the frog's ecology is Normal. But then we've just seen that if you *are* taking it for granted that the frog's ecology is Normal, the requirement that its beliefs should operate in conjunction with its desires to produce successes isn't strong enough to motivate unique assignments of intentional content to the beliefs. Dilemma.

20. Strictly speaking, given the possibility of higher-order conditioning, it may be that getting an organism to respond to the triangularity rather than the greenness of green triangles doesn't depend on *green* and *triangle* being dissociated in the course of training, so long as *some* colors are dissociated from *some* shapes. A general habit of responding to shape rather than color could perhaps be established by differential reinforcement in those cases. I have no idea whether this would actually work, and, anyhow, it's just a curiosity; it suggests, contrary to fact, that if "green iff triangular" is reliable, it can't be that an organism is responding to triangularity rather than greenness unless it has a disposition to respond to shape rather than color *in general.*

21. It should now be clear that the argument against Darwinian theories of content was, in effect, that "Mother Nature" can select for organisms that snap at flies—as opposed to organisms that snap at fleebees—only if she can perform a "split stimulus" experiment; i.e., only if she can contrive to present the frog with fleebees that aren't flies; i.e., only if she can contrive to present the frog with fleebees that are bee-bees; a fortiori, only if "all the fleebees are flies" isn't reliable in the frog's ecology.
22. I'm very grateful to David Rosenthal for a conversation that helped to get this sorted out. The saccharine case isn't exceptional, by the way; any example of what ethnologists call a "supernormal" stimulus serves to point the same moral.
23. Till now I've been arguing that appeals to selectional history can't distinguish an organism that *represents things* from *F* from an organism that represents them as *G* in a world where it's counterfactual supporting that all and only the *F*s are *G*s. A parallel line of argument secures the present claim that appeals to evolutionary history can't distinguish selection for *being F* from selection for *being G* when *F* and *G* are necessarily coextensive: If you always get *F*s and *G*s together, then a mechanism that selects one *thereby* selects the other, so the utility of being *F* and being *G* always comes out the same.

    This has philosophically interesting consequences. For example, even assuming that it's a law that hearts and only hearts make the noises they do, still it's intuitively plausible that the function of the heart is pumping the blood, not making the noises. If the line of argument I've been selling is right, then appeals to selectional history do not, in and of themselves, underwrite this intuition. This does not, of course, imply that it's false that the function of the heart is blood pumping; it only implies that facts about function don't reduce to facts about selectional history. Dennett (1987) says that "if you want to maintain that it is perfectly respectable to say that eyes are for seeing . . . you take on a commitment to the principle that natural *selection* is well named . . . there is not just selection *of* features but selection *for* features . . . without this 'discriminating' prowess of natural selection, we would not be able to sustain functional interpretations at all" (p. 316; his italics). But no argument is given for this, and, as we saw above, it could turn out that function gets an analysis in terms (not of selectional history but) of *counterfactuals*. The governing intuition is, perhaps, that it would be OK if the heart stopped making noise as long as it kept pumping, but not so good the other way 'round.
24. Similarly, mutatis mutandis: Teddy bears are artificial, but *real bears are artificial too*. We stuff the one and Mother Nature stuffs the other. Philosophy is *full* of surprises.
25. The idea that "the" function of perception is to guide movement rather than to fix belief is also a main theme in the American Naturalist tradition; and in what is sometimes described as the evolutionary approach to the mind (see Patricia Churchland, 1987). For discussion, see chapter 9.