

Representation and Rationality

Anthony Dardis
University of California, Berkeley

August 19, 1990

Representation and Rationality

By

Anthony Buckelew Dardis
B.A. (Columbia University) 1977

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

PHILOSOPHY

in the

GRADUATE DIVISION

of the UNIVERSITY OF CALIFORNIA at BERKELEY

Approved: Donald Davidson (Chair), Bruce Vermazen, Lotfi Zadeh

For Marla and Eli

Acknowledgments

Many people have provide essential help to me in writing this dissertation. In addition to those mentioned [in the text] I would like to give special thanks to Tom Dardis, Jane Buckelew, Sue Larson, George Myro, Bruce Vermazen, Donald Davidson, Lotfi Zadeh, Jerry Fodor, Fred Dretske, Alistair Hannay, Bonnie Glaser, Eric Goodwill and Robert Ruggiero.

Contents

1	Content Rationalism	3
1.1	Introduction	3
1.2	Content Rationalism	3
1.3	Fodor on Meaning Holism	6
1.4	A Map of the Territory	9
1.5	(1) From Confirmation Holism to Meaning Holism	11
1.6	(2) From the Theory of Meaning to Content Rationalism	13
1.7	(3) Axioms of Decision Theory and Constitutive Concepts	16
1.8	(4) From Action Theory to Content Rationalism	17
1.9	(5) Testing Decision Theory and Content Rationalism	20
1.10	Conditional Arguments	24
1.11	Content Rationalism and Naturalism	25
1.12	Conceptual Connections without Analyticity	27
2	Kitty Thoughts	29
2.1	Introduction	29
2.2	Exegetical Preliminary	30
2.3	Animal Thought and Intensionality	31
2.4	The Central Arguments	37
2.5	Believing Requires The Concept of Belief	39
2.6	Having the Concept Requires Speech	43
2.7	Conclusions	48

3	Causal Semantics	51
3.1	Introduction	51
3.2	Stampe's Causal Theory	53
3.3	Motivations	56
3.4	The Singularity Condition	60
3.5	The Content Condition: Function and Causation	61
3.6	Stampe's Theory Applied to People	65
3.7	Conclusions	66
4	Dretske's Theory of Content	69
4.1	Introduction	69
4.2	Dretske's Theory	70
4.3	Indeterminacy	76
4.4	Indication	80
4.5	Needs, Life, Value	84
4.6	Applying the Account to Propositional Attitudes	85
4.7	Conclusions	88
5	Biosemanitics	91
5.1	Introduction	91
5.2	Millikan's Theory	93
5.3	Function and Philosophical Methodology	97
5.4	Problems for Millikan's Account of Content	105
5.5	Millikan on Other Functional Accounts	110
5.6	Millikan and Psychophysical Laws	113
5.7	Conclusions	115
6	Psychosemantics	117
6.1	Introduction	117
6.2	Fodor's Theory of Content	119
6.3	Perceptualism	124
6.4	Fodor's Theory for Perceptualist Contents	125

6.5	Fodor's Theory and Non-perceptualist Contents	127
6.6	Fodor's Attacks on Teleological Accounts	130
6.7	Intensionality	135
6.8	Semantic Atomism and Content Rationalism	136
6.9	Conclusions	140
7	Psychophysical Laws	143
7.1	Introduction	143
7.2	The Argument for Anomalous Monism	145
7.3	The Argument for Anomalism	146
7.4	Support for the Failure of Lawlikeness	152
7.5	Rationality in Physics	155
7.6	Conclusions	159
8	Causal Relevance	161
8.1	Causal Relevance and Conceptual Connections	162
8.1.1	Introduction	162
8.1.2	Metaphysical Background	163
8.1.3	Notation and Terminology	165
8.1.4	A Weak Condition	166
8.1.5	Justifying the Condition	168
8.1.6	Details and Elaboration	170
8.1.7	A Stronger Condition	173
8.1.8	The Strongest Condition	175
8.1.9	Conclusions	178
8.2	Causal Relevance and Strict Law	179
8.2.1	The Basic Problem, and the Solution	179
8.2.2	Sosa's Principle and Supervenience	182
8.3	Causal Relevance and Externalism	189
8.3.1	Externalism	189
8.3.2	Fodor's Argument	192
8.3.3	Screening Off Revisited	196

8.3.4	The Mental Isn't Screened Off	197
8.3.5	Rejecting the Principle	201
8.3.6	Probabilistic Causality	203

Introduction

How are we related to the physical world? We see, hear, touch the world, believe things about it, want things from it, take joy in it. Our physical theories describe the precise interactions of kinds of events. These theories never speak of objects that see, believe, or desire *as* seeing, believing or desiring. How then are these characteristic aspects located in the world described by physical theory?

We must, I believe, avoid two answers. We literally do have these characteristic aspects: we cannot simply deny that they exist. Further we are part of the world, not somehow separate from or outside the world of physics.

How can it be that we are part of the world described by physics, yet physics never describes us? The answer depends on the nature of these characteristic aspects of ourselves. Seeing, hearing, believing and desiring are all intentional states, ones which are about other things and which may represent them correctly or incorrectly. The concept of representation essentially involves the concept of rationality. This concept plays no role in understanding the physical world, but it is crucial to understanding ourselves.

This conception of our place in the physical world is central to Donald Davidson's philosophy of mind. In what follows I argue for this conception and discuss various of its difficulties.

I begin by rejecting one set of arguments to the conclusion that thought is rational. I construct a new argument based on recent work on "naturalizing" representation. "Naturalizing" in this case means stating an account of representation that does not use the concepts of value or rationality. These theories aim at an account of representation that applies to many kinds of representational systems other than human beings. I show that these accounts must use the concepts of value, and that in the case of human thought they must use the concept of rationality.

Chapter 7 evaluates the prospects for one of Davidson's arguments that there cannot be a rigorous science of psychology, on the ground that the concepts of

representation and rationality are connected. If this argument succeeded it would constitute a proof that physical science could never include psychology as a part. Although I do not believe the argument succeeds, I think it shows that we have overwhelmingly good reason to believe its conclusion.

Many have felt that such a conclusion leaves us too far outside the world of physics: if there were no rigorous science of psychology then we must admit that our mental lives are as “motes above the stream,” an offshoot of the basic causal processes of the world, but an offshoot which has no causal bearing on anything else. There is a variety of reasons to believe this; in Chapter 8 I consider several and conclude that none are good.

Chapter 1

Content Rationalism

1.1 Introduction

Donald Davidson holds that application of the predicates of propositional attitude is governed by the “Constitutive Ideal of Rationality:” if a creature has a single propositional attitude then it has indefinitely many beliefs, desires and intentions, and the resulting cognitive set must be rational. The intuitive pull of this claim is not great. In this Chapter I describe Jerry Fodor’s reasons for doubt (Fodor, 1987, Chapter 3). I survey the arguments Davidson presents for the claim and show why they fail to convince a skeptic of Fodor’s persuasion. This survey shows why attempts (like Fodor’s) to give a reductive naturalistic account of content have a certain plausibility. In later Chapters I give a new argument that content must be rational: these reductive naturalistic accounts all fail by begging the question about rationality. I end this Chapter with some remarks about the apparent inconclusiveness of the arguments on both sides.

1.2 Content Rationalism

I shall call Davidson’s view about rationality and the propositional attitudes “content rationalism.” In this section I describe the view.

Content rationalism is a claim about how people are related to propositions. I do not take propositions to be metaphysically basic in any sense; the point of talk of persons and their relations to propositions can be honored by talk of per-

sons and their relations to things. People are related to propositions through their propositional attitudes, like believing, desiring, intending and saying. Propositional attitudes can be separated in concept into two components: the attitude, like that of saying or intending, and the content or proposition. Many things in the world are related to propositions only through their relations to persons. A sentence on a page has a certain meaning in virtue of having been written by a person; a painting represents a face or a political situation in virtue of its connection with the painter. Talk of meaning, content and propositions obeys some clear principles of interchange: an utterance has a meaning, a thought has a content which is the same as the meaning of the utterance that would express it, and an utterance that means that *p* expresses the proposition that *p*.

What *kind* of thing is a propositional attitude? Propositional attitudes are events: particular occurrences that have many properties and can be described in many ways. Believing seems to be a state of a person, but it is a state that a person begins to have and will leave off having. We can think of the state of believing as the event of coming to believe. This event may occupy a long stretch of time: until the person changes her belief.

Content rationalism is the claim that if a person has a propositional attitude then she has an enormous number of propositional attitudes and the relations among the attitudes are rational. The person who has the attitudes may not be rational. This can happen if the person has several (perhaps substantially overlapping) sets of propositional attitudes, where each set is rational, but where the sets conflict. The claim is still quite strong in requiring that each propositional attitude is surrounded by a rational set.

What does “enormous number” mean? Take the belief that it was raining on Tuesday. For anyone to have this propositional attitude she has to “command” the various concepts involved (of rain, of Thursday, of weeks, of days, of time, of water, etc.). For her to command any one of these concepts, say, that of rain, she has to command many others (the concepts of stream, lake, ocean, sky, cloud, clear colorless liquid) in ways that ensure that she has this concept rather than some other, perhaps broader in application. For her to command these concepts requires that she have beliefs about these other things. Hence in order to have a single belief a person needs to have some beliefs about a large number of the kinds of things that the single belief involved and some beliefs about the kinds of things anyone should know about if she is to count as having a belief about the things the single belief is about. Hence she would need to have some beliefs about most of the things we all ordinarily run into in going about our lives.

What does it mean to say that the sets are rational? What follows is a list of some things that should be true; it's not meant as a theory of rationality.

Calling a set of attitudes rational is appraising it, but this does not mean that the agent who has it believes what we think she should believe. I may think smoking cigarettes is clearly a dumb thing to do, but someone may know the things that I know about smoking cigarettes and also value the pleasures of smoking far more than avoiding the dangers. The appraisal is meant to be internal: as far as each set goes, the thought and actions that it involves make sense from the perspective of that set.

A rational attitude set is one which displays various kinds of consistency. The beliefs should be logically consistent. Beliefs arrived at through experience should be inductively sound. Beliefs about one's own beliefs and desires should be correct. If an agent believes she ought to make a certain inference, the presence of the inference makes the attitude set more rational. Conflicting desires will always coexist in agents, but a particular desire can be irrational if the agent has good reason to think she is worse off if she has it.

Attitude sets can be assessed for the rationality of the way they change through time and result in action. Decision theory gives a clear and precise model of practical deliberation. Dutch book arguments provide one way to test the adequacy of the model against our intuitions about rationality.¹ There must be no way for the agent to act on her beliefs and desires and lose out in the long run. Losing out means she fails to realize certain of her values, solely because of the way she thinks about things and values them. Let a person think a set of alternatives is exclusive and exhausts the possibilities, and assign (act as if she assigns) likelihoods to the possibilities that do not add to 1. There is a way to present her with gambles on those possibilities such that she will end up losing all her money.

It may be that certain beliefs or strengths of belief are the "right" ones to have. For instance causal decision theorists criticize Jeffrey's theory on the ground that the Bayes act will be irrational in certain decision problems if the agent's subjective probabilities are set by observed relative frequencies. A reasonable response is that these are not the right subjective probabilities.² Since Jeffrey has the agent's entire preference ranking determine subjective probabilities, it's likely that they will end up differing from what they would be if they were to match what

¹For doubts on the strength of the connection between Dutch book arguments and rationality, see (Adams and Rosenkrantz, 1980).

²For Jeffrey's theory, see (Jeffrey, 1983); for the objection, see (Skyrms, 1980, 128–139); for the response, see Piers Rawling, unpublished.

the agent thinks the relative frequencies are, since the agent will take the relevant causal relations into consideration in her preferences.

Finally, content rationalism is a modal thesis: what it is to be a propositional attitude is to be something connected with many others in the ways stipulated by the norms of theoretical, practical and epistemic reasoning.

1.3 Fodor on Meaning Holism

Is content rationalism true? In later sections of this chapter I will survey some grounds for a positive answer. In this section I want to describe one negative answer.

Jerry Fodor thinks content rationalism is false.³ He is vexed by arguments, like Stich's,⁴ that purport to show that there is no proper science of propositional attitudes and which suggest that, well, really there are no such things. Fodor finds the inference plausible⁵ and he thinks there surely are such things, so he takes pains to show there can be a proper science. He thinks the most serious threat to the scientific character of the propositional attitudes is posed by a loose collection of doctrines he calls "Meaning Holism." He characterizes the whole family this way:

When an intentional system takes the semantic value [truth] of P to be relevant to the semantic evaluation of Q, ... P is an epistemic liaison of Q ... Meaning Holism is the idea that the identity—specifically the intentional content—of a propositional attitude is determined by the *totality* of its epistemic liaisons. (p.56)

If Meaning Holism is true, since each person's epistemic liaisons differ, no two people share a belief. Meaning Holism thus has the consequence that we could never hope for any interesting empirical generalization about the behavior of the propositional attitudes.

³It is important to note that he thinks belief or attitude rationalism is true. The reason is that he thinks beliefs or propositional attitudes are contentful states of which decision theory is true. Thus in some sense Fodor does think content rationalism is true, but the reason is just that that's the way he defines belief, rather than the intimate connection between content and rationality.

⁴See (Stich, 1983). (Dennett, 1987c) has a good summary of other positions inspired by Quine's skepticism about meaning.

⁵He shouldn't. There is no reason to hold that instances of a category of things are real only if there is a science of their behavior that meets certain criteria.

Why would anyone hold such a crazy position? Fodor proposes that the general structure of all arguments for Meaning Holism take the following form (this is the “Ur- Argument for Meaning Holism”, p.60):

Step 1. Argue that *at least some* of the epistemic liaisons of a belief determine its intentional content.

Step 2. Run a “Slippery Slope” argument to show there is no principled way of deciding *which* of the epistemic liaisons of a belief determine its intentional content. So either none does or all do.

Step 3. Conclude that they all do (1, 2: modus tollens).

Stich has several arguments that Step 1. is correct. Mrs. T, for instance, is a woman who still responds, “Oh, McKinley was assassinated,” to the question, “Mrs. T., tell me what happened to McKinley,” but she is unable to say where McKinley is now, whether he is alive, who he was, etc. Stich claims, and Fodor agrees, that Mrs.T does not believe that McKinley was assassinated. Stich goes on to say that the reason she does not believe this is that she *lacks* the sort of beliefs someone must have if she is to have the belief that McKinley was assassinated. Fodor says this is the fallacy *post hoc, ergo propter hoc*: it does not follow from the fact that she lacks this central belief *and* that she lacks these others, that she lacks this central belief *because* she lacks the others. Fodor draws the obvious conclusion: that step 1 is false, and that none of the epistemic liaisons of a belief determine its intentional content.

“But look, if you radically detach content from functional role, then why does one have to draw *any* consequences from one’s thoughts? On your view, entertaining (as it might be) the thought that three is a prime number could constitute *an entire mental life*?” This too is satisfactory as a reductio ad absurdum argument only on the assumption that the conclusion is false. But its conclusion doesn’t strike me as *self-evidently* false. (p.89)

Fodor considers 3 other arguments for Meaning Holism (pp.62-84): one from confirmation holism (a style of argument favored by Quine and Putnam), one from functionalism in the philosophy of mind, and one from functionalism in semantics. The first argument is interesting and important because it touches on the ground many philosophers actually have for holding something like Meaning Holism. I shall remain silent about the other two.

The first argument works with a cartoon version of Quine's answer to positivist theories of meaning. The positivists held that what any sentence means is set by its relation to experiences that would confirm it. Some sentences are observational; they are confirmed by experiences only, and never by other sentences. Their meaning is that the relevant experience is occurring. The sentences that are not observational are confirmed on the basis of other sentences. Ultimately the chain bottoms out in experience. The meaning of non-observational sentences is set by the range of observational sentences that contribute to their confirmation. These cartoon positivists hold, for the non-observational sentences, that (a) what meaning they have is determined by their (confirmation) relations to observational sentences, and that (b) to each non-observation sentence there corresponds a distinct and determinate set of observation sentences that determines its meaning, i.e., the relations are *direct*.

Quine, following Duhem, pointed out that confirmation relations are never direct in this sense. Any observation can be relevant to any sentence. An observation of a black raven can be taken to confirm, "All ravens are black," or it can be taken to disconfirm that sentence (on the assumption that observation conditions were improper for viewing what was probably a white raven). Quine did not, however, abandon the idea that the meaning of these non-observational sentences was somehow determined by this much looser connection with experience. This is a characteristic feature of doctrines of meaning holism that people have actually held: what a sentence means should be some sort of holistic function of its confirmation relations to observable events.

Fodor has a peculiar response to this cartoon. Let us call the positivist position "Semantic Localism," since the observations relevant to the meaning of any sentence fall in a narrow range for each sentence. The holist position is "Semantic Holism," which says that the meaning of any sentence may be relevant to the meaning of any other sentence. Quine is famously skeptical about meaning, and famously skeptical about most things philosophers have said about meaning; Fodor thinks Quine is skeptical about meaning *tout court*. This makes for a third position, "Meaning Nihilism;" philosophers like Stich seriously attempt to take this position. But apparently there is a fourth position to take: deny that confirmation relations between sentences are relevant to meaning *at all*. This is Semantic Atomism. In the philosophy of mind this is the claim that the content of a belief or desire does not depend on the relations that attitude has with any other.

I will return to Fodor's take on the argument from confirmation holism to meaning holism, and on Quine's views on semantics, in section 5. Fodor has, I

think, done us the service of indicating a thesis which could use some support, the thesis that the epistemic connections of a belief are relevant to its content. This thesis is not self-evident and it is not entirely easy to see why it should be true.

Before I go on I should note the relations between Meaning Holism and content rationalism. Content rationalism is the claim that a necessary condition for the existence of one propositional attitude (with a certain content) is the existence of a whole collection of others, all of which are interrelated in the way our attitudes are interrelated. These ways properly include the epistemic connections definitive of meaning holism. Meaning holism is the claim that the intentional content of a single propositional attitude is determined by the epistemic connections it has with others. The same idea is often put this way: the content of a single attitude is constituted by these connections. (The phrase, “The Constitutive Ideal of Rationality” encourages this formulation.) Hence content rationalism is stronger than meaning holism in including more rational relations, and weaker than meaning holism in failing to make the claim about the constitution of content. I do not think there is more to the claim about constitution than a necessary condition: If A is constituted by relations R then if something is A then R holds of it. Hence I think if meaning holism is wrong then content rationalism is too. Both meaning holism and content rationalism are typically parts of a more complex story about the patterns of interaction between contentful states and other things that makes these states have the content they do. The thought behind both claims is similar; it is the thought that content is the sort of thing that comes about by way of a large pattern of interrelated events.

1.4 A Map of the Territory

There are two different distinctions we might make about content; these generate a space of 4 possibilities that I think is useful in sorting out what Fodor claims. We can either accept or reject content rationalism. And we can either accept or reject the claim that there is a natural account of attitude content.

For now I want to remain fairly vague about what an “account” of content is; I will discuss several such accounts in Chapters 3 through 6 below. I have in mind any explication of the ways non-intentional and non-semantic facts determine what content a thought has, or determine that a thought with a certain content exists. The account can be reductive, like the ones Stampe, Dretske, and Fodor offer; or it can be non-reductive, like Davidson’s. The alternative position

is that there are basic, non-derived facts about content.⁶

The first of the four possible positions is that of the natural rationalist, who holds that there can be an account of what it is about a person and her relations with things that makes it true that she has thoughts, and holds that those thoughts are by and large rational. Davidson is one such theorist. I think Fred Dretske is another; he holds that a reductive account can be given and that we can *explain* why content rationalism is true in terms of the reductive account.

The natural irrationalist holds that we can give an account of content, but that contents need not be rational. This is Fodor's official position.

The *sui generis* rationalist holds that there is no account to be given of content, but that content rationalism is true. Perhaps George Bealer is the clearest example. I think John Searle is another.⁷

Finally the *sui generis* irrationalist holds that content is a basic fact about things, and further holds that contents can occur in wildly irrational collections. This is atomist realism about content. It is the claim that there *just can be* free-floating instances of content properties. It's hard to be certain whether anyone ever held such a theory. Descartes' theory of mind and God might be a candidate. He is clearly a *sui generis* theorist; he holds that it is impossible to explain what it is to be a thought in physical terms.⁸ It may seem odd to call the father of Rationalism an irrationalist. He also holds that God could have made the eternal truths of reason differently than they are. Equally He could have made me such that my reasoning does not proceed at all in accord with the eternal truths (or even in accord with what I think are the eternal truths). This possibility is what makes for the worst of First Meditation skepticism, and what, for instance, makes it mandatory that the *cogito* not be an inference, since after all the principles of

⁶Typically externalist theories of mind and meaning are naturalistic in this sense. There is at least one externalist position that could be taken to be non-naturalistic. John McDowell (McDowell, 1986) holds that the identity of certain thoughts is determined by external objects. But for all McDowell says this could be a brute fact about how things are, not susceptible of any further explication, for instance in terms of the relations of the person who has the thought to the object.

⁷Searle's position is not completely clear. He holds that intentionality is a biological phenomenon. This suggests that some sort of biological account of content might be forthcoming. But I think Searle holds that content is really *sui generis*. See, for instance, his (Searle, 1984, 98–99) where he says that we have to realize that the sciences of man have to use metaphysically basic notions of teleology and intrinsic Intentionality. He is a content rationalist, although his version is considerably weaker than Davidson's, since unlike Davidson he holds that irrationality requires no special sort of explanation.

⁸See Margaret Wilson's comments on the end of the Second Meditation (Wilson, 1978, 92–99).

inference are among the candidates for doubt.

I think the main thing to be said for *sui generis* irrationalism is that it is not logically inconsistent. It just doesn't seem to follow from the statement that there is a thought that three is a prime number that there are certain other thoughts that stand in certain relations with this one. This is part of Fodor's official justification for his stance on content; it's just not self-evident to him that it is *absurd* that there should be a mental life comprising nothing but this thought.

A position that is not logically inconsistent may still be incoherent or unintelligible, or it may be inconceivable how it would be realized. It is not logically absurd that there should be such a mental life, but it is incomprehensible how it could happen. I will return to the bearing this reflection has on the truth of content rationalism below in section 12.

I suspect Fodor overlooks the distinction between *sui generis* irrationalism and naturalistic irrationalism. *sui generis* irrationalism is a hyper-realistic stance on content properties which can be maintained in the face of any argument for content rationalism I know of. The semantic atomism he counters to holism, localism and nihilism is just that stance. But Fodor is not a *sui generis* theorist. And the price for an acceptable naturalistic account will be meaning holism. Hence I think Fodor's challenge to meaning holism is unsuccessful. I'll argue this in more detail below in Chapter 6.

Here's how I am going to proceed in the rest of this chapter. Fodor seems to miss the force of the various arguments Davidson has given for content rationalism. Others inclined to attempt a naturalistic account of content miss it as well. Why is this? Imagine an irrationalist (I'll call her the Skeptic) who isn't sure whether content rationalism is true. She goes to Davidson's papers for an argument that will rationally compel her to believe it. I'll describe 5 different arguments the Skeptic finds, and the reasons she remains unmoved. This will set the stage for chapters 3 to 6 by showing why someone might think that an account of content might be given that denies content rationalism or attempts to explicate its truth.

1.5 (1) From Confirmation Holism to Meaning Holism

Fodor doesn't deny confirmation holism. His polemical strategy to show that semantic atomism is a viable position is to note that Quine doesn't argue from confirmation holism to meaning holism: Quine moves straight to meaning nihilism.

But Fodor overreads Quine's skepticism. Quine thinks that what many philosophers have said about meaning must be wrong, and he doesn't think there is any hope for a science of meaning. But the second chapter of *Word and Object* is precisely about meaning: it sets out what Quine thinks there is to be said about meaning. And what there is to be said patently includes meaning holism. Perhaps Fodor has underestimated Quine, then; perhaps there is an argument from confirmation holism to meaning holism.

How might such an argument go? Translation is a relation that preserves meaning. In translating we attend to confirmation relations and settle on the meanings of terms by seeing how they function in inferences from observation. Confirmation holism requires that translation of a single term must be sensitive to its actual and potential confirmation relations with everything an agent thinks. Hence confirmation holism together with plain fact about translation shows that content rationalism is true.

The trouble with this strategy is that confirmation holism alone does not entail meaning holism or content rationalism. Confirmation holism is compatible with semantic atomism. Hence any argument from confirmation holism will use some additional premise, like the one I just gave about the connection between translation and meaning, that the skeptic will reject. For instance the skeptic will respond that the norms that govern translation practice embody good epistemic policy but do not constrain semantic possibility. Attending to confirmation relations is a very good way to find out what someone means, but what the facts are is independent of the confirmation relations.⁹

⁹The skeptic may have trouble saying *why* this is good epistemic policy. Skepticism about content rationalism is in this respect like other kinds of epistemological skepticism. Skeptics generally respond that we should not posit very strong connections between concepts just because it is very difficult to see how we know the things we do. This pattern of argument and response is familiar in other debates about realism. See, for instance, David Armstrong's rejection of "sophisticated" Humean theories of natural law (Armstrong, 1983, Chapter 5, 71–2). He complains that the "sophistication" confuses epistemology with metaphysics, that it makes the epistemology of laws constitutive of lawhood rather than just good policy. A natural law, on his view, is a necessitation relation that obtains between universals. Whether such a relation obtains is logically independent of the sort of evidence we would use to decide whether it obtains.

1.6 (2) From the Theory of Meaning to Content Rationalism

Davidson suggests that a theory of meaning for a speaker's utterances should take the form of a theory of truth for a language. This suggestion provides a remarkably valuable framework within which to raise questions about language, thought and meaning. The suggestion is also intended to tell us about the nature of meaning. Davidson asks what knowledge someone could have that would suffice to interpret another's utterances. Knowing an interpretive truth-theory, and knowing that it *is* such a theory, would do the trick; what is it, then, to know such a theory and know that you know it? We can answer these questions by seeing how to confirm that a given theory of truth interprets a speaker. Seeing how to confirm a theory of truth is seeing how facts about meaning depend or supervene on facts that can be known in advance of knowing how to interpret a speaker. Hence an inquiry into how a truth theory is confirmed is an inquiry into the nature of meaning.

The appearance that we are on the verge of repeating the move of the last section is partly dispelled by the doubly hypothetical character of this investigation: it is an investigation into how someone could confirm a theory which could be used to interpret. The appearance is not entirely dispelled, though. To show this I'll need to elaborate how the argument to content rationalism is supposed to go.¹⁰

Theories of truth can be given for all kinds of languages, most of which are never used. How might we go about telling whether a theory of truth for a language is a theory of meaning for a particular person? The basic strategy is as follows. The theory entails sentences of the form,

(T) An utterance of S by A at t is true iff p.

We find some empirical condition such that if A satisfies the condition at t then an utterance of S by A at t has occurred. We survey what we believe to be the facts. If other information we have about A suggests that this utterance of S is specially likely to be true, and we believe that p, we have reason to suppose that this consequence of (T) is confirmed. Clearly we would not have very much reason; confirmation of a theory of truth is based on an enormous amount of information of this sort. Considerations of structure constrain the number of theories that are

¹⁰This argument runs through Davidson's papers on radical interpretation in (Davidson, 1984b), especially "Truth and Meaning", "Belief and the Basis of Meaning", "Radical Interpretation", and "Thought and Talk".

confirmed; so does the demand that the theory should support counterfactuals. It is to be expected that if one theory can be confirmed infinitely many can. The hope is to find a variety of constraints that narrow the kinds of acceptable theories to those that capture the intuitive notion of meaning well enough. Since there will be infinitely many left over anyway, this way of thinking about the nature of meaning and content makes them out to be somewhat indeterminate.

One important constraint is that agents should be interpreted as rational. We are seeking to confirm a theory of the meanings of all the sentences of a speaker's language. Davidson also holds that we can treat the propositional attitudes as sentential attitudes: believing that *p* is a matter of holding a sentence true that means that *p*, desiring that *p* is a matter of desiring a sentence true that means that *p*. The theory of meaning thus interprets all the speaker's attitudes as well as her utterances. Since our attitudes are connected with each other and with perception and action in the rich and familiar patterns of practical, epistemic and theoretical reason, we know that a better interpretation is one that will display the agent as more rational in what she does and what she thinks.

The conclusion is that the more sense we can make of an agent the better the interpretation, and the more ground we have for thinking a theory of truth is confirmed for the agent's language. So if we can confirm a theory of truth for an agent the agent is more or less rational.

The skeptic protests that the conclusion so far is conditional. If the agent is not rational at all then no theory will be more confirmed than any other.¹¹

Clearly if meaning must be accessible to outside investigators then there must be some route to an agent's meanings using observations of her interactions with what is going on. Similarly, since grossly irrational agents would be indistinguishable from things which are not agents at all, if thought must be accessible to outside investigators then agents must be fairly rational. The trouble is the strong claim that meaning and thought content must be accessible to outside investigators in the way sketched.

Davidson offers a couple of considerations favoring this strong claim.¹² We are

¹¹The Cartesian skeptic will have a similar worry about the needed assumption that we can identify a truth-relevant attitude toward sentences and succeed in untangling the ways the person has gone wrong about what is around her.

¹²Perhaps the strongest reason is that meaning is public and accessible: it is our medium of communicative exchange, and so it must be accessible to outside investigators. Unfortunately this just begs the question against the Skeptic. She thinks Descartes' First Meditation shows that meaning isn't necessarily public.

1.6. (2) FROM THE THEORY OF MEANING TO CONTENT RATIONALISM 15

entitled to assume that most of an agent's beliefs are true because agreement and disagreement only make sense against a background of widespread agreement; similarly,

To see too much unreason on the part of others is simply to undermine our ability to understand what it is that they are so unreasonable about. ("Belief and the Basis of Meaning", p. 153)

Undoubtedly this is correct as a remark about what we do when we make sense of others, and about how we do it. But it will not persuade the skeptic. Suppose seeing too much unreason undermines my ability to understand what another is thinking about. Then I fail to understand what she is thinking about. It cannot straightforwardly follow that she isn't thinking about *anything*. Davidson claims that it does:

The methodological advice to interpret in a way that optimizes agreement should not be conceived as resting on a charitable assumption about human intelligence that might turn out to be false. If we cannot find a way to interpret the utterances of a creature as revealing a set of beliefs largely consistent and true by our own standards, we have no reason to count that creature as rational, as having beliefs, or as saying anything. ("Radical interpretation", p.137)

The skeptic has two complaints. First, why should we hold that if we cannot find a way of interpreting a creature we have no reason to count it as having beliefs at all? Second, suppose we don't have reason to count a creature as having beliefs at all. There are two ways to gloss this lack. The stronger is to say we do not have and there could not be reason to count it as having beliefs. The weaker is simply that we do not have reason to count it so. It does not follow from the weaker gloss that it is not true that the creature has beliefs. It appears to follow from the stronger gloss, given a principle that what we cannot have reason to believe (about another) cannot be a fact. But it is unlikely that any principle strong enough to do the work here required is defensible.

1.7 (3) Axioms of Decision Theory and Constitutive Concepts

Davidson's strongest reasons for holding that content rationalism is true come from considerations about decision theory. In this section I will describe one clear but inconsequential argument. In section 9 I'll describe a more persuasive argument.

There is a tight connection between decision theory and a theory of the propositional attitudes. Decision theory is intended to be a mathematical and precise version of the relations we find among beliefs, desires and rational decision. It describes an elegant structure that captures central and essential features of the norms by which we live. Hence it gives a structure that anything that deserves the name of belief should satisfy.

Decision theory assigns numbers to valuations and states of belief by assuming that there is some empirical test that gives us insight into these states, and that there is structure in the results of the test. Jeffrey's theory takes preferences between propositions as the basic fact. If certain things about the field of preferences are true then it can be proved that relations among preferences can be represented in a pair of real-valued functions *prob* and *des*. Stating what needs to be true about the field of preferences is solving the "existence" problem: formulating the axioms of decision theory. Jeffrey's first axiom is that the relation \succsim , "is preferred at least as much as," is transitive and connected (p.145); that is, for all propositions A, B and C, if $A \succsim B$ and $B \succsim C$, then $A \succsim C$, and furthermore for all propositions A and B, either $A \succsim B$ or $B \succsim A$. This axiom captures an ideal of rationality: our preferences should be transitive. If our preferences are not transitive then a Dutch book can be made against us.

If some agent is such that there is a *prob* function that assigns a number to a single proposition entertained by that agent, and the *prob* function is derived from preferences in the way Jeffrey describes, then *all* her preferences are rational in the way demanded by this first axiom. This sort of consequence follows from the attribution of any one probability or valuation in any version of decision theory. Davidson points to Ramsey's theory: if Ramsey's theory applies at all then there is an ethically neutral proposition for that agent. If there is an ethically neutral proposition for the agent then a great deal is true about the gambles between which the agent is indifferent:

if it is ever reasonable to assert, for example, that one event has a

higher subjective probability than another for a given person, then there must be good reason to believe that a very strong theory is true rather than false. (“Psychology as Philosophy”, p.236)

The sense Davidson attaches to the claim that the norms of rationality are constitutive of the propositional attitudes is that if a term from the unified propositional attitude theory applies to an agent then the norms of rationality must as well, since the theory is a theory of things which are connected with one another in this way.

A weak skeptical stance to take to this argument is to emphasize doubts about the relations between decision theory and propositional attitude theory. For instance all decision theories have what might be called the “too many sandwiches” problem.¹³ For the field of preferences to determine real-valued *prob* and *des* functions there must be infinitely many preferences between “nested” choices. Ramsey’s theory, for instance, requires being able to find gambles exactly equidistant in value between any two others. It’s hard to imagine that any actual agent is like this.

This is not, however, a particularly good response, since there are relatively natural ways to count an agent’s preferences as satisfying the axioms in virtue of approximating the ideal. It’s easy to fill in the numbers provided the agent’s actual preferences give us enough to get started.

A stronger skeptical stance notes that once again the conclusion is conditional. If one of the notions of decision theory applies, then many do, and the agent’s preferences will be rational, or will approximate well to rationality, in some appropriate way. But what is needed is an argument that there cannot be any such thing as an agent’s preferences that don’t even come close to satisfying the axioms of some defensible decision theory.

There are other, and better, arguments from decision theory to content rationalism than this one, and we will return to them in section 9.

1.8 (4) From Action Theory to Content Rationalism

Plausible accounts of the nature of action and the relation between reasons and action support content rationalism. An action is a bodily movement which is caused

¹³In honor of the illustrative examples that describe preferences toward many kinds of sandwiches; see Jeffrey, pp.44-58.

in the right way and rationalized by reasons. To be rationalized by reasons requires that the bodily movement have some aspect such that the agent could have reasoned from her beliefs and desires to the conclusion that a bodily movement with this aspect is desirable or valuable in some other way. It is not enough that there should be *some* path of reasoning that leads from certain of the agent's attitudes to what she does. There is such a path between any primary reason and any action. Rather, the agent must have actually followed such a path, or the path was just the one she would have taken had she given it a moment's thought. Finally it's not enough that there should be a suitable path from one single primary reason to the action; the path must take into account pretty much everything the agent thinks and wants.

Here's an example to make this concrete. A woman buys some broccoli at the grocery store. Her primary reason for the act of buying the broccoli is that she wants some broccoli and she thinks buying some broccoli is a good way to get some. She undertook the action knowing that she had enough cash to pay for the broccoli, that the store was likely to have some, that she didn't have some at home, that she wouldn't rather have snapper and rice for dinner, and so forth. Undoubtedly very little of all this "went through her mind" when she acted, but it is all clearly relevant to what she did.

If any event is an action it is caused by a vast and vaguely bounded collection of beliefs and desires and the action is rational in light of that collection. It follows that if an agent acts then content rationalism is true of the attitudes behind the action. What I've said leaves open whether this is *all* the agent's attitudes.

The same result holds for attitudes behind irrational actions. Irrational actions are not events which are random with respect to what an agent thinks. Suppose I smoke a cigarette, having quit years ago for all the standard reasons. I act against my best judgment, but I have reason for what I do. I anticipate the pleasure of smoking, I know that this action will not immediately cause results so intolerable as to outweigh this pleasure (I'm not on an airplane in California, for instance) and so forth. Irrational actions are thus backed by an equally vast and unbounded collection of attitudes, and the action is rational in their light. The only difference is that somehow that collection fails to include my best judgement.

The same result holds for attitudes behind the existence of each intention. An intention may be thought of as a kind of desire or valuation of an action; it is special in being the valuation on which the agent has decided to act. Hence an intention is something which is the outcome of a rational chain of causation from a large number of the rest of the agent's attitudes.

A similar result holds for many emotions. Pride is that emotion caused by appreciation that one has done well (Davidson, 1980e). Appreciation requires a rational appraisal of what one has done that involves a great deal of what one thinks. Similarly we feel anger for reasons, and these reasons can be rational or irrational.

The skeptic's response is again that the conclusion is conditional. Any belief or desire which stands in these relations with actions or intentions or cognitive emotions will be such that content rationalism is true of it. That does not show there could not be a belief or desire which does not stand in these relations. There are two kinds of case to consider. First, cognizers are possible who never act, form intentions or feel emotions. We have a strong sense that there is no such possibility, but I think we have no logical ground for rejecting it. Second, it does not seem to follow even given a normal agent that all her attitudes are behind an action. Davidson argues (Davidson, 1982) that irrationality always demands a very special reason explanation, one that involves thinking of the set of the agent's attitudes as segmented, one part representing the agent's best judgment, the other the worse judgment that somehow overcomes the force of the better. But it is not clear that this special explanation is always required. My best judgment may be left out of what yields my action in a particular case through fatigue or injury or confusion. Hence it can happen that occasionally attitudes are simply left out of deliberation, caused to be left to one side but for no reason.

We might claim in defense of content rationalism that all propositional attitudes are mental acts and hence stand in the same sort of relations to prior attitudes and perceptions as do proper actions. But our thoughts are not actions. We never have direct choice over the next thought. We have our thoughts and we undergo changes in what we think which are (if we are careful and fortunate) rational changes.¹⁴

Davidson, in responding to Hempel's view of the explanation of action, notes that "the looniest action has its reason." ("Hempel on Explaining Action", p.237) We might say, just as every action has its reason, however loony, every reason has its reason, however loony. I think this is a central assumption of psychotherapy: reasons don't just happen to us. There is always something around that provides

¹⁴This does not mean our thoughts are completely out of our control. One aspect of rationality is our capacity to alter habitual transitions of thought: we can, for instance, learn and practice epistemic skills. As we'll see in Chapter 4 on Dretske, if we aren't skeptics and aim at an account of content, this constraint is important: nothing merits the title 'belief' unless it is subject to this sort of higher order control.

a rational ground for what we think. And again it is hard to know what to think of the idea of a thought that is neither preceded nor followed by any rationally related thought. It might be that the agent who had the thought wouldn't know that she had it, since we might hold that the kind of relation that holds between thoughts in self-knowledge is a species of rational connection.

The skeptic still has responses. Even where therapy works and an agent "uncovers" reasons for her unhappy state, and perhaps manages to do something about them, we do not have to accept the claim that the reasons that are "uncovered" are reasons that the agent had. We can always find reasons for holding some belief; having found them, we may accept them and act on their consequences.

None of these considerations on action appears to rule out the possibility of an event of coming to believe something could just occur, without any of the sorts of connections that content rationalism requires. We can, it seems, think of events of coming to believe something or coming to desire something as having their propositional attitude properties intrinsically.

Finally, I doubt we can make an argument for content rationalism by appeal to the right analysis of self-knowledge. Perhaps that analysis requires that where there is self-knowledge an agent is caused to believe that she has a belief by that belief, and caused in a way that is rational. Still, we can conceive of the existence of a belief by itself that isn't known by its owner by way of self-knowledge. The existence of the belief wouldn't thereby be unknowable; perhaps we could come to know about it through some other means.

1.9 (5) Testing Decision Theory and Content Rationalism

The strongest argument for content rationalism comes from considering the empirical status of decision theory. I'll start by reviewing three reasons for thinking that decision theory is an empirical theory, and then sketch the argument that content rationalism is true that stems from the prospects for testing whether it is true.

First reason: Hempel was led, by the possibility that people can have the same reasons yet do different things, to think that reason explanation must appeal to more than the reasons we cite (Hempel, 1974). He proposed that there must be a lawlike generalization about what people do given certain reasons together with a statement of how rational they are. Every reason explanation then has the form of

1.9. (5) TESTING DECISION THEORY AND CONTENT RATIONALISM 21

a deduction from a law: the antecedent conditions are the belief, the desire, and the degree to which the person is rational (perhaps with respect to those very reasons), and the law is a quantitative statement of what people do with their reasons when they are rational to that degree. If Hempel is right then we must be able to make sense of the idea of testing how rational people are, and even perhaps how rational they are with respect to certain reasons.

Second reason: Ramsey's and Jeffrey's decision theories require very strong conditions on preferences for representation theorems. Both require that preference should be transitive. But it is *prima facie* unlikely that our preferences are transitive. The question appears to be empirical; we can imagine asking people about their preferences and discovering whether they are rational in this sense. Of course we can do better than imagine, since an enormous amount of energy has been expended precisely on determining to what extent and in what respects people typically fail of decision-theoretic rationality (Kahneman et al., 1982).

Third reason: Decision theory is meant to be a mathematically precise version of something we are doing every moment of our lives: making decisions and acting in the face of less than perfect certainty about what will happen. Psychology includes the empirical study of the ways we acquire and change our beliefs and desires and act on them. Psychology thus appears to include the empirical study of human approximation to decision theory. The generalizations it yields should look something like the generalizations of decision theory. Continuing in this line of rumination it's easy to come to the following conclusion: decision theory is best thought of as an empirical theory which captures the ways that people and their propositional attitudes behave. For instance Fodor finds that Shakespeare's Hermia (in *Midsummer Night's Dream*), in thinking about what Demetrius says about Lysander and about what Lysander might have done, must be working with the causal generalization that

If x wants that P and x believes that -P unless Q, and x believes that it is within his power to bring it about that Q, then *ceteris paribus* x tries to bring it about that Q (Psychosemantics, p.13)

Fodor holds that this generalization is more or less true and that it is a contingent, empirical generalization; that is the force, it seems, of calling it a *causal* generalization. As I noted above Fodor holds that the conditions on being an intentional system are stronger than the conditions on being something that has states with content. Intentional systems have states with semantic content which are furthermore related in a certain way. Being a belief is being a state with content that

stands in a special relation to perceptual 'inputs', behavioral 'outputs' and other contentful states. Which special relation?

suppose that a belief state is by definition one that causally interacts with desires and actions in the way that your favorite decision theory specifies (p.69)

The distinction, in the set of all items that have the content that P, between being a belief and not being a belief, is an empirical distinction and a matter of a certain sort of empirical complexity. Hence according to Fodor decision theory states a collection of empirical laws which are such that if something satisfies them it is a believer.

Here then are some reasons for thinking decision theory is an empirical theory. If it is then we should be able to test whether it is true.

Much empirical investigation has been performed in an effort to determine whether decision theory is true. Davidson and Tversky draw conceptual conclusions from the empirical work. Consider the simplest sort of example. A subject expresses a preference for a hamburger over tuna salad, a preference for \$10.00 over a hamburger, and a preference for tuna salad over \$10.00. Her preferences appear not to be transitive. Shall we conclude that she is not rational? Not at all. She prefers a hamburger over tuna salad, given *that* choice, as far as the taste goes; the money over the hamburger, given *that* choice, for its buying power; and tuna salad over money because her cat wants it. The general trouble is that whenever we find what appear to be violations of the axioms of decision theory we can usually also find something that explains the breakdown in purely rational terms. The trick is to find what propositions the agent is considering in expressing her preferences.

Notice also that even a run of complex failures of perfect rationality hardly shows that the subjects are even mildly irrational. These subjects are rational enough to get themselves involved in the experiments, to understand what they are to do, and to produce sensible reports of their responses to the experimental situations. And the more complex the failures the harder it is to draw any consequence, since agents must be quite rational before they are even in a position to fail in such tasks (Kahneman and Tversky, 1982).

Davidson concludes that all the tests of decision theory can just as well be taken to measure the adequacy of the theories' interpretations of the term 'prefers' as they can be taken to test decision theory. He concludes that decision theory has almost no empirical content:

1.9. (5) TESTING DECISION THEORY AND CONTENT RATIONALISM 23

I am skeptical that we have a clear idea what would, or should, show that decision theory is false . . . the theory . . . is so powerful and simple, and so constitutive of concepts assumed by further satisfactory theory . . . that we must strain to fit our findings, or our interpretations, to preserve the theory. (“Hempel on Explaining Action”, p.273)

We have a quasi-empirical argument to a quasi-conceptual conclusion.¹⁵ Since we can make no clear sense of the idea of testing decision theory, the axioms of decision theory are true (“constitutively true”) of anything to which the concepts of decision theory apply. The parallel argument for psychology is that, since we can make no clear sense of the idea of testing how rational a person is, the norms of rationality are true (or nearly true) of anything to which the propositional attitude concepts apply.

The skeptic has at least three things to say in response.

First, she will emphasize the differences between the point about decision theory and the point about propositional attitudes. Decision theory is an idealization of ordinary practice; it aims to interpret a single predicate in a way that yields functions from propositions (or sentences) to numbers. There is no representation theorem for psychology that shows we can measure propositional attitudes with anything so tractable as numbers, and there are grounds to think there couldn't be one.

Second, in thinking about ourselves we apply a scheme of explanation with constitutive elements, but we have a large degree of tolerance for deviation from the ideal. For instance we can reinterpret an apparently intransitive triad of preferences so as to understand the person as rational. We can imagine further tests by which to test the reinterpretation. And, I think, we can imagine them failing. An agent can be understood as irrational, and the irrationality can be mild or severe.

Third, she will protest that Davidson's skepticism results from too narrow a view on the possible data. The tests of decision theory that have been tried all work with behavioral data of greater or lesser subtlety. Perhaps it is true that there just could not be a genuinely plausible test of decision theory that uses such data. From that fact it wouldn't follow that decision theory has no empirical content

¹⁵I think it is easy to miss the force of this argument. People say Davidson's argument for anomalism is a conditional we can traverse in either direction: Davidson infers that psychology is not a science from a commitment to thinking of belief in a certain way, but we can just as well infer that we shouldn't think of meaning in that way. (See for instance (Johnston, 1985, 422–426)). But Davidson's choice of direction comes precisely from the effort to get the science off the ground and the resulting pattern of failure.

or that it is analytically true (or something like it). All that follows is that the empirical content is not exclusively behavioral, or that we can't find the cases that show why it is not even something like analytically true.

The naturalistically inclined skeptic thinks it is not at all hard to imagine a test which uses other than behavioral data. She claims that what content thoughts have is a function of the relations between internal states of the person and things outside, where these relations may be underdetermined by behavioral events. The *sui generis* theorist will say that we have always known that inference to what someone thinks from behavioral data is an iffy business at best; she may go on to say that the only good way to know of the existence of a thought is to have it.

1.10 Conditional Arguments

All the arguments I've surveyed fail in the same way to convince the skeptic. Each has the form, if we think of meaning and content in a certain way then content rationalism is true. The skeptic responds, why should we think of meaning and content in that way? In this section I consider reasons to think the skeptic doesn't realize the cost of denying the antecedent.

Much of Davidson's writing on meaning assumes certain basic facts about meaning. Meaning is basic to communication; it is what we grasp in understanding others. We understand others by going on behavioral evidence: by looking at what they say and do. We should, it seems, understand meaning and thought content as available on this basis.

Thinking of meaning and thought content as essentially bound up with this sort of evidence has several philosophical advantages, besides placing meaning and thought in what seems to be an optimally appropriate conceptual context. Epistemological skepticism gets a relatively clear response: since anything which is a believer is rational and a believer of truths, I can know of myself (since I know I am a believer) that most of my beliefs are true. The problem of other minds seems to evaporate, since thought content just is what can be gathered by interpreting an agent's attitudes toward sentences.

The trouble with this strategy is that it makes the connection between meaning, content and a certain sort of theory too strong. The solution to the other minds problem seems to be bought at too high a price: the solution entails that it is just not possible for someone to have thoughts forever beyond the reach of a theory based on behavior. Certainly if we think of someone else and ask why we should

ever believe this possibility obtains (and ask the parallel question what sense we can make of the possibility) it is hard to imagine any satisfying answer; but I think we have a clear understanding of what it would be in one's own case to have a belief like this.

We can try for a stronger response to skeptical problems about knowledge of other minds. Akeel Bilgrami has suggested that we *cannot* ensure the public availability of meaning except on condition of accepting an externalist position on meaning (Bilgrami, 1989). Analogously perhaps we *cannot* ensure the public availability of meaning except on condition of accepting content rationalism, since a wildly irrational agent would be one we could not understand in the way we do understand agents. Perhaps this is right, but Bilgrami has not offered any considerations that will compel the skeptic. We may consistently believe that thought and meaning are publicly available and that externalism is wrong or that content rationalism is false.

1.11 Content Rationalism and Naturalism

My aim in discussing Davidson's arguments for content rationalism has been to set the stage for what I'll be doing in chapters 3 to 6. I've suggested that the skeptical responses to these arguments flow from two sources. The *sui generis* theorist is hyper-Realist about content; to him it doesn't matter if what he says is no more than logically consistent. The naturalistic theories suppose that there is some further ground for testing decision theory than the ones that have been investigated by psychologists. The naturalistic theorist aims to be a slightly milder sort of realist about content; he is worried that if we cannot give a causal account of content in non-question-begging terms then somehow content will not be real. If his project succeeds he may be able to show us that content rationalism is an *empirical* thesis, true in virtue of how beliefs actually interact rather than in virtue of the concepts of belief, desire and rationality.

I'll be arguing that the naturalistic theorist has overestimated the payoff of a causal theory of content. It is important and interesting to show how the concepts of content and causation are related. But I do not believe the causal theories really show that content rationalism is an empirical thesis. Each theory has to use some key notion in stating the causal account of content that begs the question. I am inclined to treat this systematic failure as evidence that the causal theories do not genuinely provide further ground for testing decision theory.

My argument for content rationalism is in a sense a naturalistic one: it is an argument from the systematic failing of naturalistic theories to content rationalism. All the arguments I've just surveyed are also naturalistic in spirit; they say that content rationalism follows from the best naturalistic accounts of content we know how to give. But we might wonder whether naturalism is the price we must pay for an argument for content rationalism.

The *sui generis* irrationalist is not, of course, going to give an argument for content rationalism. I cannot think what might convince such a theorist. His position is consistent and unintuitive and he will accept the unintuitiveness. The *sui generis* rationalist refuses to give an account of content; he may as well refuse to give us grounds to think content rationalism is true. I don't think there is anything that would rationally compel such a theorist to try to say something.

Can a *sui generis* rationalist give an argument for content rationalism? I suspect he cannot, but I have no conclusive argument to show this. Perhaps he will appeal to one of the arguments I've surveyed, especially the one based on the difficulties in making sense of a test of rationality. The trouble will come in defending some appropriate connection between what content is and the sorts of possibilities we find intelligible. Surely there are connections strong enough to do the job, but they violate the *sui generis* theorist's inclination to say that content is the sort of thing of which no account can be given.

If this is right I suspect the reason it is right is connected with problems with the distinction between analytic and synthetic truths.¹⁶ Content rationalism is not even analytically true, but it appears that if it is true it is more than a contingent truth. Necessity of this sort is best explained by appeal to the way the concepts are connected within the theories we accept or the frameworks we currently think will inform any acceptable theory. Hence an argument to show that this sort of necessity obtains has to appeal to considerations about theories of the items in question.

Putting this point this way leaves an avenue open for the *sui generis* theorist. Perhaps when we are finally satisfied that we understand meaning and thought we will have a theory of these items that does include content rationalism but which treats meaning properties as primitive and hence not determined by other facts in the way the theories I consider claim. I remain agnostic on the prospects for such a theory.

¹⁶Claudine Verheggen suggested this explanation to me.

1.12 Conceptual Connections without Analyticity

It might be objected that the strategy I pursue for showing content rationalism true does not establish my conclusion. Perhaps naturalism is just not the right way to think about content. Even if it is, maybe there is a genuinely explanatory reductive account of content right around the corner.

There is more to my strategy than this assessment suggests. My overall argument has two aspects. First, the arguments I've examined so far are negative in character; they say that we can't make sense of the idea of a belief off by itself or in a highly irrational cognitive set. Second, the arguments I will be presenting show that a certain positive project is in serious trouble; none of the attempts to provide a genuine alternative to content rationalism works.

The skeptic cannot say very much in response about the positive projects. She has temporarily run out of proposals and neither she nor I can fruitfully speculate on what will happen as she attempts to generate more. But she may still balk at the negative claim. What metaphysical conclusion ever follows from our inability to make sense of a certain possibility? Maybe it is actual, despite the limitations on our comprehension. Perhaps we will come to be able to make sense of it. We aren't confronted with a flat inconsistency; the conceptual connections are asserted to be constitutive, not logical or analytic.

The skeptic is right that the connection has rather less modal force than a logical or analytic connection. We should be careful that the claims we make reflect this modal force. We are not entitled to claim that content rationalism could not possibly be false, or that the reasons we've seen show that we could not possibly make sense of a cognitive set that is not rational. All we are entitled to claim is that we do find it so difficult to make sense of the purported cases that we are left unsure whether they are real possibilities.

If we are only entitled to make such a weak claim, does the skeptic win out? I doubt it. The choice is between an intelligible theory which fits our experience as well as any alternative, and a theory which is logically consistent but at present unintelligible.

I think the situation with content rationalism is not unlike the situation with other philosophical projects of explication of conceptual connections. Take for example Davidson's principle of the nomological character of causation, the thesis that if a singular causal claim is true then the events involved must have aspects which make them instances of a strict law. Anscombe asks why we should believe

this, and finds no answer. But not having compelling reason to believe it is not the same as having reason not to believe it. I doubt anyone has ever described a clear case where the principle fails. Now we might accept the possibility as part of our metaphysics if we had some good reason to do so.¹⁷ But we are then accepting a possibility we do not understand. There is nothing logically improper in doing this. Clearly we must be ready to accept some actual facts which we don't understand. But I think we should aim to accept such facts as infrequently as possible. If there is reason, albeit inconclusive, to accept a position, and the alternatives are not plausible, we should accept the position.

The skeptic would have a better time of it if there were some explanatory point or force to the denial of content rationalism. I can think of two, both involving a science of psychology. First, Fodor and others worry that if content rationalism is true then there will not be a proper science of psychology. As I've said, and as I'll argue in some more detail below in chapter 7, I don't think this is a serious problem. Nothing, for instance, about the reality of our mental lives would follow from the unscientific character of psychology.

Second, many have thought that if there is no proper science of psychology—if the anomalism of the mental is true—then our mental lives will be causally inert. One might attempt to build an argument to the effect that the stronger the conceptual connections among properties in a set the less ground we have for claiming causal connections among the properties. I think there is some force to this idea; I'll be returning to it at length in Chapter 8. But it won't, I think, help the skeptic much, because the skeptical challenges can be met; there is simply no good argument to causal irrelevance based on anomalism or conceptual connections.

Since I can think of no other point to the denial of content rationalism I think we should accept it.¹⁸

¹⁷For instance David Armstrong suggests this possibility provides an “independent fix” on the notion of necessitation needed to explicate the universals theory of natural law.

¹⁸I'd like to thank the participants of a Chat on the topic of this chapter at UC Berkeley in December 1988.

Chapter 2

Kitty Thoughts

2.1 Introduction

Can my cat Grushenka think? Here are some of the things she does. She seeks me out. She looks me in the eye. She hollers at me. She looks at the string. She crouches when I reach for it, pounces when I pick it up. She chases it. I stop; she hollers at me again.

When I come home she runs down the stairs, smells my shoes, rubs against my legs. Another day I arrive wearing a motorcycle helmet. She gets halfway down the stairs, glimpses the helmeted figure, and runs back up the stairs to hide.

She's doing something much like acting on reasons, expressing desires, indicating objects, and making mistakes. Is she *actually* acting, thinking, erring?

(Davidson, 1984c, 1985b) argue that speech is necessary for thought (I'll call these 'TT' and 'RA'). Grushenka doesn't speak. If Davidson is right, she literally does not have any thoughts at all.

The theories of content we will be examining in Chapters 3–6 below apply most naturally to very simple creatures, like bacteria, like frogs, like cats. These theories all assume that these simple creatures literally have content-bearing states. If all attributions of content to non-linguals are metaphorical then these theories are doomed.

In this Chapter I consider the details of Davidson's arguments. I think Davidson does not produce a convincing case that thought requires speech, and hence I think that attributions of content to non-linguals are sometimes literally true.

Davidson's views on animal thought have shifted somewhat since "Thought and Talk" was published. In (Davidson, 1985c) he writes,

My suggestion comes to this: we will in any case continue to talk as if animals have propositional attitudes. We can do so with good conscience if we keep track as best we can of the level of significance of such talk. (p.252)

I take essentially this position in what follows. We can describe the minimum structure required for representational content. The content of many things that satisfy this minimum is quite different from the content of any of our thought, but it is still literally content.

Certainly there remain many distinctions to be made: we may decide against attributing *beliefs* to my cat (certainly to bacteria) but continue to attribute representational content. Hence there are at least two conceptual boundaries that we need to mark: one to indicate what is required for the simplest possible representation, and one to mark the difference between representation and belief. In this thesis I work out the beginnings of a way to mark the first boundary; I will have nothing to say about the second.

2.2 Exegetical Preliminary

TT covers a great deal of ground on the way to its conclusion that "a creature cannot have thoughts unless it is an interpreter of the speech of another" (p.157). Should we consider all this to be one argument, so that the discussion of each topic yields another premise? Or is the argument made in the last page or so (which yields the conclusion) meant to be relatively independent?

The structure of RA suggests a way to read TT, even if it is not exactly the way Davidson thought of it when he was writing TT. RA makes two preliminary points against Malcolm's claim that dogs have beliefs: first, attributions of content to dogs are not intensional, while attributions of propositional attitudes clearly are intensional; second, any one thought requires a "rich supply" of related beliefs. But Davidson notes that these points do not demonstrate that animal thought is not possible: "Indeed, what these considerations suggest is only that there probably can't be much thought without language." (p.477) He follows these preliminary points with an expanded and somewhat revised version of the argument of the last page of TT. I suggest we read TT as presenting several interrelated views of

the relation of thought and talk, but the argument of the last page is essentially independent of the rest. I'll call this argument and its fellow in RA the "central arguments."

Hence there are roughly three distinct arguments to consider: one about intensionality, one about the holistic character of thought, and one about what is needed for thought that only language can supply.

I agree with Davidson that intensionality is a crucial issue, and I agree with him that it does not decide the question whether animals can think. In section 3 below I argue that the issue of intensionality is something of a red herring: if a creature has mental representations (in a sense I will describe) then ascriptions of thought to it *are* intensional and may be *as* intensional as attributions of propositional attitudes.

I will not discuss the argument about holism here. It is true that very many thoughts are such that something cannot have just that thought unless it has many others. But I think this is not a universal phenomenon: some thoughts are such that a creature can have that thought and have no other. I discuss this point at some length in Chapter 6 below, section 8.

In section 4 I will describe the central arguments. Each has two premises; in section 5 I examine the claim that if something has a belief it has the concept of belief, and in section 6 the claim that if something has the concept of belief then it communicates with another.

2.3 Animal Thought and Intensionality

Attributions of thought are intensional; if what we say about the thought of dogs, cats, frogs and bacteria is not intensional then we are not attributing thought.

Davidson has two kinds of arguments to show that attitude attributions to animals are not intensional. Davidson concedes that these arguments are not specially conclusive, but I want to demonstrate why they do not make their point.

The first argument appears both in TT (p.163) and RA (p.474). Malcolm urges that the dog thinks the cat went up the oak tree. But does the dog think the cat went up the same tree it went up yesterday? that Grushenka went up the tree? that the 7 year old domestic shorthair that Tony owns went up the tree? We hardly know where to begin in thinking about answering these questions. Dodging them by claiming the descriptions and predicates we offer are in transparent position

doesn't help, since if a *de re* attribution of thought is true then a *de dicto* one must be as well. (Some, like Burge and Dretske, will urge that this is incorrect; but *some* kind of *de dicto* attribution must back up a fully transparent one, even if it is only one that attributes the predicate.) Since we can't decide which descriptions matter and which do not, it looks as though it doesn't matter how we describe the doggy thought.

The last step in this argument is a mistake. If dogs have thoughts they are clearly not much like ours. Their "form of life"—their sensory capacities, cognitive capacities, motor capacities, reproductive cycles, dietary needs—is extremely different from ours. We should expect that if they have concepts they will not be much like our own. The trouble is not that attributions of doggy thoughts are semantically transparent, but rather that they are sensitive to different substitutions than similar attributions to persons would be. My cat's behavior shows this: she runs from the helmeted figure, and she wouldn't run from me (if I weren't wearing the helmet). We don't have an articulated set of concept-terms for dogs. Even if we did it would remain a curiosity for ordinary explanatory purposes (although absolutely fascinating on other grounds: it would say what it is like to be a dog), since the much more powerful scheme of concepts we employ for ourselves works just as well for dogs. We see a similar phenomenon at work in what we say about each other. We may know that Jones has a very idiosyncratic understanding of the relations of nations, so idiosyncratic as to show, for instance, that what she calls 'détente' is not détente; yet we might for all that report one of her beliefs using the term 'détente'. She has *some* concept, one for which we have no simple term; so we use another which is close enough for explanatory purposes.

The second argument occurs only in TT. One way to investigate the nature of a kind of thing is to examine the theories that talk about that kind of thing, in particular for invariants in the structure of claims about the kind of thing. If a theory preserves its explanatory power under many more transformations than the theory of belief-desire psychology permits, then it is not a theory of belief. Davidson traces a series of refinements made to a very weak specification of teleological explanations of behavior, and suggests (but does not claim) that a theory that does not interpret speech is essentially different in respect of invariance than propositional attitude psychology.¹

¹I am somewhat hesitant whether this really is the structure of the argument. What Davidson says satisfies this form up to the concluding claim; but the form of the conclusion is not that a theory permits too much variation, but rather that if it doesn't interpret speech it fails to capture certain facts relevant to the success of its explanations.

I will summarize the series of refinements, then criticize the suggestion by showing how the needed invariance can be provided without speech.

We explain our actions by talking about our reasons. We say someone wanted a certain outcome and believed that by acting thus she would obtain that outcome, and that's why she acted thus. A solitary reason attribution, however, explains nothing, since reasons are explanatory only given strong background assumptions, as for instance that the agent didn't have a better reason to do something else. So reason explanation works by citing, implicitly or explicitly, a whole pattern of reasons that cohere in a certain way and which together explain a series of actions.

This constraint of pattern is a considerable improvement on solitary attributions, but it still leaves room for unlimited alteration in the descriptions we offer in explanation of an action. We can show this by describing a mathematical analogy with no more structure than we have introduced so far. Suppose all actions are ordered, so that we can assign an integer to each one based on its position in the ordering. Suppose what it takes to *explain* an action is two real numbers which when multiplied together yield the integer. Of course for any integer there are infinitely many such pairs. Now suppose the trick is to explain a *series* of actions, and a constraint is that the number-reasons for earlier actions can never exceed the number-reasons for later actions. This would constrain the possible attributions of number-reasons, but the possibilities are still infinite.

Decision theory finds more structure in the field of beliefs and desires. Davidson considers Ramsey's theory. Ramsey showed how both utility and subjective probability can be determined on the basis of information about preferences among a variety of gambles. This extended preference ranking determines a probability function uniquely and a collection of equivalent utility functions. They are equivalent up to a linear transformation, in Ramsey's theory; this entails that given one probability function each utility function yields the same ranking of *acts*. This in turn means that we can explain an act by showing that it was the act that came out on top, and by citing the reasons the agent had for that act. It's still true that if one reason attribution is explanatory then there are infinitely many others, but we now have a principled way to characterize the multiplicity. Furthermore, the alternatives are generated by altering the entire attribution, rather than single attributions piecemeal.

Where do the preference rankings come from? In actual practice, claims that someone has a preference ranking are supported by data about linguistic behaviour. Information about choices between gambles isn't enough, since the bare fact of a choice never determines what it was about the object chosen that was

chosen. Suppose Jones pulls a \$10.00 bill from her wallet and pays for a book. She chose one object among others. What was she thinking? Did she choose the leftmost thing in her wallet? The thing with Hamilton's picture on it? Her favorite bill? There's no way to tell *simply* from seeing what she did.

To fix her preferences uniquely we should have to interpret what she says about them. Without speech "the evidence will not be adequate to justify the fine distinctions we are used to making in the attribution of thoughts" (TT, p.164). Evidence about choice behavior alone also doesn't seem sufficient to settle other fine distinctions in the thoughts we have. It seems hard to imagine how to "distinguish universal thoughts from conjunctions of thoughts . . . how to attribute conditional thoughts, or thoughts with, so to speak, mixed quantification. . . ." (TT, p.164).

Interpretation of speech solves the problem of making extremely fine distinctions by two means. First, it settles the interpretation of non-logical terms by concentrating on general truths about their relations with things. Actual generality isn't needed, since sometimes we can gather evidence from a single use about how a term would be used in other circumstances. Second, interpretation finds a repeatable structure in the language (e.g., phrases like 'if', 'all', 'she') that can be used to generate the interpretations of extremely complex constructions.

The trouble I see for this kind of argument is that there is a way to solve these problems for non-linguals: *mental representations* (internal structures with the sort of structural complexity that language has) provide the resources.

What follows is a brief exposition of the sort of theory I have in mind. First I'll give some reason to think that the problem for the determinateness of the content of terms can be solved, and then describe where logical structure fits in. What I say here is somewhat sketchy and incomplete; we will be considering the details of much more articulated theories like this in the next 4 Chapters.

We start by supposing a creature C has a certain *type* of state S. It's important that S be a non-intentional type, since the project is to give a theory of content that doesn't assume categorizations of things that depend on their content. S's are caused by various things, among them F things. When S's occur, the creature is caused by them (and other internal and external conditions) to execute certain movements. Sometimes these movements result in some benefit to the creature that is contingent on the F things being around, and contingent on their *being* F. This benefit, over time, controls the production of the state-type S. For instance a history of benefit contingent on the production of the state when an F thing is around might increase the reliability of the connection between F things and

instances of the state type S, so that the likelihood that there is no F around when an instance of S occurs goes down. If the state type S has this relation in the life of the creature to the environmental type F, then an occurrence of S, when there is no F around, is wrong or incorrect.

Instances of state type S are representations. Of course we must be very careful what we mean when we say this. S has no structure, so there is no sense in which it refers to F, and there is no sense in which S is a concept, if we mean by a concept something which can be predicated of an object. The most natural way to express its representational content is with some such phrase as, there's an F, or, F here now, but we still have the problems we had with Malcolm's dog, now far worse, since we have no reason to think that creature C even has the internal resources for predication of a demonstratively indicated object or time or place.

Certainly these are serious problems, but I don't see why they are principled objections to the claim. We know what the differences are. We keep them in mind in making attributions of content to a simple creature like C, and we avoid drawing rash conclusions from the attributions. This situation is familiar in the theory of measurement. We measure hardness of minerals, pitch, temperature and length using numbers. Each domain has its own level of complexity, which is mapped by a certain set of features of the numbers. We know precisely what we are measuring in each case, so we know that certain entailments from the numbers are not licensed for the objects measured. From the fact that the measure of the hardness of one mineral is half that of the hardness of another it doesn't follow that the one is half as hard as the other. The reason is that the Mohs hardness scale is calibrated against 15 standard minerals, and a given hardness only reflects the fact that a mineral can be scratched only by minerals with higher numbers.

The sketch of the theory so far solves the problem of making fine distinctions in attributing content to unstructured representations or to elements of representations. The attribution that S is about F is made relative to a certain kind of explanation: C receives a benefit from an F's being F when S is produced. The explanation depends on certain nomological truths about F's and about C. They might support counterfactuals like, things that are otherwise like F's but which are not F would not generate a benefit for C if C is caused by an S to produce its typical bodily movement, and things that are otherwise unlike F's but which are F would. These nomological truths provide the requisite fine distinctions among content attributions. (See below, Chapter 6, section 7, for more detail on this point.)

What about structure? There might be some further state which C comes to

produce only given a fair sampling of S-type states, and then only if S-type states always occur with S'-type states (which we'll assume represent G, in the way S represents F). This cautiously produced state might count as a representation that all F's are G's. C might use this representation, along with others, to generate a state that leads it to move to a region of its environment that lies outside of direct sensory range: it might thus have a representation that there is a G behind that rock.

We can imagine finding some repeatable component of these internal states such that the way the states interact with one another and with the analogues of perception and action indicates that the component behaves much as some sentential connective behaves in our language. We can imagine finding structure in these states that is analogous to the quantificational structure of our language.

I am certain that describing the empirical basis for such ascriptions of structure is a difficult task. We know what the detail must look like, though, since we are looking for the sort of pattern of behavior and internal change we know we would have if we acquired thoughts of the relevant types and acted on them. It's unlikely that we could succeed in describing the interactions for complex thoughts and continue to honor a strict constraint that the descriptive resources are naturalistic. In fact I doubt the description of S that shows it to be an F representation honors it. But C doesn't speak, and that is the constraint we are presently investigating.

The general strategy is very much like Quine's strategy in investigating meaning by investigating radical translation.² There are three central differences. First, we aim at no skeptical conclusion; the idea is to show how we can get attributions of content that are as good as those we can get in radical translation. Second, we do not work with query and assent; we need to look directly at the way the inner states change in relation to stimulations from the world and in relation to what the creature needs and wants. (Here we relinquish one of Davidson's central assumptions, that the only data available to a radical interpreter are of the ordinary sort that communicators have about one another.) Third, this project is in a sense constructive, where Quine's is descriptive. We aim at describing various levels of complexity in the patterns of relations between inner states and states of the world that would suffice for translations into content ascriptions with lesser or greater complexity. Perhaps some creatures have something like a sentential logic but no quantificational logic. Or perhaps some creatures lack certain kinds of logical structure our language has; they might be incapable of adverbial modification.

²I am indebted to Klaus Strelau for this way of describing the strategy.

I conclude that we can find the kind of evidence we would need “to justify the fine distinctions we are used to making in the attribution of thoughts” even for creatures that do not speak. Things don’t *have* to have mental representations to have thoughts—in particular, if they speak they don’t—but we have seen no reason to think that the thoughts creatures can have *with* mental representations are any less fine-grained than our own.

2.4 The Central Arguments

TT and RA conclude with similar arguments that non-linguals lack thought. In this section I describe the two arguments and locate the differences between them.

The argument in TT runs like this:

1. If something has a belief it has the concept of belief.
2. If something has the concept of belief then it has the concepts of interpretation and error.
3. If something has the concepts of interpretation and error then it is an interpreter.
4. Therefore, if something has a belief it is an interpreter.

The stated conclusion of the considerations advanced in TT is that “a creature cannot have thoughts unless it is an interpreter of the speech of another” (p.157). Davidson does not make the modal force of his claim precise. It might be that anything that has a belief is something capable of interpretation. Davidson aims at the stronger claim that anything that has a belief does correctly interpret the speech of another. The stated conclusion is non-modal, in the way calling someone a programmer very strongly, perhaps with logical force, implies that the person has actually written working programs. But the remarks that are intended to support it point in both directions. “It follows that a creature must be a member of a speech community if it is to have the concept of belief” can be read as requiring that a believer must do more than live in a speech community. But “. . . we can say more generally that only a creature that can interpret speech can have the concept of a thought” (both passages from p.170) puts the point much more weakly.

The real heart of the argument lies in the first two premises: having a belief requires having the concept of belief, and having the concept of belief requires relating belief to language in the ways we think belief and language are related.

The argument of RA is similar.

1. To have a belief it is necessary to have the concept of belief.
2. To have the concept of belief it is necessary to have the concept of objective truth.
3. To have the concept of objective truth it is necessary to communicate with another.
4. Therefore, to have a belief it is necessary to communicate with another.

Davidson does not put the argument so strongly. He notes that Premise 3 is acceptable rewritten as a sufficient condition, and that he does not know how to demonstrate that it is a necessary condition, although he thinks it is.

TT does not make the claim that if something is an interpreter then it speaks, although Davidson hints that “there may be good reason to hold this” (p.157) RA makes (somewhat implicitly) both this claim and the still stronger claim that if something has thoughts then it is interpreted (not just, interpretable). The reason for the stronger claim is that a creature that is not interpreted lacks something essential for having the concept of objective truth: the sort of “triangulation” on the objective world that communicative interaction with another provides. Hence there is no similar modal ambiguity in the argument of RA: if something has a belief then it can and does both speak and interpret.

There is an interesting difference in the support offered for the second premise in TT and RA. In TT Davidson rests on the strong claim that belief is a theoretical notion essentially wedded to a social notion of interpretation:

We have the idea of belief only from the role of belief in the interpretation of language, for as a private attitude it is not intelligible except as an adjustment to the public norm provided by language. (p.170)

(Compare “Belief and the Basis of Meaning”, p.153: “belief is built to take up the slack between sentences held true by individuals and sentences true (or false) by public standards.”) There are two ways to understand the term “public”: it might require a linguistic community of many speakers, or it might only require two people involved in the existence and maintenance of the standard. Davidson was inclined to the former sense in his papers on radical interpretation written around the same time as TT, but he argued decisively for the latter sense in the later “A

Nice Derangement of Epitaphs” (Davidson, 1986, 433–446). His earlier stance cannot be understood as much more than an inclination, since the central thought of the later paper is strongly suggested by a remark in TT, p.157: “Two speakers could interpret each other’s utterances without there being, in any ordinary sense, a common language.” RA argues squarely from the perspective that no more than two communicators are needed and that no common language is needed either.

This difference is interesting, I think, because it points to a central problem in understanding meaning, one that has plagued many investigators and one we will consider at length in succeeding chapters (especially Chapter 5): the source of the standards or norms that distinguish things with content from other things. Kripke holds that Wittgenstein held that the standards are public in the former sense, and Davidson holds they are public in the latter sense. I believe that they can be public in a still weaker sense: there can be something about what a single organism is and does (something which can be determined by an observer) that sets the kinds of standards that are needed for meaning and belief.

Despite these differences the heart of the argument in RA is essentially the same as the heart of the argument of TT. Let us turn now to the first premise of these arguments, the claim that if something has a belief then it has the concept of belief.

2.5 Believing Requires The Concept of Belief

My remarks have three stages. First, I criticize Davidson’s reasons for this premise. Second, I offer a weaker claim which I think is a necessary condition on representation and which seems to me to capture what is right about the premise. Third, I survey three other arguments for the strong premise and show why they fail to establish it.

TT offers the following reason for the premise: “Someone cannot have a belief unless he understands the possibility of being mistaken, and this requires grasping the contrast between truth and error—true belief and false belief.” (p.170) No reason is offered for the claim that if someone has a belief then he understands the possibility of being mistaken. But surely there is some conception of belief that is very close to, if not the same as, our own conception of belief, according to which something can have a belief but simply not grasp the idea that it might be wrong.

RA offers something more, as follows. If something has a belief, it must be capable of surprise. Being capable of surprise in turn requires that it must be able,

if the belief is false, to reflect that what it believed and what it now believes are different. This capacity clearly requires the concept of belief.

The trouble comes again at the beginning: it does not seem necessary that if something has a belief it must be capable of surprise. Something which made no affective or cognitive response to a false belief would be decidedly odd, but it's not clear why we should think it is not a believer.

It is one thing to fail to make a cognitive response to a particular false belief, but it is quite another to make no response to any false belief. What I think is right about the first premise of the central arguments is that it must matter to a creature that represents things if its representations are wrong. We can put this idea in a way that is independent of the issue whether believers must have the concept of belief:

If something has states with representational content then it must be capable of a corrective response in the face of their falsity or misrepresentation.

I intend for this somewhat vaguely expressed condition to cover a range of cases from the simplest representational systems through human beliefs. Consider the case of very simple representation described in Section 3 above. What makes the state *S* a representation of *F* is that there is an explanation for the presence and production of its instances that refers to a benefit the creature receives when *F* is there. The explanation should support counterfactuals like, if a string of tokens of *S* were produced and no benefit resulted, then the creature would leave off producing *S*'s.

People, who do possess the concept of belief, often satisfy this condition by reflecting that certain of their beliefs must be false. But often enough we change what we think without reflecting. We need to be careful to distinguish three kinds of change. Some mental changes are reflective, and some are not reflective, even though they are rationally appropriate. Finally perhaps some mental changes occur by accident. My weak condition on representation requires either the first or the second kind of change.

A believer that lacks the concept of belief is something extremely different from us. Is there any way to develop this difference into an argument that it is not, after all, a believer? Here are three attempts, and the reasons why I think they fail.

(i) An intentional action is one caused in the right way by a reason that rationalizes it. It is rationalized by the reason provided the agent could have reasoned

(from the reason) that the action has a desirable aspect. A creature without the concept of belief cannot reason that an action has a desirable aspect, since it cannot have the concept of action (since having the concept of an action requires having the concept of reasons). Hence a creature without the concept of belief performs no actions.

Searle's conception of the relation between intentions and actions (Searle, 1983, Chapter 3) has a similar conclusion: an action is a bodily movement caused in the right way by an intention whose conditions of satisfaction include that that very intention cause the bodily movement; but lacking the concept of belief (and intention) a creature cannot have an intentional state whose conditions of satisfaction are like this.

The trouble with this suggestion is that it is unclear why bodily movements caused by rationalizing reasons are not actions, even if the creature that has them is not capable of grasping what makes the reasons rationalizing, or even capable of representing anything about a reason. This difficulty shows up in Searle's theory of action as a difference between the content of an Intentional state and its conditions of satisfaction: it does not always seem necessary that people should be capable of thinking that their perceptions are related to what they perceive in the complex causally self-reflective way Searle claims they must be.

(ii) It has been suggested that for a reason to cause an action "in the right way" (one condition on a bodily movement being an action) the way the reason causes the action must correspond to the idea the agent has of the relation between his reasons and his actions. But an agent without the concept of belief has no ideas about the relations of her reasons to her actions. Hence no bodily movement caused by the reasons of such an agent is ever caused in the right way by the reasons.

There is a parallel argument to be made about the kind of causal relation that must obtain among a sequence of beliefs if that sequence is to count as a course of reasoning. If an agent doesn't (because she couldn't) have views about why one belief follows from another, then the sequence isn't reasoning.

Both these suggestions are taken from responses to a different problem, the difficulty that certain sequences of events do not count as acting on a reason (Davidson, 1980c, 78–9) (or as reasoning) if the sequences are in some sense accidental. Both attempt to distinguish accidental from non-accidental causal chains by considerations about which chain the agent has in mind. But I think the difference between an accident and a non-accident is preserved in the sort of account

of representational content I sketched above. S represents F because there is an explanation for why S occurs that refers to the benefit C receives from F when S occurs. If there is an occasion on which C receives a benefit from an F when an S occurs, but the explanatory path to the benefit is different from the one that generally explains the occurrence of S's, that occasion is an accident. (As we'll see in succeeding chapters, drawing the accidental/non-accidental distinction in this way does not begin to solve the problem Davidson sees for the causal theory of action, since for human agents the relevant explanations are ones that involve the notion of rationality.)

(iii) Putnam's views about Twin Earth and meaning (Putnam, 1975, 237–8) can seem to suggest that the meaning of our words depends on a sort of semantic ideology; we insist that our beliefs should be about things in virtue of what they are rather than how they seem. Consider Archimedes and his word for gold. Putnam considers the objection that perhaps Archimedes had an “operational definition” for gold, so that whatever meets the tests he has for gold is gold as far as he is concerned. Putnam's response is that perhaps our natural kind terms and Archimedes' natural kind terms differ somewhat, but *our* natural kind terms are sensitive to what really is out there, whether we can distinguish it from other things or not. We are realists, not operationalists.

Now I doubt Archimedes' term would be a natural kind term at all if the definition is operational in the strong sense that whatever produces a certain kind of experience is in the extension of ‘gold’, and in particular I doubt his word would mean gold. Clearly more is relevant to the meaning of a term than just the things to which a person has applied it; we need to consider in addition what the agent would apply the term to. This suggests that in order to have thoughts about natural kinds at all an agent has to be able to think that differences in what is really there in the world matter to whether what he thinks is true or false, and to have that kind of thought one needs the concept of belief.

The suggestion is too strong, however. A somewhat weaker way to say why the operationalist's concept differs from the realist's is that the difference between gold and something else *matters* to the realist. We can see that it matters from Putnam's description of what he thinks Archimedes would do if we performed the relevant tests on the non-gold stuff for him. But certainly it can *matter* to Archimedes even if he has no thoughts about why or how it matters. All that is necessary is that he is disposed to undergo the relevant alterations in what he thinks when presented with the relevant evidence.

These three arguments and their problems suggest a general conclusion. The

capacity to reflect on our beliefs as beliefs is very important to our cognitive life. But there doesn't seem to be anything logically wrong with the supposition that all the changes that occur in our beliefs occur without the mediation of reflection.

2.6 Having the Concept Requires Speech

What of the other central premise in Davidson's arguments, that if something has the concept of belief it must actually be an interpreter or a speaker? In this section I will give my reasons for thinking this premise is somewhat shaky. I will also describe an extended argument that is suggested by certain of Davidson's remarks, and show how the argument doesn't establish the point.

In TT the argument for the second central premise reads as follows:

We have the idea of belief only from the role of belief in the interpretation of language, for as a private attitude it is not intelligible except as an adjustment to the public norm provided by language. It follows that a creature must be a member of a speech community if it is to have the concept of belief. (TT, p.170)

Two ideas make up this argument. First, if a concept is analytically related to certain other concepts, then if someone has the one then she has the others. Second, if a concept is analytically related to certain other concepts, then if someone has the one then she actually applies the entire set of concepts.

If the reasoning of the last section is successful then the concept of belief is not, after all, analytically connected to the concepts of language and communication. But let us suppose that it is.

The first idea is very plausible, but not, I think, compelling. If Burge is right (Burge, 1979), someone might command the concept of arthritis yet fail to understand that it is conceptually necessary that arthritis occurs only in the joints. Suppose the basis for holding that there is an analytic connection between concepts is the sort of basis we have for such judgments: a long history of wide linguistic usage. Then a particular person might learn one of the concepts and not the others. But Burge's "social externalism" about meaning is not necessary to make this point. I think that if someone consistently applies a word in a certain way (to arthritis, for instance) and otherwise performs as we would perform with this word, the fact that she lacks certain concepts we think are strongly tied to that

word is not sufficient ground to deny that her word means what we mean by it. (On the other hand evidence that someone lacks these other concepts is usually extremely good evidence that the person is *not* able to apply the word as we do.)

The difficulty with the second idea is the modal ambiguity we canvassed in section 4 above. It is certainly possible for an agent to recognize that one of her concepts has certain analytic connections with other concepts, and for her to fully grasp and understand the concepts, without actually applying them. Hence there appears to be no reason to deny that someone could have the concepts of belief and language but not be a speaker or interpreter.

The argument in RA is much more subtle and much more difficult to evaluate. To have the concept of belief one must have the concepts of truth and falsity, since beliefs are the sort of thing that can be true or false. To have the concept of truth one must have the concept of the distinction between how things seem to one and what is objectively so. Davidson puts this variously as the concept of objective truth or the subjective-objective contrast. He concedes he does not know how to show that possession of the concept of intersubjective truth (a requisite for communication) is *necessary* for possession of the concept of objective truth (RA, p.480). In place of an argument he offers an analogy:

If I were bolted to the earth I would have no way of determining the distance from me of many objects. I would only know they were on some line drawn from me toward them. I might interact successfully with objects, but I could have no way of giving content to the question where they were. Not being bolted down, I am free to triangulate. Our sense of objectivity is the consequence of another sort of triangulation, one that requires two creatures. Each interacts with an object, but what gives each the concept of the way things are objectively is the baseline formed between the creatures by language. (RA, p. 480)

I'd like to sketch one way to develop the analogy and show why the sketch falls short of demonstrating the point.

The analogy is based on a relatively precise claim: that a surveyor who is unable to move is unable to "give content" to statements of how far an object is from her. "Giving content" means, I think, being able to apply an empirical test that grounds a measurement scheme. We can measure weights with real numbers because we have empirical tests (using an equal-arm balance) for when two things weigh the same and when one thing weighs precisely twice another. Triangulation begins with measurement of a baseline, adds to it information about the angles of

vertices of the triangle formed by the baseline and the object under inquiry, and computes distance using geometry. Triangulation requires two empirical tests: one to determine the length of the baseline, and one to determine the angle formed by the baseline and the sides of a triangle. (This test must be applied twice, for the two ends of the baseline.)

I can see no principled reason why a surveyor who is “bolted to the earth” could not apply both these tests. The surveyor might possess a meter stick; then any point at the far end of the meter stick, when the near end touches the surveyor, is one meter away. Any line between her and one of these points can serve as a baseline. She can determine the angles at the ends of the baseline with a theodolite; she need not actually be at the far vertex.

It is surely possible to improve the description of “bolting” so one or the other of these tests becomes impossible for the surveyor. The surveyor might nevertheless appreciate what she is unable to do; her statements about distances would then have content, but a content she is unable ever to verify. The only way for her statements literally to lack content is for there to be some principled difficulty in the way of an empirical test, for instance if it is nomologically impossible for the relevant measurements to be made.

If we apply the analogy to the case of possessing the concept of objective truth, the results so far are disappointing, and reflect problems we have seen before. We don’t yet know what the relevant empirical tests are for applying the concept of objective truth, or why a non-communicator would be unable to apply them. But we clearly do know that whatever they are they can be applied, since we do apply them. So at best the analogy so far suggests a non-communicator might understand the concept but not be able to apply it.

Here is another other way claims made using a measurement scheme might be said to lack content (Suppes, Patrick and Zinnes, Joseph, 1963, 11–15). The measurement scheme might be said to measure empirical properties on a scale of a certain type, for instance an interval scale. If the empirical basis for making the measurements does not support the claim about the type of scale then claims of measurements on a scale of the relevant type lack content. Distance is measured by an interval scale; the analogy suggests there is a way of describing being bolted to the earth that shows that measurements made from this basis could only be “weaker” scales in the sense that they are subject to more transformations.

Let’s try to develop the analogy in this direction. I’ll describe a series of refinements to a too-simple causal theory of content, where each refinement is

prompted by considerations that appear to be ones Davidson has in mind in RA and his “Reply to Burge” (Davidson, 1988). The last refinement uses communication to provide a solution to a difficulty about change of belief.

We start with the simplest possible causal theory of representation: a type of state of a creature is about whatever causes it. There are two problems. First, these representations can never be wrong, since their cause is always actual. Second, the content is fixed by consideration of the cause of each token of the state. But what is the cause of a particular event? There appears to be no principled way to select any particular region of the light-cone preceding the occurrence of an event. Hence the content of the state seems to be more or less arbitrary.

A first suggestion is that a state-*type* of a creature represents that *type* of feature of the world with which the state-*type* is best correlated over time. While singular causal claims are interest-relative in the sense just mentioned, general causal claims are not; they are supported by considerations about what is common to all possible light-cones that produce the relevant type of effect.

There is still a problem about error. If it is possible for instances of the state-type generally to be produced by more than one environmental type (which must be possible if error is possible) then there is a type of state such that the internal state type is better correlated with this type than any environmental type: the disjunction of states of the surface of the creature that lead to the production of the internal state type. (This disjunctive state is clearly not any “natural” way of categorizing the world, but the theory with which we are now working does not require “naturalness.”)

This problem *isn't* solved if we make the causal relations to the world more complex. We might say that the creature manages to *improve* the reliability of the representation-type, and thereby alters the proximal states that give rise to it. But there is still a more complex, *changing*, state of the world with which the representation-type is better correlated than any in the creature's environment; it's again the surface states of the creature, now considered as one object.

What's needed at this point is some way to select between things in the world and surface stimulations. Benefit solves the difficulty. The inner state type is about that feature of the world which produces it and which accounts for the success the creature has when it produces it. The condition of the surface of the creature does not account for success, since without the environmental cause there is no success. (Benefit usually forces the choice of some distal cause, but this is a contingent fact about how goods are distributed in our world. If one could benefit simply from

the state of one's surface, the content of the inner state would be that the surface state obtained.)

There is, however, at least one difficulty remaining, and communication appears to provide a solution.

The sketch of a causal account so far makes the content of a type of state relative to what provides a benefit. But the relation between the external type and the benefit has not been stipulated to remain constant. If this relation changes the content of the representation can change.

Suppose that over time what the creature needs in order to survive alters, and the creature makes suitable alterations in what proximal states yield the representational state type *S*. The content of the representation drifts as time passes. The relation can change from the world side as well, if what property is available to provide the benefit alters. Suppose the benefit is initially provided by things that are *F*, but in time *F* things disappear; but during that time *G* things become more available, and provide a similar benefit. So long as the difference doesn't matter to the creature, there is simply a smooth shift in the content of the representation, from *F* to *F-or-G* to *G*. We may call this the problem of content drift. It appears to be a serious problem, since beliefs cannot change their content, whereas the state-type *S* apparently can.

If the creature communicates, though, we can describe a clear way in which a drift in content makes a difference. Recall Davidson's claim that belief is "not intelligible except as an adjustment to the public norm provided by language." The problem of content drift is that the standard provided by the notion of benefit is so weak that representational content does things that belief content cannot do. But communication provides a much stronger standard.

Two distinct creatures are essentially causally independent of one another. Their histories are different. Usually they come from different parents and have a different internal constitution. Even if they are twins they differ physically in a variety of ways. They are in different places. They go different places. Each is subject to different pressures from heritage and environment.

Since they are different in these ways, if the content of their representations drifts it is possible for it to drift independently. In fact it is likely that it will.

Now suppose they are trying to communicate. Suppose for simplicity's sake they are reporting their beliefs honestly and sincerely. Each belief report expresses the content of a representation. If the contents are changing then what the reports are keyed to changes as well. If each creature changes independently, their word

use changes independently. Neither can predict what each new report might mean. In this case they cannot communicate. Communication requires that communicative interactions yield or have the potential for yielding benefit to the participants that derives from the exchange of information. Neither can bring the other to believe anything about the world by way of getting the other to respond to the content of its states.

If they are communicating then they are holding their use of utterances relatively fixed with respect to their other changes; fixed enough for one to take the other as using the same utterance form again for the same things. This relative fixity is what provides a standard against which to measure their beliefs. It is a standard over which neither has complete control: the other may at any point fail to catch on. That's what makes it a standard; and it only continues to be a standard if they continue to communicate.

That completes my sketch of one way to understand Davidson's analogy.³ I think Davidson is correct that this kind of causal interaction does settle some question, but I do not think it makes a sharp distinction between things that have concept of objective truth and things that have no representational contents at all. The difference I have indicated is a matter of degree. Our language is subject to content drift as well. It simply drifts more slowly and under different sorts of pressures. The appearance that the content of non-communicators really does something different from the content of beliefs comes from a mistaken concentration on the non-content properties of the representations, the property of being state S. When the content of the representation changes, we should say the creature has a different representation, since certainly beliefs, and other representations, are individuated according to their content properties, not their non-content properties.

2.7 Conclusions

I have characterized Davidson's position on animal thought as consisting of three types of argument, one about intensionality, one about holism, and one about what is needed for something to be a believer, and what is needed for something to have the concept of belief. I have argued that the first and last type of argument fail; I will return to holism in Chapter 6 below.

³I hasten to add that it is unlikely that this is Davidson's way of understanding the analogy. His suggestion is extremely intriguing, but I am unable to make better sense of it than what I have written.

Despite these problems Davidson is surely right that there are very important differences between believers like us and other things. We may be inclined to build these differences into the definition of belief.⁴ For my purposes that is an acceptable stance to take; what I have been concerned to show is that Davidson's arguments have no force against the claim that things that do not communicate could have states with semantic properties, and in fact could have propositional contents as complex and fine-grained as the contents of our propositional attitudes.⁵

⁴On this suggestion the difference between a believer and something just like a believer, except in speaking, would in some cases be rather like the difference between a statue and the bronze of which it is made. Something might be just like me and fail to be a communicator; our sense that it and I are psychologically importantly similar might correspond to the sense that even if the statue and the bronze are different they are surely not distinct things.

⁵Dugald Owen, Claudine Verheggen and Bruce Vermazen each read an earlier version of this chapter and patiently corrected various mistakes I made.

Chapter 3

Causal Semantics

3.1 Introduction

In Chapter 1 I considered the claim that the propositional attitudes must be rational. My conclusion was that if we are willing to regard semantic properties as *sui generis* there is no compelling conceptual reason to hold that the attitudes must be rational.

There is something deeply unsatisfactory about holding that semantic properties are *sui generis*. One version of this view claims that semantic properties are properties to which we will appeal in our basic understanding of how the world works; in effect, they are a new kind of physical property. But it is extraordinary to suppose philosophical rather than physical investigation should force us to make such a claim.

The alternative is to hold that we can “give an account” of what semantic properties are. If we make this assumption I believe we can show that the propositional attitudes are rational, or nearly so. “Giving an account” is saying how semantic properties are related to other properties. Such an account is a second-order theory; it specifies what sort of relations are relevant to the construction and confirmation of a first-order theory that specifies what semantic properties a particular creature has. Since what is most puzzling about semantic properties is that things that have them are about things, one kind of investigation into the nature of semantic properties examines a variety of relations things have with one another. Since our thoughts can be about anything in the world whatsoever, we are looking for a pervasive relation between things. We are also looking for a

relation with explanatory power, since one point in attributing semantic properties is to explain what people do. Causation, the “cement of the universe,” is a very appealing candidate for the central relation.

In this and the next three chapters I will investigate attempts to work out a more or less causal account of content. Each of these accounts makes the plausible but controversial assumption that creatures much simpler than ourselves have semantic properties; in the last Chapter I showed why one set of reasons advanced for the claim that only people really instantiate semantic properties is not persuasive.

I take it that very simple creatures like bacteria and frogs are not rational. Hence even if a causal account shows they have semantic properties, they could not show that if something has semantic properties then it is rational. But I will argue that we can show, by considering these causal accounts, that if something has semantic properties then it is a member of a class to which rational creatures also belong, and that membership in this class by creatures with semantic properties like our own is earned only by being rational or nearly so.

What class is this? Rational action is action explained in terms of a benefit at which the agent aims. Rational thought is thought explained in terms of the benefit of following the norms of rationality. Common to various kinds of rationality is what we may call a benefit-involving explanation. My claim is that causal accounts of content show that having semantic properties is essentially a matter of having a benefit-involving explanation. Causal accounts show this in two ways. First, they may be patently inadequate unless supplemented by claims about such explanations; this is true of the accounts we will consider in this Chapter and in Chapter 6 below. Second, the account may simply make the claim; this is true of the accounts we will consider in Chapters 4 and 5.

Having a benefit-involving explanation is being a member of a wider class to which rational agents and their propositional attitudes also belong. If that were all we could say about the relation of content and rationality, there could be very complex creatures with states with semantic properties (and benefit-involving explanations) who were not rational. I will show, however, that there is more to say: benefit-involving explanations for the presence and activities of the cognitive states of very complex creatures simply amount to explanations which show these creatures as rational.

I begin with Dennis Stampe’s theory (Stampe, 1979). In the next Section I’ll describe his view. In Section 3 I will discuss and criticize his motivations for his theory. Section 4 briefly considers Stampe’s appeal to causation to settle which

object is the object of a representation. Section 5 considers at length Stampe's way of setting the content of a representation. This Section shows why representations are items with benefit-involving explanations, even though Stampe does not make this claim. Finally in Section 6 I will describe how his theory might be used to determine the content of our thoughts, and show how a first order theory can make correct particular attributions of propositional attitudes only if the second-order theory that describes its construction and confirmation makes use of the notion of what is rational to believe.

3.2 Stampe's Causal Theory

Stampe aims to "naturalize" content (p.90):

Representation is an altogether "natural" relation; there is nothing essentially conventional about it. There is nothing essentially mentalistic about it; it *may* be a wholly physical relation. Neither is there anything essentially *semantic* about it ... (p. 87)

Stampe is not careful to state what makes a relation naturalistic or mentalistic or conventional or physical or semantic. Hence it is difficult to know what his project is or whether he has succeeded at it. I will be claiming that his account requires a notion of function that involves a relation to benefit. Then whether Stampe succeeds in "naturalizing" content depends largely on whether benefit is a natural notion.

The theory has two components. The first is designed to give a causal answer to the question what thing in the world a representation is about. The second shows how to determine the content of a representation using information about the causation of the representation.

Call the first component the "singularity condition". Let 'R' and 'O' name a representation and an object, respectively. Let 'pR' name an ordered set of properties that R has, and 'pO' an ordered set of properties that O has. Let 'CR' name a two-place relation on properties, such that properties x and y satisfy CR just in case x is causally relevant to y.¹ Let 'CRC' name a 4-place relation on two objects and two ordered sets of properties, such that $CRC(x,y,z,w)$ is satisfied by a

¹Causal relevance is an explanatory relation on properties that is general and supports counterfactuals: if $CR(x, y)$ then the general causal claim "x's cause y's" is confirmed, and statements of the form, if some particular event had been x then its effect would have been y, are reasonable to

pair of objects x and y if and only if x causes y , x has the properties in z , y has the properties in w , the cardinality of z and w is the same, and, letting i range from 1 to the number of properties in z , and letting ' z_i ' name the i th property in z , $CR(z_i, w_i)$.

Then R is a representation of O (or, O is R 's object), just in case $CRC(R, O, pR, pO)$.

Now of course every particular object stands in this causal relation with many objects. There are many causal chains affecting each object, and each object stands in this relation with each link in each chain. So the singularity condition is only one necessary condition for R to represent O ; the second condition, the content condition, will be used to isolate one of the links of one of the chains.

The singularity condition is a relation on two particulars and two sets of properties. The content of the representation is determined with respect to some general features of the representation, i.e., some facts about what happens to things of R 's type. Now of course there are many things that share some aspect with R which are not representations; presumably the general truths that determine content are truths about a special sort of type. Stampe does not attempt to specify what sort of type this is.

Things of R 's type are produced in many situations, and presumably could be produced in others as well. Call the set of conditions that could be relevant to the production of an instance of R 's type, the "production conditions." Call a special set of these production conditions the "fidelity conditions." Then there is a counterfactual governing the production of things of type R : if the fidelity conditions were to obtain then something of type M would cause R . (We shall need to add something like, "would be the most likely cause," since usually more than one thing would cause a given type of occurrence.) The causal claim should be understood as a claim about causal relevance: the property of being an M is causally relevant to the property of being an R . This counterfactual doesn't specify what *does* cause R , either when the fidelity conditions obtain or when they do not; rather it specifies the property it is most reasonable to think would cause R given

assert. Stampe doesn't cast his theory in quite these terms, but I think it is what he has in mind. He writes, "The causal relation we have in mind is one that holds between a set of properties $F_{(1...n)}$ of the thing (O) represented, and a set of properties Φ ($\phi_{1...n}$) of the representation (R)." (p.85) Of course, causal relations cannot hold between sets or properties, so there is no such causal relation. In Chapter 8 below I describe some features of a theory of a causal relevance relation on properties.

the fidelity conditions.² Then the content of R is M, or, that something is M, or, M is instantiated, or, M there.

Stampe is somewhat vague on the nature of the fidelity conditions. He writes,

But one doubts whether statistical normalcy will get us far in dealing with living systems and with language, or generally with matters of a teleological nature. Here, I think we shall want to identify fidelity conditions with certain conditions of well-functioning, of a functional system. (p.90)

Statistical normalcy apparently gets us *somewhere*, if not far; this suggests that the notion of function isn't essential to representation. Conditions of well-functioning for something with a function are those conditions such that if they obtain, the thing performs the function (p.90). The fidelity conditions relevant to content are conditions of well-functioning for some object which performs a function when suitably related to a representation, and which performs it only when the representation, as it were, "gets the facts right." Then the fidelity conditions are some subset of the conditions of well-functioning.³ Here is Stampe's statement:

Now consider a mechanism, the function of which is fulfilled in part through the generation of such representations, and fulfilled only if those representations are more or less accurate ones. A subset of the well-functioning conditions of such a system will be conditions such that if they obtain, then the accurate-making state of affairs would be capable of causing the representation to occur. These are "fidelity conditions." (p.91)

I'll close my exposition of Stampe's theory by describing the ways it solves the problems he sets for it.

²Stampe says the counterfactual gives a hypothesis which is "reasonable to accept" (p.90). This is an odd notion to use in an attempt at a naturalistic account of content; perhaps there is a way to avoid the appearance that what people think is reasonable plays a part in determining content.

³Stampe's theory has a striking resemblance to Millikan's (see (Millikan, 1989a) and Chapter 5 below). The content of a representation is determined by something that holds when something other than the representation successfully performs a function. Perhaps motivated by the difficulties in making out what Stampe means by "conditions of well-functioning" (see section 5 below) others, notably Fodor, have tried to understand content in terms of the conditions of well-functioning of some internal mechanism. The virtue of this approach (that we can say what it is for an internal mechanism to be in top condition) is its vice as well: an artifact or a creature may be in top condition yet unable to perform a function, because the world doesn't cooperate; conversely it may be in very poor condition yet able to perform the function because the world is helping out.

The singularity condition is designed to determine the object of a representation, particularly in the case where (apparently) the content does not determine it, as for instance a representation of one, but not the other, of two identical twins. The solution comes with the CRC relation among properties: the object that R is about is the one that causes it such that if that object were different then the representation would be different in certain ways, and the difference would depend on the difference in the object. Hence if one twin takes a photograph of the other twin the photograph is a photograph of the one who is such that, if he were different in certain ways, the photograph would be different in certain ways.

The content condition is designed to solve three problems. First, it must isolate one of the links in one of the causal chains that yield the representation. The solution is that the representation is about the link that contributes to the performance of a function by something connected with the representation. Roughly put, the one it's about is the one it's supposed to be caused by. Second, the content condition is designed to provide a specification of the content of the representation; we have seen how it does this. Third, it is designed to give a causal account of the possibility of error. This is a critical problem for causal accounts, since if the content is set by actual causes then error is impossible. R is correct (or not in error, or true, if it is a thought or an utterance) just in case the object it is about has the properties the content condition gives as its content. (Stampe is preoccupied with reference; he will need to weaken this claim to handle representations that do not so clearly have a single object they are about, as for instance utterances of existentially quantified sentences.) R is in error provided no event among its causes has the configuration of properties its content requires.

3.3 Motivations

Stampe is concerned both to articulate a causal account of representation and to justify the idea that representation is essentially a causal phenomenon. If he could justify the idea, we might avoid the conditional conclusion of Chapter 1: we would have reason to think content could not possibly be *sui generis*. In this section I show why Stampe does not provide this justification.

He begins with the following thought. He notes that it can't be a coincidence that there are causal theories of knowledge, memory, belief, evidence, proper names, common names, and reference. In particular, if these theories turn out to be acceptable then there must be some causal account common to all of them.

Evidently what is otherwise common to all of them is that the things analyzed have semantic properties, or involve representations that “involve an object” which may or may not exist; perhaps then what is common is a causal account of representation.

I doubt this is a good reason for thinking there must be a causal account of representation. A simpler explanation for this commonality could well be as follows. We are caused to believe things by the world, and we cause things to happen in the world. This causal contact with the world runs through states with semantic properties. This causal contact is absolutely central to the way we live our lives. It is no wonder that we can articulate theories of various semantic properties that relate them to causation. But that fact does not license the claim that semantic properties are essentially causal.

Stampe is clearly motivated by the desire to provide a causal alternative to what we might call the “classical” theory of reference, a Fregean account that holds that the object, if any, to which a thought refers is the unique object that has the properties specified by the thought. He is confident that Donnellan and Kripke’s work on reference (Donnellan, 1966; Kripke, 1972) shows that such an account could not possibly be correct. I think Searle has shown that a Fregean account could possibly be correct, and hence I do not think we can find a persuasive justification for a causal account of representation in this direction (Searle, 1983, Chapter 9).

Stampe offers what appears to be a more fundamental or foundational reason for thinking that representation should be a causal phenomenon:

These states of ours that have objects—these perceptual states, these beliefs, desires, intentions and fears— are largely responsible for the success with which we occupy the world. But that harmony would be inexplicable were it not that by and large the objects “involved” in those states are the very constituents of the world—the world therefore *determining* that these states exist and have by and large a character that promotes our survival in it. Thus it cannot be by luck that our beliefs and desires correspond as they do with the facts and the characters of things; it must be determined that they should often so correspond, and determined *by* those facts or things. The idea that this determination is *causal* determination is, if not inevitable, only natural. And only it perhaps, is Realistic. (p.82)

There are any number of ideas contained in this passage; I seriously doubt any

of them contribute to justifying a causal theory of representation. Perhaps most salient is an idea about Realism. Philosophers of science argue that we cannot explain the success of science without supposing that the things our theories claim exist really do exist; if there are no electrons, the idea goes, then we have no explanation for why we get such good predictions with our atomic theories. In causal semantics the parallel argument would be that we cannot explain our success in the world unless the world really is the way we think it is. Both arguments presuppose there is something to be explained: successful science, and actual harmony in the world.

The grain of truth behind the denial of Realism, I think, is that the falsity of our theories is consistent with their predictive success, if that success is couched in terms of observations. This grain of truth is not, however, a good reason to hold that there really are no electrons. A similar difficulty attends the claim that we cannot explain our success in life except by appeal to the truth of our beliefs: a successful life can be led by someone most of whose beliefs are false. There is no justice in representation.

There are further difficulties familiar from the philosophy of science. Perhaps it is true that an explanation of success cannot be given if a certain condition does not hold, but unless we know an explanation *must* be given we cannot infer that the condition does obtain. And, as Stampe is clearly aware, we cannot infer that one condition that would explain success does obtain, if there are other conditions that would explain success as well.

Even if we were confident that our success in the world requires an explanation of the sort Stampe has in mind, it's hard to see how the fact that the things we represent cause our representations of them would be much help in explaining that success. Things in the world could so cause our representations of them that our success was very unlikely—if they present appearances very different from the truth, for instance.

The explanation that Stampe envisions has to do with objects in the world “*determining*” the existence and character of our mental states. Stampe does not say what sort of determination he has in mind. It is doubtful he means no more than that the existence and character of our mental states should depend causally on the existence and character of things in the world. Certainly one way to know a bear is in the vicinity and that it presents a danger is to be caused to believe this by the bear; but that's not the only way to be a success in the world.

Stampe's theory suggests that he means by “*determining*” that the semantic

character of our thoughts is *logically* dependent on their causes. He doesn't provide any reason to prefer this suggestion to the alternative that content is logically *sui generis* and that something else explains why we tend to have true thoughts and perform successful actions. He may have in mind that there cannot be any such alternative explanation. The parallel thought in epistemology is that there is no way to avoid skepticism about the external world short of claiming that what things our thoughts are about is logically determined by what actual things they are caused by. It is an open question how well this idea works for guaranteeing the truth of some or many of our beliefs. For the idea to work at all for the success of our lives it seems we should have to make success a logical determinant of content. This is, in fact, what Stampe seems to have in mind, since the content of a representation is related to the actual performance of a function, and presumably the performance of a function has something to do with success. As with the epistemological suggestion, it is hard to get any clear guarantees. The nature of the connection between content and success might allow a representational system to have very little success in the world.

Similarly it doesn't appear to be necessary, for an understanding of how we succeed in the world, to claim that the "objects of thought" are the actual objects in the world, if this means more than that our beliefs are sometimes true—as for instance that our beliefs relate us to singular propositions that *include* actual objects.

I think Stampe is on to an important connection between content and success, but in this paper he has it badly confused. He is clear that we must use the notion of function to get an account of "psychological content and linguistic meaning." He writes,

The causal conjecture springs most deeply from the desire to understand how our psychological states and their linguistic expressions contribute to our survival in a world of things distinct from them and us—or, to give the question a functionalist twist, how those states and expressions fulfill their functions in our lives. (pp.83-4)

I think he has the thought exactly backwards. The causal conjecture and its functionalist twist spring most deeply from the nature of content, not from something content does. The central issue in understanding content is understanding the possibility of error. What Stampe misses is that only the connection with teleology and with some notion of success permits an understanding of error. The connection between content and success doesn't begin to provide a guarantee or

explanation of actual success; rather it is only against the background of the possibility of success that we can make sense of one thing being right or wrong about another.

I think this conceptual link is as strong as conceptual links that are not logically necessary can get. Again, the conceptual link is not so strong as to provide an argument that we could not possibly treat content properties as *sui generis*.

3.4 The Singularity Condition

We have seen that the singularity condition isn't meant to be a sufficient condition. Is it necessary? Stampe considers the objection that the condition need not be causal, since a nomic relation between properties of the object and properties of the expression would do as well. He responds as follows:

we could not answer the question of singularity. We could not do so, because our only resources are relations between sets of properties, but those sets of properties even if nomically related in such ways as these could be instantiated by a plurality of objects. The introduction of a causal relation, however, allows us to form the singular proposition that we need to answer the singularity question. (p.87)

I take it the singular proposition in question is one of the form, the *a* caused the *b*, where '*a*' and '*b*' name events involving the object and the expression respectively. The argument is not clearly a good one. We don't need a singular proposition; a proposition that specifies a unique object will do. But such a proposition need not state a causal relation. For instance the reference, if it has one, of my phrase, "the object nearest me in space and time yet beyond informational range" is the unique object, if there is one, nearest me in space and time yet beyond informational range, yet there cannot be a causal relation.

Stampe is surely right that causation is important in some way to reference. I doubt we can make sense of the idea of a creature all of whose representations are about things outside informational range of it. How would it ever tell whether it got things right or wrong? What difference would being right or wrong ever make to it?

Despite the importance of the connection between reference and causation I do not think reference is best understood as a causal notion. Stampe's own

cases make this clear. Consider twins, one of whom takes a snapshot of the other. Certain features of the resulting photograph depend causally on certain features of one of the twins and not the other; if someone with different hair color had been where the one was, the print would have had a different color. But clearly there are other features of the photograph that depend on the other twin: where the camera was pointed, how blurry the print is, etc. Our interests provide the fidelity conditions that determine what the photograph is of: we built the camera to provide a very controlled way of exposing film to light reflected from an object related to the camera in a very precise way. There is clearly nothing naturalistic or physical about the causal relation between photograph and subject; if we wanted to use cameras as sobriety testers we could take a print's degree of blurriness as a representation of the inebriation of the person who held the camera.

3.5 The Content Condition: Function and Causation

Stampe's content condition relativizes content to function and causation. Evaluating the condition requires understanding what notion of function might be involved. Even though Stampe's own views on function are seriously incomplete I think content must be relativized either to function or to some teleological notion. Stampe's causal condition, on the other hand, is neither necessary nor sufficient for content. Most of this section will be devoted to function, and I'll argue this claim about causation at the end.

Stampe has extremely little to say about function. He feels that philosophers are wary of talk of function because things that have functions don't always, or ever, perform them, and this makes the notion empirically suspicious. But this is not a genuine problem. Dispositional properties are similar in not always being actually manifested.

Stampe's solution to his problem runs like this:

In honor of these doubts we may erect a principle that in part vindicates the empirical significance of function ascriptions, laying down a procedure for their empirical disconfirmation. A mechanism can truly be said to have a function only if there are specifiable conditions of well-functioning, such that were they to obtain, then the function of the mechanism would be fulfilled by its operation. (p.90)

I understand Stampe's principle to make the following claim: mechanism *M* has the function *F* just in case there are conditions *C* (the conditions of well-functioning) such that if *C* obtains then *M* performs the function *F*. The principle is problematic on several grounds. First, it yields no account of what it is for a thing to have or to perform a function. Second, a broken can-opener has the function of opening cans, but there are no conditions in which it will open cans. Third, the conditions under which a device that *can* perform a function *will* perform the function are not finitely specifiable except trivially. A functioning thermometer will tell the temperature correctly in normal circumstances. Suppose a distant star goes nova and emits a beam of radiation that boils the mercury. That's not a normal circumstance. We can trivially specify normalcy as conditions in which *M* can perform its function; a non-trivial specification must describe all the infinitely many extraordinary circumstances which would block function performance and state that they do not obtain.

Stampe's examples of items with functions—thermometers and barometers—are all artifacts, and hence are of little use in getting a naturalistic account of function. His single more or less natural example is the rings in a tree-stump; he claims that if fidelity conditions are “normal climatic conditions” then the number of rings represents the age of the tree. But the tree doesn't represent its age for itself; we take the number of rings to represent the age because we are interested in the age. We must find a notion of function such that the functioning items function for their possessors and do so regardless of what we think about them. One such notion relates function to survival, as follows.

Consider a class of creature members of which need various substances in order to survive. These creatures have an internal state which is causally correlated with the presence and relative direction of one of these substances. This internal state in turn produces characteristic movements that lead to useful interactions with the substance, i.e., ones that promote survival. Then these movements function to obtain this substance, and function for the creatures, regardless of what we think about them. Furthermore, the internal movement-controlling state is a representation of the presence and direction of the substance.

Supplemented in this way I think Stampe's condition is sufficient for a very simple sort of content. There are difficulties remaining (distinguishing representations from other inner states that are necessary for function performance, for instance) but I think the suggestion is on the right track.

Is this relation to function necessary for content? The answer depends on the notion of function and the sort of representational system we are considering.

This survival- based notion of function is not necessary for human thought content. But some connection with a teleological notion is necessary, whether it is a simple, more or less naturalistic notion like this one, or some more complex notion suitable for understanding human thought content. I think we can show this for the very simple system at hand by asking whether states that lack this direct connection to survival and benefit are representations.

Let us suppose that there is a very straightforward selectional explanation for why these creatures are capable of representing the necessary substance in this way. They stem from a population otherwise like them but which lacked the internal movement-controlling state. We may suppose these ancestors simply ran into the needed substance often enough to permit survival. A mutation occurs; the creatures that have the mutated gene are able to represent the substance and as a consequence are better able to survive than the ones that do not.

Any single mutation is liable to have many morphological consequences. Let us suppose that the original mutation produced a change in the creatures that introduced a certain sort of mechanism that leads from the representational state to the bodily movement that sometimes yields the beneficial substance. Now suppose instead that the mutation yielded two states, the original as before, and another. The second is caused by an environmental substance in much the way the first is caused, and it causes a bodily movement by the same mechanism. The bodily movement yields contact with the second environmental substance in much the same way as the first. The difference is that interactions with the second substance are not good for the creatures; but they are not so bad that the harm they yield outweighs the benefit the first representational state yields.⁴

This second state resembles a representation in many ways, but lacks the direct connection to benefit that a representation has. We can increase the resemblance by increasing the complexity of the representational system. Suppose there are many representational states and a complex mechanical system that leads from “perceptual” states to “actions” in the creature. Suppose on balance the operation of this system yields a benefit, but there are one or more internal movement controlling types that do not and cannot yield a benefit, even though their interac-

⁴I intend for my example to resemble anti-adaptationist claims in evolutionary theory (Gould and Lewontin, 1979). Two ideas should be distinguished. First, a particular feature of an organism’s structure may be selected but not selected for; it may be present only because it comes along with other features of the organism that are selected for. Second, I do not claim that something is a representation only if it is selected for or if it is an optimal solution to some environmental problem. All representation requires is some relation to benefit, however minimal.

tions with the environment and bodily movements are just like those of their more useful companions. These “rogues” look just like representations.

I think that the mechanical similarity to real representations should not be reason to think these “rogues” are representations. They lack the connection to benefit that is required for content. Clearly they do have some benefit- involving explanation, since they only exist because they have to accompany the other states that do confer benefit. But for representation we need a more direct connection to benefit.

How should the connection to benefit be articulated? Suppose the creatures or the environment to change, so that one of the rogues becomes useful. In a particular creature it becomes a representation when one of its occurrences is explained by its connection to benefit, if, for instance, it is reproduced because an earlier instance was useful. In the class of creatures it is harder to say when a kind of rogue becomes a representation. It might be enough for it to be a representation in one creature of the class if it is generally true that it confers benefit, or if it becomes easily possible for the rogues to confer benefit, even if they have not yet done so.

What about the relation Stampe requires between content and causation? A suggestion of Ruth Millikan’s shows it is neither necessary nor sufficient for content, even for the simple representational systems we are now considering.⁵ Stampe claims a representation is about what would cause it if conditions of well-functioning are met. But why should it be about what would cause it at all? Why is it impossible for a representation to be about some aspect of things that is never causally responsible for the representation? For concreteness let us consider an actual very simple representational system. Venus Fly Traps shut their leaf traps on small objects that land on the villae within the trap. The function of the shutting mechanism is, I suppose, to obtain nourishment. Some state of the villae or the path to the device that pulls the leaves shut is the representation. It represents the presence of something nourishing. Perhaps we can refine the ascription further: flies are what would cause the trap to shut that would allow the shutting mechanism to fulfill its function. Yet what would cause the mechanism to shut is the same whether the function is performed or not: light pressure of an ambient object on the villae. What is causally relevant to the production of the representation when it contributes to function performance isn’t the fact about the fly landing that it is a fly, rather it is the fact that something touched the villae with a certain force. (See below, Chapter 5, for a case where the object that causes the representation

⁵(Millikan, 1989a) I discuss the point below, Chapter 5, Section 5.

is distinct from the object represented.)

Hence Stampe's causal condition on content is not necessary, since the trap-shutting mechanism represents flies even though what makes it shut (the property causally relevant to its shutting) is not flies; it is not sufficient, since the trap-shutting mechanism is controlled by light pressure but it does not represent light pressure.

3.6 Stampe's Theory Applied to People

To apply Stampe's content condition to people we need some way to understand the notions of function and conditions of well-functioning for people. The simple survival-based notion of function will not do, since people do not always aim at their own survival. But again some relation to benefit, however understood, is needed. For purposes of evaluating the causal aspect of Stampe's condition we can help ourselves to a thoroughly non-natural notion: the equivalent of fulfillment of function is successful performance of an action. Stampe's conception of conditions of well-functioning would then require a specification of conditions under which successful performance of an action is guaranteed. Again there is no hope of non-trivially specifying such conditions. So we treat the suggestion as an idealization; if there were such conditions, the content of a person's thought would be what causes it when the conditions obtain. The question then is whether this suggestion gives the right answer about the content of a thought. I think it does not.

Picture a person in a café at a table. There is a cup of coffee in front of her. She has good eyesight, thinks clearly, the lighting is good. She has the thought that there's a cup of coffee in front of her. She has this thought because she sees the cup. Periodically she performs actions involving the cup: she takes a sip. Let us suppose the conditions of well-functioning are satisfied, and hence that these actions are successful. We now read the content of the thought from a general truth about what would cause the thought in conditions of well-functioning, i.e., conditions under which actions caused by and rationalized by this thought succeed.

Cups on the table would cause the thought that there is a cup in front of her. But that isn't all that would cause the thought. Suppose we put a piece of cardboard between her and the cup. If she had not seen the cup beforehand, she might not think there is a cup in front of her. She might, though, if for some reason she

thinks there is a cup glued to the back of the cardboard.

The difficulty is that what would cause the thought in the conditions of well-functioning is whatever she takes as evidence that there is a cup there in front of her. There's some sense in which the evidence most likely to yield the thought that there is a cup in front of her is the presence of a cup. But we know this because we know what we think and why we think it. If our task were independently to confirm an hypothesis about what would cause this thought, there's no guarantee that it would be a coffee cup. Perhaps our subject always requires some independent confirmation of her perceptual beliefs before she acts on them.

Clearly knowing a lot about what else a person thinks would be a big help in deciding the content in those cases where the person isn't going on simple perception of the object of her thought. We might envision a causal theory of belief content that is holistic, one that takes a lot of information about the causes of representational states and somehow computes the content of each in the context of all the rest. The trouble is that the computation would surely beg the question whether the account really is naturalistic. How could we decide what the content of a particular thought is by consideration of the thoughts that yield it in any other way than to decide what the most reasonable thing for someone to think would be if she believed just those things?

I conclude that Stampe's content condition is neither necessary nor sufficient for the content of human thoughts. It's not necessary since what would cause the thought that there's a coffee cup in front of her might not be a coffee cup. It's not sufficient because even if that is what would cause a type of thought it might not be the thought that there's a coffee cup in front of her; she might be disposed to mistake coffee cups for small hats.

3.7 Conclusions

Stampe's causal theory of representation is clearly flawed in several ways. Yet it captures something interesting and important. One might have thought that there could not possibly be a way to understand content in causal terms. Despite the flaws Stampe shows that it is plausible that some causal relation underlies facts about content.

Stampe aims at a naturalistic account of content. He is not entirely clear what "naturalistic" means, so it is difficult to determine how well he succeeds. If function attributions are essentially interest-relative then he does not succeed. In sec-

tion 6 I argued that his structure is neither necessary nor sufficient for correct attributions of human thought contents, and that possibly the only way it might work requires a notion of what is rational to believe. Hence I suspect Stampe's own aims are not met by his theory.

Stampe's account does provide preliminary confirmation of my claims about content. We still have no argument that content is something of which we must be able to "give an account." But if we think we can give an account of it, we will find that items have content only if their presence and activities are explained in some way by reference to a benefit for the creature in which they occur. Stampe says very little about the notion of function, and apparently thinks there can be representation in very simple systems even without function. I think function is the minimum we need. My exposition of a survival-based notion of function and content showed how states of creatures are representations only if they stand in a certain explanatory relation to a good for those creatures. Finally I showed that to get an account of content like Stampe's to give the right answers about the content of human thought the account would need to use a notion of rationality and would need to find its subjects rational.⁶

⁶I am deeply grateful to George Myro for a series of conversations we conducted on an earlier version of this chapter in the Fall of 1987.

Chapter 4

Dretske's Theory of Content

4.1 Introduction

Suppose there were an account of content along the lines of Stampe's theory, one that uses the notions of function and of causation, but one that overcame the difficulties of Stampe's account. Suppose furthermore that the notions of function and causation were clearer, better understood, and broader in application than the notions of rationality and correlation that we use in interpreting one another. This theory would give a substantive explanation for why, when we interpret, we endeavor to see the other as rational (why we use the Principle of Charity). We might even be able to show why the Principle of Charity isn't really a necessary aspect of content, why, for instance, the thesis of content rationalism as described in Chapter 1 above is false.

I believe that Fred Dretske has such hopes for his theory of content.¹ In this chapter I give my reasons for thinking that these hopes will not be realized.

In the next section I'll present his theory. Since Dretske places great stock in the difference between evolutionary content (roughly the sort I described in Stampe's defense) and content that stems from learning (Dretske's central contribution to the contemporary discussion) I'll describe both theories and Dretske's

¹The theory is presented in *Explaining Behavior* (Dretske, 1988). More detail on the theory of content itself is given in his "Misrepresentation" (Dretske, 1986). I will also refer to two earlier works on information and information-based semantic theories: *Knowledge and the Flow of Information* (Dretske, 1981) and "Précis of Knowledge and the Flow of Information" (Dretske, 1983). I'll call these works 'EB', 'MR', 'KFI' and 'Précis'.

motivations for preferring learning-based theories of content. In section 3 I criticize these motivations at length: there are important differences between the two kinds of content, but Dretske's reasons for holding that the evolutionary account is inadequate are poor. In section 4 I question whether Dretske's proprietary notion of indication is really what we need in a theory of content. In section 5 I argue that the central obstacle in the way of claiming that the theory is naturalistic in the required sense is to make out a naturalistic account of value properties. Section 6 shows why the theory uses the notion of rationality when applied to the propositional attitudes: this is where I argue that Dretske's hopes for explicating the rationality of the propositional attitudes are empty.

4.2 Dretske's Theory

Dretske's aim is a naturalistic account of content that shows how "physical systems ... misrepresent the state of their surroundings" (MR, p.17). It should be naturalistic in the sense that it does not appeal to our own powers of representation; i.e., it should be an account of a *non-derived* capacity for representation. We are physical systems, so the account should be extendable to cover human language and thought. I assume that Dretske's aim would not be satisfied if he were forced to use the notion of rationality in the account; that would clearly undermine the ambitious hope sketched above.

The central relational concept in Dretske's theory is that of indication. Indication is a natural dependency relation between particular events (MR, p.20). It is modeled on Grice's notion (Grice, 1957) of natural meaning, or meaning_n; it is the relation we report when we say that *that* smoke means fire, or that *those* storm clouds mean rain. By definition there can be no misindication; if one event *e* indicates another, *c*, then the indicated event *c* must exist.

Indication is not simply a relation between two particulars, however. Some aspect of the indicating event indicates an aspect of the indicated event. The aspect of the indicating event must depend on the aspect of the indicated event. The dependency must be grounded in some causal or nomic relation. (Dretske's account is not, therefore, strictly a causal account of content, although most indicators are related to what they indicate by causation.) This relation must support counterfactual assertions, to the effect that if the indicated event had differed in certain ways the indicating event would be different.

Finally indication is a *necessitating* relation, in the sense that one event's being

F is not an indication of another event's being G if being G is not necessary for being F. Consider the spots on a child's face. These are an indication of measles if measles is the only thing that would lead to the spots. If the likelihood that the child has chicken pox is quite close to the likelihood that she has measles then the spots do not indicate measles. (They do indicate the disjunctive property of having chicken pox or measles.) The necessitation need not be strict: the spots on a child's face might indicate measles even though it is possible for an Evil Demon to place such spots on a child and leave her completely healthy. Dretske offers another example (MR, p.34; EB, p.58, note 3). Suppose a light is controlled by two switches wired in parallel. Then the event of the light going on does not indicate that the first switch was closed, even though the closing of the first switch causes the light to go on, since the light could have been made to go on by closing the second switch.²

Dretske claims (MR, p.19, note 3, and EB, pp.58-9) that the natural meaning of an event (what it indicates) is essentially the same as the information it carries, where by information he means the notion he described in KFI. If a signal is an event then

A signal r carries the information that s is F =_{df} the conditional probability of s 's being F, given r (and K), is 1 (but, given K alone, is less than 1). (Précis, p. 57, KFI, p.65)

K stands for knowledge we might have that would alter the information a signal might carry: this knowledge limits the range of possibilities left open by the occurrence of r and hence changes the information that r carries. For a naturalistic account K must be empty. Information is defined in terms of the notion of channel conditions, which limit the possibilities relevant to the conditional probabilities: "If every logical possibility is deemed a possibility, then everything is

²Dretske's relation of indication is close to the causal relation Stampe invoked in explicating reference, but there are differences. Indication need not be causal. Stampe's relation is necessitating in the same way as indicating is necessitating: "Ordinarily, if O's being F causes R to be phi, R is phi only because O is F, and R wouldn't be phi, were it not for the fact that O is F" (p.86 of "Toward a Causal Theory of Linguistic Representation"). He also uses the phrase "normally, if not necessarily" in place of "ordinarily" (p.85). Hence Stampe's relation has the vagueness of an unexplained ceteris paribus clause. This vagueness is shared by Dretske's account in as far as the necessitation of indication need not be strict. Stampe does not use such a necessitating relation in explicating content: what a representation is about is what would cause it in certain conditions, even if it could well be caused in other ways as well. This last difference is important in evaluating Dretske's theory; see section 4 below.

noise" (Précis, p.60). This relativization corresponds to the less than strict necessitation that indication demands.

Indication may be related to meaning in some way, but the relation must be distant. Meaning essentially involves the possibility of falsity: if *S* means that *p*, then unless *S* is a logical truth it is possible for *S* to be false. Put slightly differently, representation depends on the possibility of misrepresentation; the trouble with indication is that there is by definition no such thing as misindication, no such thing as an event's indicating that *c* is *F* when *c* is not *F* or *c* did not occur.

The bridge between indication and representation is made with the concept of function. Dretske offers an intuitive justification for this idea. Many artifacts are devices that are designed to indicate something, or devices whose function is to indicate something. A thermometer is a device that we construct so that there is a certain dependency relation between one of its aspects (the height of a mercury column, for instance) and the ambient temperature, such that any particular reading of the thermometer should be an indication of the temperature. Occasionally the reading of the thermometer fails to indicate the temperature, when, for instance, the bulb is irradiated with laser light. The thermometer continues to indicate lots of things. But it also now misrepresents the temperature. (How exactly we understand this claim is not particularly important; what is important at this point is that there is some familiar notion of misrepresentation and function according to which this claim is correct. I think there is: it is literally true that the broken scale on the IRT subway platform at 42nd Street and 7th Avenue misrepresents my weight.)

We need a way to discharge the appeal to derived functions: we need a notion of function which applies to things in their own right rather than one which depends on our interest in one particular way a thing is related to other things. Dretske's first suggestion is that we should look to evolution and adaptation for a "non-derived" notion of function and the possibility of a non-derived account of content:

The obvious place to look for natural functions is in biological systems having a variety of organs, mechanisms, and processes that were developed (flourished, preserved) *because* they played a vital information-gathering role in the species' adaptation to its surroundings. An information-gathering function, essential in most cases to the satisfaction of a biological need, can only be successfully realized in a system capable of occupying states that serve as natural signs of external (and some-

times *other* internal) conditions. (MR, p.25)

Dretske introduces the now-famous magnetotactic bacteria. These organisms have a magnetosome, an organelle that contains tiny bits of a magnetic material. This organelle, in tending to orient itself with the earth's magnetic field, tends also to rotate the bacterium. In turn, because of the way the organelle is situated, this rotation tends to make the bacterium propel itself toward magnetic north. Since these bacteria live in the northern hemisphere, magnetic north is, relative to the ocean's surface, *down*; swimming down, as it happens, is a good thing for these bacteria, since oxygen is toxic to them, and swimming downward into the anaerobic mud at the bottom of the ocean gets them into the environment they need.

The magnetosome indicates many things: among others, the direction of magnetic north, the direction of the north pole of the Earth, the direction of anaerobic water. In the life of the bacterium it functions to indicate the last, since that is what the bacterium needs. Intuition suggests that the magnetosome represents the direction of anaerobic water. This intuition is bolstered by the fact that if bacteria from the North Atlantic are brought to the southern hemisphere, where the relation between magnetic north and anaerobic water is reversed, they move upwards, into oxygenated water, and they die. They are making a mistake; the magnetosome is misrepresenting the direction of anaerobic water.

The biological notion of function is not enough, however. Dretske claims that the function of the magnetosome is indeterminate (MR, p.28-32, EB, pp.64-70). There are, I think, two reasons Dretske offers for this claim. First, an item that has a function may fail to perform the function while functioning as well as it can. A can-opener may be in top condition yet unable to open a given can, since as it happens the can is made of quarter-inch steel plate. Suppose we hold a bar magnet near one of the bacteria and get it to swim in the wrong direction. The magnetosome is doing precisely what it is supposed to do (align the bacterium with magnetic north). If there is a teleological failing it lies outside the bacterium.

Second, Dretske holds that the way to determine what an indicator functions to indicate is to ask *how* the indicator performs this function:

given that a system needs *F*, and given that mechanism *M* enables the organism to detect, identify or recognize *F*, *how* does the mechanism carry out this function? Does it do so by representing nearby *F*s as *nearby F*s or does it, perhaps, represent them merely as *nearby G*s,

trusting to nature (the correlation between F and G) for the satisfaction of its needs? To describe a cognitive mechanism as an F -detector (and, therefore, as a mechanism that plays a vital role in the satisfaction of an organism's needs) is not *yet* to tell the functional story by means of which this mechanism does its job. (MR, p.31)

The answer to such a *how*-question appears to depend on what is causally relevant³ in the transaction between environment and indicator. Causal relevance is different from indication in several ways. The salient difference here is that causal relevance involves stricter regularities than indication. What makes the magnetosome twist is the magnetic field. It does this whether or not there is oxygenated water around. It would not be made to twist by the direction of oxygenated water if there were no magnetic field present.

Dretske calls this the problem of indeterminacy. We might conclude from Dretske's reasons that there is no fact of the matter what the function of the magnetosome is, hence that the function is indeterminate.⁴ The real problem is not indeterminacy, however. The causally relevant factor that makes the magnetosome twist is perfectly determinate: the direction of magnetic north. Hence the function is perfectly determinate. But the magnetic field is pervasive (on earth, anyway), so the magnetosome is always indicating just what it functions to indicate. Hence so far we have not shown how the magnetosome misrepresents anything:

If we describe the function only in the latter way [indicating the direction of the surrounding magnetic field] it becomes impossible to fool the organism, impossible to make it misrepresent anything. (MR, p.32)

as long as there remains this indeterminacy of function, there is no clear sense in which misrepresentation occurs. Without a determinate function, one can, as it were, always exonerate a RS [a representational system] of error, and thus eliminate the occurrence of misrepresentation, by changing what it is supposed to be indicating, by changing what it is its function to indicate. (EB, p.69)

³See Chapter 8 below for the outlines of a theory of causal relevance.

⁴This is essentially the stance that Dennett takes in a paper that exercises Dretske: "Evolution, Error and Intentionality" (Dennett, 1987a). The point is taken up, though to a somewhat different end, by Jerry Fodor, in a recent manuscript; I address Fodor's use of it below, Chapter 6, section 6. As I'll argue below in section 3, and in response to Fodor in Chapter 6, I believe this whole line of thought is completely misguided. I am grateful to Elisabeth Lloyd for helping me to sharpen my response.

In MR Dretske holds that this is a conclusive objection to the claim that an evolutionary or adaptation-based account of function gives an account of representation. MR offers a different account of function, to which I will turn in a moment. EB claims that this is a serious objection, but then goes on to agree that the evolutionary account can be seen as an instance of the MR solution, albeit with an important difference. The problem is one that Fodor labels, in its various incarnations, the “disjunction problem” (Fodor, 1987, 102): unless we have a principled and naturalistic way of distinguishing among the variety of properties that are causally relevant to the production of a type of representation, we are forced to say it is about the disjunction of all of them if it is about anything at all.

The MR solution to the indeterminacy problem is an elaboration of a suggestion already present in KFI: that learning imposes function on things. The simplest sort of system that displays the sort of functioning we need has a complex connection between its sensory structures and its behavior. (Following Dretske we allow for kinds of behavior that are not caused by reasons, so as not to beg the question how we are describing the movements of systems.) Suppose it has a behavior B (a kind of behavior; all the schematic letters in this sentence stand for kinds), a need N, a series of sensors (indicators) I_1, \dots, I_n , a central state R which causes B, causal links between instances of the I_n and R, and finally the capacity to alter the causal relations between the I_n and R. The process of making such alterations is influenced by the “success” the system has had in satisfying its need N by issuing B's when R's occur.

The capacity to alter the connections under pressure of need is crucial. To show this, we consider a creature whose connections do not change. It has series of indicators that cause the central state R, rather than the single indicator the bacterium has. We avoid the simplest sort of problem of indeterminacy of function that came up with the bacterium, but we certainly don't make the problem go away (MR, p.34). Suppose this creature, too, needs to be in anaerobic water, and has a magnetosome, but also has some sort of lack-of-oxygen indicator. The central state R cannot indicate magnetic north, since R will also occur whenever what suffices for the lack-of-oxygen indicator to occur is present. But R will still occur, without malfunctioning, whenever either sensory indicator occurs, regardless of what it is caused by. Hence the answer to the question *how* it functions is still to be given in terms of the disjunction of these surface states. Hence its function can only be to indicate that disjunction, and so it still cannot ever misrepresent the direction of oxygen-free water.

A solution to this problem must entail that there is no disjunction of surface

states by means of which R functions to indicate F. Suppose that the creature can learn, in the sense that it is capable of altering the connections between R and its senses in order to improve its rate of success. Through time the system attaches further F-indicators to R, and perhaps detaches sensory states that no longer indicate F.

At any point in time there is a range of things in the world that will produce R, some of them F-things, some of them not. As usual, at any point in time there is an involved disjunction of proximal stimulus states which is guaranteed to have occurred if R occurs. If R has a function it functions by being caused by an instance of this disjunctive property. As usual, therefore, at any point in time, R indicates F and this surface state, among other things, and it functions to indicate only this surface state. Hence it does not represent F.

But since this system learns, it is false that R, *through* time, without malfunctioning, functions by being caused by any one of these disjunctions of proximal stimulus states, since the disjunctions are changing. Each alteration is performed by adding or deleting an F-indicator, and is performed in the service of satisfying need N. This, according to Dretske, implies that R has an indicator function; and the only function it could have through time is to indicate F. So when it is caused by something other than F, it misrepresents it as F.

We have now given a non-evolutionary account of the function of an internal state like R, and we've shown how the set of properties it functions to indicate is a subset of the set of all properties that could produce it; hence we've given a solution to the problem of the indeterminacy of function.

4.3 Indeterminacy

Dretske claims that evolution provides only indeterminate characterizations of function, while the learning account does better. In this section I criticize his reasons for this claim. Before I begin I want to mention a reason I think is a good one for preferring the learning account. Clearly not all our representational powers stem from our evolutionary heritage. We think about the U.S. Constitution, about coal, about our computers. These things were not on the scene when we evolved our present structure. And clearly we do gain representational powers by learning about things. Hence a learning based account of content has much greater promise to tell us something about ourselves than an evolutionary theory.

In presenting Dretske's theory I located two reasons he offers for his claim

that biological functions are indeterminate. Neither is good.

There are two ways to approach the function of a part of an organism, depending on the breadth of our focus in describing it. We may concentrate on the part of the organism itself, and look for a characterization of the aspects of the part that permit it to perform the function. Or we may concentrate on the conditions required for function performance, whether these lie within the organism, or must be sought in the environment. I noted above in Chapter 3, note 4, that writers on function and content divide on this issue: Stampe and Millikan take the broader view, while Fodor and Dretske take the narrower.⁵ Dretske's first reason for thinking that the function of the magnetosome is indeterminate, that the magnetosome may orient the bacterium in the wrong direction while all along functioning perfectly well, clearly stems from the narrower view.

The trouble is that there is no reason to prefer the narrower view. The definition of the notion of function is up to us as theorizers about content. We are not forced to define the notion of function in such a way that we end up with the problem about indeterminacy. We can instead characterize the notion of function in terms of what is needed for function performance. In the case of the magnetosome function performance depends on the bacterium residing in anaerobic water, and the magnetosome works toward this end; we should, I think, say its function is to indicate anaerobic water.

Dretske's second reason concentrates on *how* the function is performed. I agree that the magnetosome performs its function by aligning with the lines of magnetic force. That is what is causally relevant to what it does. But, again, we should not let this fact dominate our characterization of the function of the organelle.

F is not, by hypothesis, causally relevant to R. (If anaerobic water were the causally relevant factor controlling the bacterium's sensorium, there would be no problem of indeterminacy.) F always controls R through some other properties. No matter how complex and changing the causal link between them, F is never causally relevant to R. F can therefore never be the answer to the question how R performs its function. So the learning account at best entails that R has *no* indicator function, since there is no single answer to the question how R indicates F.

Dretske claims there is no single answer, but I think he is wrong. Since some-

⁵Millikan shows how important it is to take the broader view in her "Biosemantics" (Millikan, 1989a).

thing other than F causally controls R there must be some property that is causally relevant to the production of R. The fact that R's relations to F are mediated by different things does not show that there is no such property. At any moment there is some disjunction of states that is causally relevant to R, and no one of these disjunctions remains in causal control of R through the life of the creature. But surely there is a longer disjunction that does: the disjunction of all the momentary disjunctions. This disjunction may seem too heterogeneous to count as an answer to *how* R performs its function; it doesn't seem to be a properly explanatory property. We can do better: suppose we have a true theory that describes the conditioning capacities of such a creature, one that specifies precisely what kinds of relations it *can* establish between its sensory surfaces and its representational states. This theory entails an exhaustive specification of how R performs its function.

The problem Dretske set for himself was to give a "*principled* way of saying what the natural function of a mechanism is" (MR, p.32). It seems to me that Dretske's solution, if it worked, would provide a principle, but one of the wrong sort. The problem was that there is always a determinate disjunction of surface states that is more causally relevant to the representation than the environmental state. The solution is to find a way to get rid of the determinate disjunction. But the solution looks to something that seems to be irrelevant to an understanding of content. If we wanted a relation that is satisfied only by the representation R and the environmental state F, surely there are many such relations. But virtually all of them have nothing to do with what R might mean.

The virtue of Dretske's solution lies elsewhere. It makes explicit something he left implicit in his description of biological function: something has a function only if there is something it does that is best explained by reference to a good it may yield. The learning account of indicator functions says that something has an indicator function provided there is an explanation for what it does that connects the occasions on which it indicates with a good it sometimes yields. A function for the state is determined by considering what controls the state that yields a benefit.

If we regard the solution to the indeterminacy problem in this light we can see why there was no problem to be solved in the first place. The function of the magnetosome is to indicate anaerobic water. This follows from the fact that there is an explanation for why the magnetosome is situated in the bacterium, and hence controls the movements of the bacterium, in the way it does, that depends on the relation that its placement and effects and anaerobic water have to a good for the bacterium.

I think Dretske agrees with me on this point in EB. He raises the indeterminacy

problem, but never describes a solution, and ends up conceding that evolutionary function is sufficient for representational content after all:

If we suppose that, through selection, an internal indicator acquired (over many generations) a biological function, the function to indicate something about the animal's surroundings, then we can say that this internal structure *represents* (or *misrepresents*, as the case may be) external affairs. This is, in fact, a representation of Type III [i.e., a representation the function of which is not derived in any way]. But it is *not* a belief. (EB, p.94)

The real point of introducing the notion of learning into the account of content is to make the distinction between representations and beliefs.

This distinction involves another poor motivation for the distinction between two kinds of content. Dretske holds that if content is a relational property of things that have more intrinsic explanatory properties, then there is a serious problem whether content is a causally explanatory property. The trouble comes with the consideration that items with the intrinsic properties are able to produce much the same kinds of effects whether or not they have the relational properties. One central concern of EB is to show how content really does explain something. The solution depends on the structure of control that the learning account requires. The inner representational state R has the content that something is F in virtue of the fact that R's naturally meaning F, i.e., the fact that R indicates F, causally controls the fact that R controls behavior. In this sense the meaning of the representational state is causally engaged in what it does. The reason evolutionary content is not causally relevant, according to Dretske, is that what an evolutionary representational state indicates does not control what it does. Rather the explanation for why the representation causes behavior is the genetic material which structures the creature.

I will discuss the problem of causal relevance at length in Chapter 8 below. I am confident that the consideration about relatively extrinsic and relatively intrinsic properties has no important bearing on causal relevance. If there were a problem, though, Dretske's solution is little help. The problem should be whether what I believe, i.e., the content of my thoughts, has any causal relevance to what I do. Dretske's solution is to point to a way we might explain the causation of my actions by my beliefs that depends on what *other* beliefs have *indicated* (not on what their content was).

There is one other thought that appears to motivate Dretske in working on the difference between selectional and learning-based content. In MR he writes,

I have confined the discussion to simple organisms with primitive representational capacities. It is not surprising, then, to find no clear and unambiguous capacity for misrepresentation at this level. For this power—and, presumably, the dependent capacity for belief—requires a certain threshold of complexity in the information-processing capabilities of a system. (pp.32-3)

Clearly we are extraordinarily complex physical systems, and anything that can do the things we do will be similarly complex. I think that the considerations of this section show that complexity has little or nothing to do with semantic powers. What matters is an explanatory relation to a benefit that a system may yield; very simple systems can satisfy this relation.

4.4 Indication

Indication is an explanatory relation on two particulars and two properties possessed by the particulars. In this section I consider two kinds of objection to Dretske's theory that concentrate on this relation. The first kind looks at the explanatory notions that are required, like the notions of law, counterfactual, etc., and objects that their use entails that content is derived, in Dretske's sense, after all. I do not believe this is a good objection. The second kind of objection allows Dretske his basic explanatory notions and shows why there doesn't seem to be reason to prefer indication to other, more familiar relations. I believe the second objection is a good one, though not telling to the spirit of Dretske's theory.

Representation, as Dretske describes it, depends on two explanatory relations: indication, backed by nomic relations, and the relation between an indicator and the causation by a representation of a bodily movement. The most objective explanatory relations to which we might appeal in articulating such a theory are the (more or less) strict laws of physics. It is clear that we will not be able to do this, however. The explanations Dretske cites as examples involve very contingent, local regularities. They appeal to counterfactuals about what creatures would do in otherwise normal conditions. We might worry that the reason Dretske's proposals have any plausibility at all is that this completely unprincipled appeal to natural

generality begs all the important questions. Perhaps what is normal is tacitly determined by something we know about content. Or perhaps the only way to render the generalities explicit is to insert a place for explanatory interests.⁶

I do not think this is a good objection to Dretske's position. The trouble is that it is far too general. Suppose we are convinced by Nelson Goodman's work that lawlikeness is interest relative in some important way. That would certainly show that we cannot give a theory of content in Dretske's form without a tacit appeal to what we think is interesting. But the same point goes for absolutely everything we say about the world, from the most subjective psychological theorizing all the way to physics. To make the objection apply in an interesting way to Dretske's theory we should have to be able to demonstrate that the theory could not entail correct consequences about content even if it used explanatory generalizations gathered with no thought to content; and I do not think we can give such a demonstration.

A much better objection asks why indication, rather than causation or correlation, should be the right concept to use in analyzing content. I'll begin by trying to clarify what relation indication is, then go on to show that it is *not* the right notion.

Indication is logically a somewhat motley relation. Consider the relation that makes for content. An internal state R has the content that an object is F provided there is an explanation for why R causes bodily movements that depends on the fact that R indicates F. Internal states that cause bodily movements like this must not persist through the life of the organism, otherwise the organism would be unable to fail to produce the bodily movement. Hence the letter 'R' must stand for a type of state.

What does it mean, though, to say that R indicates F? Not, clearly, that every instance of type R indicates something of type F. That would make misrepresentation impossible. Could we say that R indicates F if some of its instances indicate things that are F? We can, provided we read "some" as weakly as possible. Consider Venus Fly Traps, which, I'm told, snap shut on many more non-nutritive things than flies. Some biological indicators are very reliable, some are very unreliable. There is no lower limit to how unreliable they may be and still provide the resources for representation. All representation requires is some benefit which explains the causation of bodily movements by the inner state.

So the fact that R indicates F is simply the fact that at least one R indicated an

⁶See (Loewer, 1987), especially. pp.305-7 for a such an objection to Dretske's work. The objection has been urged in still more general form by Hilary Putnam (Putnam, 1983, 1986).

F; in this sense indication is not a general property of things.

My way of understanding indication comes from the way Dretske characterizes the notion in MR and EB. I do not think his own way of characterizing it is consistent, however. He holds that the relation between indication and information is as follows:

if S (sign, signal), by being a , indicates or means that O is A , then S (or, more precisely, S 's being a) carries the information that O is A . (EB, p.59)

Information is defined as follows:

A signal r carries the information that s is F =_{df} the conditional probability of s 's being F , given r (and K), is 1 (but, given K alone, is less than 1). (Précis, p.57, KFI, p.65)

The standard interpretation of a conditional probability is in terms of relative frequencies; on this interpretation, these letters must be understood as denoting classes of events. Suppose then there is a signal of type R and a type of state of the world F . If the signal carries the information that something is F then if any instance of the type R occurs an instance of the type F must occur as well.

If this is the right way to understand information and probability, then, if a signal carries the information that s is F , the instances of the signal surely indicate that s is F . Dretske makes the converse claim, that indication entails information. But this claim confuses a particular relation with a general relation. Suppose S 's being a indicates that O is A . We cannot infer anything about what information any class of events to which S belongs might carry, since we have only a single case at hand. We could imagine a theory of probability that shows how to support claims about probabilities based on single cases; the chances for such a theory are not good.⁷

Indication is compatible with any conditional probability whatsoever except 0. Suppose on a single occasion an event e occurs which has the property P , and there is an event c which has the property $P1$. Furthermore there is some nomic or causal truth governing these two events such that were c not $P1$ then e would not be P . Then e indicates c . Suppose furthermore that this is the single occasion on

⁷Cf. (Armstrong, 1983) pp.32-35.

which a P1 event causes a P event. Then the conditional probability $\text{prob}(P1/P)$ may be as small as we like, but not 0, depending on how large the class of P events is.⁸

Peter Godfrey-Smith has argued,⁹ I believe persuasively, that if we understand indication as a statistical notion it is not explanatory in the ways Dretske needs for a theory of content. I do not think indication is a statistical notion at all. The spirit of Godfrey-Smith's objection remains, however. Indication is not necessary for content. Recall that the explanatory component to the notion of indication is a "necessitating" dependency. Consider Dretske's example of a signal that is not an indicator: a lamp is connected to a power source via two switches connected in parallel. Either will turn on the lamp. The lamp's going on does not indicate that either switch is closed; it indicates only the disjunction that one or the other switch is closed. Consider now a creature that is otherwise capable of natural, non-derived representation; suppose it has an instance of this sort of device in it. Suppose there is some difference between the two switches. The difference is causally (nomic, counterfactually) irrelevant to the lamp's going on. But it is not teleologically irrelevant to the creature: if the first one is closed, the creature benefits (sometimes), while it never benefits if only the second is closed. This benefit explains why the lamp makes the creature perform some bodily movement: it receives the benefit often enough to make acting on the lamp signal worthwhile. Finally, suppose that the difference is otherwise indiscernible to the creature: for some reason or other neither it nor its species is capable of sensing the difference other than by the benefit it yields.

I suggest the lamp represents that the first switch is closed, even though it does not ever indicate that the first switch is closed. The same point goes for information: Dretske claims (MR, p.25) that an information-gathering function is essential to the life of a species. What is essential is getting what's needed often enough, whether or not the means to this is information.

There remains a causal and nomic relation between the closing of the switches and the lamp's going on. I do not think even this kind of dependency is required for content. (Dretske does: EB, p.57.) Suppose there is an accidental correlation between an environmental state and an inner state of a certain creature. It could

⁸I think that statements about relative frequencies are logically independent of statements about laws, causation, and counterfactuals. Obviously there is some strong connection between them, since the evidence we use to make statements of the latter sort is almost always statements of the former sort.

⁹In "Indication and Adaptation", forthcoming.

happen that this correlation explains why instances of the inner state cause bodily movements that sometimes yield benefits for the creature, and explains it in just the way Dretske's theory requires. I see no reason why we should not call this state a representation of the environmental state.

4.5 Needs, Life, Value

MR shows how a certain kind of content depends on an explanatory relation that involves the satisfaction of a need. EB relativizes content to success or to benefit. Dretske's stated aim in formulating a "naturalistic" account of content is to show the possibility of representational powers that do not derive from our own representational powers. In this section I consider whether the appeal to needs, benefit or success will compromise Dretske's aim. I show that the issue reduces to the question of the nature of value properties.

A non-derived need is a need something has that it would have regardless of what anyone thinks it needs. What is required to have a non-derived need? One sufficient condition is life: a living creature needs those things without which it cannot live. The bacterium needs to be in anaerobic water; the Venus Fly Trap needs to capture a fly now and again.

Clearly the fact that something is alive does not depend in any way on our representational powers. So far, then, Dretske is on safe ground. But there is reason to think that the concept of life has a very strong connection with Dretske's concept of representation. Consider viruses. A viral particle contributes to the survival of its strain by encountering an appropriate cell and injecting its genetic material, thus beginning a sequence of events that should lead to the production of many more viral particles. The viral particle must "recognize the opportunity" to do this: some part of the particle responds to a cell-wall molecule. Presumably a virus could be fooled: caused to extrude its genetic material by something that will not provide the cellular machinery that is the selectional reason for what the virus does. It seems, then, that even the simplest forms of life have extremely simple representational powers.

Dretske is still on safe ground. Even if the concept of life analytically involves the concept of representation we can apply the notion of life without knowing anything about the representational powers of a thing. Further, life is sufficient but perhaps not necessary for needs. Perhaps it is possible for there to be an object which is not alive but such that there is something that it does which may or may

not be successful (or which may or may not yield a benefit); then this object needs the conditions required for success or benefit. (An intelligent computer would be one such object.) The pressing question then is whether there are such things as non-derived success or benefit.

The same question arises when we look more closely at the notion of life. There are relatively mechanical criteria to demarcate the living things from the rest: living things are capable of self-replication and can interchange material with their environments while maintaining a certain structural integrity. These criteria do not seem to make up a definition of life, however. I think it is clear that they are not sufficient, since there can be physical systems, like crystals, that meet them and which are not alive. Nor are they clearly necessary.

I suspect the right analysis of the notion of life shows it to be a teleological notion: something is alive only if there is an explanation of what it does in terms of some end it has. We might attempt an analysis of the notion of an end solely in terms of a certain causal structure. I doubt this analysis would succeed, because it seems that any causal structure we might describe could be realized by something that has no end. If something has an end then there is some good it can realize by achieving that end.

Hence whether a representational system is alive or not, representation depends on there being some good for it. Is what is good for a thing a non-derived property? Answering this question requires an account of value properties. If we think that value properties must derive in some way from human agents (if, for instance, we claim that a thing is good just in case we think it is good, or if we feel a certain way about it), then goodness is not a non-derived property. The alternative Dretske must take, then, is to hold that value properties are real, mind-independent, non-derivative properties of things.

This is not the place to embark on the project of determining the prospects of such a position. I suspect it would be extremely difficult to articulate and defend, but I have no reason to think it could not possibly be done.

4.6 Applying the Account to Propositional Attitudes

Dretske's account of content is designed around fairly simple creatures, like the bacterium with its magnetosome, and things that satisfy the learning account, which connects representational states directly to sensory indicators and action. But mental representation is involved in inference and practical reason in all but

the simplest creatures that have intentional states. Dretske's aim is to account for representation in terms of indicator functions. Can we stretch the account so far offered to meet this demand for greater complexity?

We need a generalization of the idea behind the learning account of representation. Suppose we have a system with a series of surface indicator states causally connected to a sequence of more central states, an array of needs, and some sort of mechanism that computes a behavioral output given combinations of the intermediate states. Then one of these intermediate states functions to indicate that *p* provided that the behavioral output it produces yield some benefit when *p*, and this fact explains why it produces the behavioral output.

The explanation must be of a special sort. It cannot, for example, run through wildly irrelevant external conditions. There could be a creature for whom each behavioral success leads to a tree-branch breaking which leads to a thump on the head which leads to a cognitive reorganization, all as a matter of complete accident. (There could be a similar sort of set of internal accidents.) The adjustment needs, rather, to be one "of the right kind." Characterizing this kind will make an essential appeal to the concept of rationality.

Consider inductive inference. I look at the sky at evening and smell the air and think of the season and decide to put the fly on my tent; I've decided it will rain tonight.

The belief that it will rain tonight indicates the fact that it will rain tonight just in case it will rain tonight, and if it were to fail to rain tonight I would not believe that it will rain tonight. That the belief indicates the fact can explain why the belief causes actions; the explanation will show how further beliefs are more likely to cause similar actions once I discover whether I was right.

But what does the explanation actually look like? What led me to the belief in the first place was a collection of other beliefs. Suppose I turn out to be right; this fact will increase my confidence in some or all of the inferential moves I made in reaching my judgement. Where, precisely, my confidence improves will depend on what I thought to begin with: perhaps I was in doubt about the appearance of the sky as a sign of rain, but now I'm more certain. Or perhaps somewhere at the back of my mind I have reason never to trust the appearance of the sky, and so my estimation of the smell of the air goes up. Then if a similar smell occurs, and other circumstances, both in the world and in my cognitive state, are appropriate, then I may make a similar decision, and perhaps more quickly or confidently.

The trouble now is that we cannot describe this process without making use of

the notion of epistemic rationality. There are many ways I may respond to the truth of such a belief that are relevant to the fact that the belief has the content that it will rain tonight. Some will actually reduce the reliability of my judgments that it will rain, or reduce the likelihood that I will act the way I do tonight. Equally there are many things that might happen to me as a result of the truth of the belief that would alter how I came to it and how it made me act, that would be irrelevant to its content. The indicator function of such a belief depends on how its indicating what it does controls what bodily movements it produces. In the simple account the control was quite straightforward: the indication caused the representational state to cause a bodily movement. For people the control can only be described this way: receiving confirmation or disconfirmation of a belief may produce revisions to the implicit theories which generated the belief, and the revisions are ones that make sense of everything the agent believes.

We can cast the point in a slightly different way. Dretske has little to say about what the control relation should be like, relying instead on examples: training neural networks, operant conditioning, etc. Suppose we had a characterization of the control relation that was naturalistic in the sense Dretske specifies and furthermore does not simply use the notion of epistemic rationality. Then the point would be that any such relation between the surface of an agent and a particular belief is neither necessary nor sufficient for the content it has. It is not necessary, since a belief may have the content that *p* by way of its connections with other facts than *p*. It won't be sufficient, because inferential connections with other beliefs may override the control relation in determining what the content is.¹⁰

It may be objected that use of the concept of rationality is sufficient for describing the control relation, but it is not necessary. I concede I have not demonstrated that it is necessary. I do not know how to demonstrate this, but I can show where two suggestions go wrong.

First, assuming a fairly standard sort of materialism, every propositional attitude has a physical description. Then clearly there is some way to describe the control relation without using the notion of rationality: we just describe the actual

¹⁰This objection is similar to the Putnam-Loewer objection I sketched (and rejected) in section 4 above: some explanatory relation Dretske relies on cannot be characterized within his naturalistic constraint. The difference between the objections hinges on the kind of content to which they are directed. If value properties are naturalistic (and for all I have said they may be) then there may well be an explanatory relation "of the right kind" for very simple representational systems that can be characterized naturalistically. Even if there is, though, the "right kind" of explanatory relation for creatures capable of practical and theoretical inference can only be described using the concepts of rationality.

causal connections. We might even find that there is a projectible causal relation: whenever, as a matter of law, someone believes that it will rain tonight a certain set (with one physical description) of causal transactions take place around this belief. (This is, of course, fantastically unlikely.) The trouble is that this suggestion gets the order of explication backwards. It begins with characterizations in terms of content and finds nomic correlations. But Dretske's theory needs to characterize a nomic relation that enters into a characterization of necessary and sufficient conditions for content, without presupposing facts about content.¹¹

Second, some forms of rational control can be given mechanical descriptions: we can describe a sound deductive inference in a particular formal language simply in terms of the shapes of the sentences. But this is not the general case. We do not have, and we are not likely to get, a syntactic theory of sound inductive inference or of practical reason.

4.7 Conclusions

Dretske aims at a naturalistic theory of content: an account that does not derive content from our representational powers. His theory generalizes the idea behind an appeal to evolutionary function: a type of state comes to have a certain content by instantiating a certain explanatory relation to environmental conditions, behaviors, and, most important, a good for the system. I prefer to regard Dretske's learning-based account as a generalization of, rather than an alternative to, the evolutionary account, since both accounts have essentially the same structure. I noted that Dretske's proprietary notion of indication does not seem to be required by the account. Finally there is a fundamental and unsettled worry about the source of value properties. These worries aside, I think Dretske's theory shows how to understand the representational content of very simple creatures.

It does not, however, show how to understand the content of our own thoughts and utterances without using the concept of rationality. I showed this by showing that there is apparently no way to describe one of the central explanatory relations (the way indication controls the causation of bodily movements by beliefs) without using, for example, the concepts of epistemic rationality. I conclude that Dretske's account does not show how thoughts can fail to be rational, or that their rationality can be understood in terms of a better-understood concept like that of function; rather it shows how any theory that we are confident is a theory of con-

¹¹George Myro helped me to get clear on this point.

tent for us shows that content is rational, and it shows this by using the concept of rationality.¹²

¹²Thanks to Ernest Adams, Fred Dretske, John Heil, Victoria McGeer, George Myro and especially Elisabeth Lloyd for reading an earlier version of this chapter.

Chapter 5

Biosemanantics

5.1 Introduction

So far we've considered the details of two causal accounts of content, Stampe's and Dretske's.

Stampe's account shows vividly that some set of notions in addition to causation is needed if we are to get a definition of content that looks remotely plausible for understanding human thought and language. He suggests the notion of function, but leaves it unclear why that notion is the right one, and leaves the account of function completely open. I argued that for systems of representation like our own it is impossible to state a correct causal theory of representation of the sort Stampe describes without using some notion like good reasons to believe, i.e., that the notion of function required is equivalent to the notion of rationality.

Dretske's account uses a much more subtle notion of function and a much more subtle conception of the causal relations necessary for content. I argued that despite this subtlety it still begs the question at issue here: any sophistication of the account that works for human thought requires a conception of the function of a thought which cannot be articulated without the use of notions like good reasons to believe. I argued also that his account of function requires an appeal to the notion of needs and the notion of life, and hence it is not naturalistic to the extent that these notions cannot be naturalized.

The state of my argument then is that these two causal accounts of representation do not show how content rationalism is false; if anything they show why content rationalism is true, since they show how we are compelled to use some

notion of rationality in understanding content.

Ruth Garrett Millikan's "biosemantic" approach aims to give a naturalistic account of content using a biological account of function. She argues that she can solve the problem that particular thoughts don't have functions that can be described without begging the question.

In this chapter I consider the details of the biosemantic approach. It has many failings, some of them familiar from Dretske's account. It does not, I argue, solve the problem about the function of particular thoughts. On the other hand Millikan does not beg the question I am interested in: she is quite clear that anything with content has very complex normative relations to various parts of an organism and things around it, and hence that a theory of content can only be stated using at least a primitive notion of rationality.

I begin with the overall structure of my discussion of Millikan's view. There are two conditions on the adequacy of a biosemantic approach to content. First, a notion of function must be given and defended that shows function to be a genuinely natural and biological aspect of things. Second, the notion must be shown to apply to the content bearing items in which we are most interested: human language and thought.

I argue that Millikan's account satisfies neither condition. We simply cannot tell from her account of function whether it is a naturalistic notion or even a biological notion. Even if her notion of function is a fully acceptable notion of biological function used in the biological sciences Millikan does not succeed in showing that the meaning and content of human language and thought are determined by function.

The organization of the rest of this chapter follows this argument: in section 2 I describe Millikan's account of content; in section 3 her account of function and conception of methodology; and in section 4 I argue that given the most sympathetic understanding of her project it still has grave difficulties. In section 5 I praise Millikan for various aspects of her account that do much better than those of other writers in this area; finally in section 6 I turn to the question of how the special character of psychological theory bears on the possibility of psychophysical laws.

5.2 Millikan's Theory

In this section I present Millikan's theory of content. First I'll sketch the "full-dress" version as it is presented in her book, then the pared down version from her "Biosemantics" paper. I will hold off presenting the account of function as it appears in her paper on function until the next Section.¹

Here and through the rest of this chapter I will concentrate on belief-like representational states, states whose "direction of fit" is mind-to-world, states that are supposed to match how the world is, rather than to make the world match them. I think most function-based accounts of content hold, implicitly or explicitly, that belief-like states and desire-like states have to come together or not at all (e.g., Millikan holds that the simplest kinds of representations are both belief-like and desire-like). Much the same kinds of problems come up for both kinds of states, and the problems do not come from failure to attend to the relations to the other kind of state, so concentrating on one will simplify discussion considerably.

Millikan has an idiosyncratic vocabulary for various aspects of meaning. The simplest content-bearing items are called intentional icons. An intentional icon is a representation only if one of its functions is to relate to other icons in such a way that the explanation for their joint success depends on the fact that they are about the same thing. (This is what she calls an "act of identifying" (LTOBC, p.242).) I doubt beliefs are merely representations, in this sense, since the criterion for being a representation is so weak; Millikan lists some other ways in which human thought differs from the rest of the simplest intentional icons at the end of "Biosemantics". Belief-like states are called indicative intentional icons, while desire-like states are called imperative intentional icons.

Millikan's theory of content is given in the form of a statement of the content of an intentional icon. Her exposition of the theory begins with some very low level definitions and only reaches the statement of the conditions on content after 100 pages. I will present the proposal in reverse order; since what I will say about it does not depend on the finer details I omit them. I will state her conditions using the term "provided" or "providing" as a place holder for the connective she intends. I will return to her conception of methodology and the force of her theory in the next section.

¹These works are: *Language, Thought and other Biological Categories* (Millikan, 1984), "Biosemantics" (Millikan, 1989a), and "In Defense of Proper Functions" (Millikan, 1989b). I will also discuss aspects of one other paper, which I will call "Laws": "Thoughts Without Laws; Cognitive Science with Content" (Millikan, 1986).

An indicative intentional icon is about or *of* the state of affairs that *p* provided the icon maps onto the state of affairs that *p* and “the *most proximate* Normal explanation for full proper performance of its interpreting devices as adapted to the icon” must mention this mapping (LTOBC, p.100). Notice that the devices have the functions that matter, not the icon (even though the icon does have functions, too), and that they are unspecified.

Something is an indicative intentional icon provided

1. it is a member of a reproductively established family having proper functions
2. it is an intermediary between two cooperative devices, where a Normal condition for proper performance of each is the presence and cooperation of the other
3. it functions to adapt the interpreter device to the condition mapped in order for the interpreter device to perform its proper function
4. the Normal explanation of how the icon adapts the interpreter device such that it can perform its proper functions refers to the fact that the icon maps conditions in the world in a specific way.

(This is a paraphrase of the four conditions she presents in LTOBC, pp.96-7.)

I’ll explain this set of conditions by commenting on four notions that seem central to understanding it.

(i) Something is a member of a reproductively established family having a proper function provided it is produced by a process of copying and the explanation for why it was produced refers to the fact that its ancestors actually performed a function. What is copied is some structural features of the entity; the explanation refers to the fact that having those structural features explains how functions were actually performed in the past. (Hence something has a proper function only if its existence has a complex explanation: what explains its presence is the fact that its structure, represented in its ancestors, explained performance of a function.)

Proper functions can be direct or derived. Something with a direct proper function is something that is supposed to perform that function itself. Something with a derived proper function inherits this function from something else that has a direct proper function. Organs or other aspects of biological systems typically perform their functions by producing items in response to particular conditions

in the world, possibly of a type never encountered before, and these items have a derived proper function in relation to these conditions. Consider a chameleon placed on a pattern of colors never before encountered by any chameleon; the particular pattern its skin produces has a derived proper function to match the pattern at hand, while the skin that produces the pattern has the direct proper function to produce skin patterns (LTOBC, Chapters 1 and 2). In Millikan's terminology, a particular skin pattern is "adapted to" a particular environmental condition; similarly an intentional icon causes an interpreter device to become "adapted to" the state of affairs the icon maps.

(In LTOBC the definition of an intentional icon reads "direct proper function," where I wrote "proper function." In LTOBC Millikan concentrates on language. Terms and syntactic patterns do have direct proper functions since we continue using them because *they* still do what they have done in the past. Thoughts do not obviously have direct proper functions in this way, since any particular thought may well not be a reproduction of any earlier thought. Millikan is quite clear, especially in "Laws", that she aims to account for their content in terms of derived proper function.)

(ii) Millikan never says much about what a "device" or a "cooperating device" is; she remarks casually that we are filled with "programs" that manipulate inner icons. Presumably a device is something with a direct proper function that works by way of exploiting structural (i.e., non-semantic, non-intentional) features of inner states. This reading is essential to make any sense of the notion of mapping.

(iii) An intentional icon maps conditions in the world in accordance with a specific mapping function provided there is a set of intentional icons (of which the current icon is a member) all of which are the same in some respect (the "invariant aspect") and a set of states of affairs such that there is a 1-1 mapping of intentional icon to the state of affairs that is the Normal condition for proper performance of the interaction between the cooperating devices; finally there is a transformation relation that takes one icon and its correlated success condition and yields another icon and its correlated success condition. (The sets are not partitions, since several of the icon transformations may be true at the same time.)

Millikan holds that the mapping relation is central to the notion of content, and that her conditions on it provide a substantive restriction on representation that should prove useful in thinking about ontology (LTOBC, pp.107-9). As stated the relation is almost vacuous. Take the sentence "Jack fell" and its success condition, that Jack fell; clearly there is a mapping transformation that takes this sentence into "For God's sake don't take the cat!" and its success condition, that the ad-

dressed person for God's sake does not take the cat, or into any other sentence and its condition of satisfaction whatever. She must have something else in mind. The example she relies on is bee dances. I speculate that the idea is that relative to bees and the way they are structured, there is a range of physical properties that bee dances have that correlate with responses bees make, and a determinate correlation between the responses and where flowers must be for the responses to yield success for bees. Applied to human thought the idea would be that there are devices in persons that are responsive to structural features of thoughts such that the thought/device interaction is Normally successful only when the normal relation between structural feature and state of affairs is realized ("Biosemantics", p.286).

Millikan's descriptions of the mapping relation suggest that she thinks of mapping as a non-normative relation, or, at minimum, a relation independent of the other things she says about intentional icons, so that perhaps something can map even if it does not satisfy other conditions on being an intentional icon. I am dubious this is so; virtually every concept she defines is normative in some respect and none can be understood independent of the others. But I also doubt this shows her account is viciously circular; all her terms must be interpreted together relative to some conception of function.

(iv) Normal explanations are explanations that show how functions were actually performed, and moreover, in "non-accidental" circumstances. A Normal explanation for how a heart pumps blood starts with structural features of hearts and blood, looks to cases where hearts pump blood in the ways typical of the cases that lead to the proliferation of hearts (call these the "central cases"), adds laws that cover conditions in those cases, and deduces the pumping. The most proximate Normal explanation is one that goes from the function-performer to the function performed and stops there; less proximate Normal explanations add explanations of the presence of various factors in the central cases, like where the heart gets its energy from, where the blood comes from, what controls the heart, etc. The conditions in the central cases that are needed for the performance of the function are Normal conditions (LTOBC, p.33-4). Normal explanations gain a measure of objectivity from the requirement that they are the explanations that govern a very large number of cases, those that explain the ways a species has proliferated.

That completes my exposition of the theory as it shows up in LTOBC. The statement given in "Biosemantics" is supposed to be a more digestible version of the LTOBC theory, with somewhat more emphasis on the idea that the content

of an indicative intentional icon is set by normal conditions for the proper performance of consumer device functions (whatever they are), not the functions of the intentional icon itself. In “Biosemantics” she uses the term “representation” to include all intentional icons, and leaves off capitalizing “normal”, while noting that “normal” still has the heavily normative sense described above.

She gives two conditions for something to be a representation.

First, unless the representation accords, *so* (by a certain rule), with a represented, the consumer’s normal use of, or response to, the representation will not be able to fulfill all of the consumer’s proper functions in so responding—not, at least, in accordance with a normal explanation. . . . Second, represented conditions are conditions that vary, depending on the *form* of the representation, in accordance with specifiable correspondence rules that give the semantics for the relevant *system* of representation. (“Biosemantics”, pp.286-7)

We might read these two conditions as a pair of necessary conditions on being a representation: If something is a representation then it is supposed to correlate with the world in a certain way, and a normal condition for the successful performance of the things that respond to it, whatever their functions, is that the world actually is as the correlation requires; furthermore, the character of the correlation is such that there must be a range of possible variations on the representation such that each variation type correlates with a different type of state of affairs, and an instance of that state of affairs is a normal condition for the performance of the functions of these “consumer devices” when presented with an instance of the representation type. The second condition corresponds to the mapping requirement in the “full-dress” version of the theory, and the first condition compresses and abbreviates all the rest. The conditions are clearly not independent: the second just explicates the gnomic “*so*” in the first.

5.3 Function and Philosophical Methodology

Millikan defends her notion of function in “Defense” in part with some remarks on her conception of what kind of project she is engaged in and how it differs from more traditionally philosophical projects. In this section I will criticize her account of function and her basic methodological stance. The central difficulty

with the account of function is that it does not address the question whether function is a naturalistic notion. The trouble with the methodological stance is that she describes no plausible project: she claims not to be doing conceptual inquiry, but she clearly isn't doing empirical inquiry either and her account seems responsive to conceptual issues even by her own lights.

I begin with two very broad remarks about the kind of project Millikan and others are engaged in. Stampe suggested that function was a good notion with which to supplement a purely causal account, but neither he nor Dretske has anything to say about why this should be so. One thing that distinguishes content from other kinds of properties is that content is normative in some way. For instance, if a content-bearing item is complex in the way that human thoughts are complex, then it involves concepts, and concepts are such that they are correctly applied to some things and incorrectly applied to other things. The precise nature and details of the connection between content and norm are obscure, however. Fodor can be read as attempting to describe a causal relation, without using any value terms, that is satisfied by things to which norms apply. The apparent complete lack of success of this attempt (see Chapter 6) is ground for thinking that the connection between content and norm is at least strong enough to require that any account of content will use some normative notion. The explanation for the use of the notion of function then is that function is a normative notion. Millikan is clearer on this issue than many others, explicitly recognizing that that is the point of using a notion of function.² She does not contribute to a clearer understanding of how exactly content is normative and why function should be the right notion to use.

Any function based account of content also presupposes that an account of content can be given, that some relation between one thing and another is constitutive of the one being about the other. To my knowledge, no-one in the literature on function based accounts of function has defended this assumption in any convincing way. Millikan is clearly aware that there is an issue here; her response is that to fail to make the assumption leads to apparently insoluble philosophical difficulties, and that making it leads to interesting and fruitful theories (LTOBC, p.87). I noted in Chapter 1 that this is the only response currently available; but that it is relatively easily countered by someone who does not presuppose an account of content is forthcoming. The counter is to gamble that our best overall account of everything whatsoever will take content properties as primitive.

If we think that an account of content can be given and that it requires a notion

²In "Speaking Up For Darwin," in Georges Rey and Barry Loewer, eds., *Meaning and Mind; Fodor and his Critics* (forthcoming).

of function to supply the required normativity, the central difficulty is to provide a naturalistic notion of function and its normativity. There are two extremes to avoid; part of the difficulty is to show how there is anything left between the extremes.

Much philosophical work has been done on the notion of function, in an effort to show how explanation in biology appeals to no special properties or forces in the world. The trouble with these proposals is that while they give a thoroughly mechanistic criterion for the application of the relevant concept, it's not a normative concept. A system may show, for instance, any degree of plasticity you wish in getting to some designated end state ("pursuing a goal") but never do anything correctly or incorrectly, well or badly.

Any account of function must bring in a value notion at some point: e.g., the goal state is good and the explanation for why the system tends toward that state is the fact that it is good.³ The difficulty on this side is getting a suitably naturalistic account of value properties. Value properties may be understood as some sort of projective property: good things are ones we are disposed to call good, or feel a certain way about; but this notion of value would be useless in a naturalistic account of content. The only possible alternative seems to be to hold that there are natural, mind-independent facts about value. The corresponding difficulty for the notion of function then is that if we cannot articulate a naturalistic conception of value properties then function attributions are interest- relative, projective attributions that do not get at how the world is independent of the minds that regard it. (I shall call these two conceptions of function "extrinsic" and "intrinsic".)

How does Millikan handle these difficulties? The account of function she defends in "Defense" is this:

for an item *A* to have function *F* as a "proper function", it is necessary (and close to sufficient) that one of these two conditions should hold. (1) *A* originated as a "reproduction" ... of some prior item or items that, *due* in part to possession of the properties reproduced, have actually performed *F* in the past, and *A* exists because (causally historically because) of this or these performances. (2) *A* originated as the product of some prior device that, given its circumstances, had performance of *F* as a proper function and that, under those circumstances, normally causes *F* to be performed by *means* of producing an item like *A*. ("Defense", p.288)

³For a careful and exhaustive defence of this claim see the work of Mark Bedau.

The two clauses correspond to the difference between direct proper function and derived proper function. There are several things wrong here.

First, the definition places a condition on a broad class that yields a narrower class: of the things that have a function, the ones that have a proper function are a certain way. Hence it leaves the question in which we are most interested, the question about the nature of function, completely untouched. (Perhaps the crucial point lies in the parenthetical “and close to sufficient” but Millikan never elaborates this hint. See LTOBC, pp.25-28 for the same lacuna in the “full-dress” version of the theory.)

Second, the difference between a proper function and some other kind might be read as the difference between an intrinsic and an extrinsic function; this would clearly show that this definition is designed to show how things have functions independent of any people who attribute function. But they are not the same distinction. All it takes for something to have a proper function is that it has a function that accounts for its continued reproduction or survival; where the function comes from is left open. This is clear from her discussion of the introduction of new terms, nicknames and the like: they get a proper function only when they are reproduced in order to get the same good things to happen (LTOBC, pp.81-82). But clearly they have a function in virtue of the uses to which people put them. Her testimony on this point is inconsistent between “Biosemantics” and “Defense”; in “Biosemantics” she writes, “Natural selection is not the only source of proper functions,” (note 7, p.284) while in “Defense” she proclaims that “I do need to assume the truth of evolutionary theory in order to show that quite mundane functional items such as screwdrivers . . . are indeed items with proper functions” (p.290).

Third, something about evolutionary theory can seem to ground a claim about intrinsic function: the account of function reaches into the deep past of the species in order to determine function. But it’s not clear why the stretch of time should show that things have these functions intrinsically; if you hold that all function attributions are interest relative, then you will hold that evolutionary theory is confirmed by seeing whether the functions you are interested in really do have the role you think they do in promoting survival and reproduction. (Evolutionary biologists who have no qualms about the notion of function will hold that whether something has a particular function is an objective fact. My point is that more is needed to base this objectivity than history: a long history of interest-relative facts is not thereby a history of objective facts.)

Millikan’s defense of proper function begins with some remarks on kinds of

“definitions”. Her “definitions” must be understood in a special way, not the way we understand traditional conceptual analyses. I turn now to these remarks; they do not settle the question about function I have raised, but they do suggest a coherent stance on it, albeit one that holds little promise for a satisfactory account of content.

Millikan’s definition of function is supposed to be understood as something like a “theoretical definition”, as “water = H₂O” is a theoretical definition of water. A theoretical definition, as I understand her use of this notion, is one that gives the natural essence of something. Chemical investigation showed that water has this structure, lapidary investigation shows that there are two distinct kinds of things called “jade,” and medical investigation shows that several different things were called “consumption.” A definition like this has nomic force: as a matter of law all water and only water is (mostly) H₂O. Hence this definition cannot be criticized by describing possibilities in which water is not H₂O.⁴ It shows rather that some phenomenon in our actual world has an underlying explanation in terms of some powerful comprehensive theory.

I don’t know if any philosophers hold that conceptual analysis is the description of “marks that people actually attend to when applying terms” (“Defense”, p.291). Conceptual analysis is standardly understood to be a specification of how terms ought to be applied, both to actual and to hypothetical cases. Millikan’s treatment of the accidental double case, in LTOBC, p.93 and “Defense”, p. 292-3, leads me to think that she is doing what people traditionally call conceptual analysis. The case is this: take any object that has a proper function, as sophisticated as you like, then consider something that is a molecule-for- molecule duplicate of it, something that has just the same internal physical properties, but does not have the history required for proper function. Her verdict is that this object does not have proper functions. She seems tempted to say it has no function, either; she does say it does not have a purpose. Finally she says that if it were a duplicate of you it would not have any thoughts. (She does appear to think that it would seem to it that it did; I will return to this below.)

The trouble with this verdict is that the case is just the sort of hypothetical she has just asserted has no bearing on her account of proper function. She has several other courses open to her: agree that the accidental double has a function but no proper function; remain agnostic, since the case is not an actual one; deny the claim that such cases have anything to show us, since our intuitions are strained.

⁴Millikan does not seem to be aware of Kripke’s suggestion that if H₂O=water then necessarily H₂O=water, and so there are no such possibilities.

She does none of these; she applies her definition as though it determined the correct application of concepts even in extremely hypothetical cases.

Millikan's interesting criticisms of Wright's account of function in "Defense" also point in this direction: she claims that we *cannot* explain how things can have functions they cannot perform if we do not use a historically based account of function.⁵

We might overlook this point and attempt to take seriously her idea that proper function is the underlying phenomenon in many instances of function, in the way that H₂O underlies the phenomenon of water. But the idea itself has difficulties.

First, the definition is still circular, in that it uses the notion of function in defining proper function. Circularity in definition is not always a vice; the definition may show something important and interesting about the way one notion interacts with others. The trouble is that the other notions here are just the ones we need to understand.

Second, the idea of a theoretical definition in her sense requires that the term being defined is a natural kind term. Pretty clearly "water" and "jade" and "consumption" are natural kind terms. But to assert that "function" is a natural kind term is simply to beg the question at issue here.

Is there any other way to understand what Millikan has in mind? I shall consider and reject three possibilities, then formulate what I think she means.

(i) In Chapter 4, section 5, I argued that the distinction between living and non-living things might be a difficulty for someone who aims to base an account of function in biological phenomena. Efforts have been made to show how all and only (actual) living things share some mechanical property; if these efforts succeed we could use this property in our definition of function. We could do something similar with molecular genetics: since (as far as we know) all and only actual genes are nucleic acids of certain kinds, we could dispense with the notion of genes and use this well-understood notion instead.

This sort of definition would be of little philosophical interest, since it says nothing about how the concept ought to be used in actual cases unlike ones we have encountered until now or in hypothetical cases.

(ii) Millikan's writing on identity (LTOBC, Chapters 16 and 17) strongly suggest that she intends what she says to have nomic force. It would be interesting

⁵I think she is right that this is a difficult problem, but I am not persuaded that a careful dispositional account (involving claims about what something would do under certain specified conditions, along with facts about history, where appropriate) could not do as well.

and important if it turned out as a matter of nomic necessity that all and only items with a function have a proper function in the way her theory suggests. But that discovery would be essentially irrelevant to the project of understanding the nature of function. Whether there can be such a law depends on just the question at issue: which things have functions and why. Such a law would not in any event show that function facts just are proper function facts; for a conclusion like that we need more than nomic necessity. It would be like answering the question, what are screwdrivers, by noting some physical characteristic that as a matter of nomic necessity all and only screwdrivers share.

(iii) I suspect Millikan is attracted by a very dramatic possibility for the ways empirical and philosophical inquiry interact. Philosophers have been forced to accept various revisions in our concepts by results in empirical inquiry; at the moment, for instance, physics appears to be forcing a revision in how we all understand the nature and scope of causal relations. A biosemantic approach might claim a similar revisionist power: empirical studies in the foundations of cognition and evolution will force us to revise cherished ideas about thought and function. There are two difficulties for such a claim. First, no amount of empirical inquiry will remove the lacunae we've seen in the notion of proper function. Second, this form of claim is at best a gamble on very long odds at this point; given the state of the proposal it seems we'd be more inclined to stick with our ordinary notions. I'll return to this point in section 4 below.

If we set the problematic details of Millikan's notion of function to one side, I think we can see what sort of stance she means to take. Suppose it is uncontroversial that biology involves a notion of function that is genuinely normative. Then one kind of project of naturalizing function and content would be to show first how as a matter of fact all or virtually all intrinsic functions are biological functions. We would then show how as a matter of fact all items with content have some important connection with biological function. This would be at root an empirical hypothesis. If the project succeeds it would not show how to reduce the biological notion of function to something better understood, or show how to understand content in terms better understood than those of biological theory. It would at best show how the theory of content is as respectable as biological theory. (This might be an important reassurance for someone who worries that if we cannot show how psychology is a certain kind of science we must conclude that psychological states and processes aren't real and that we should drop psychology, both lay and scientific.)

This is a coherent description of a project, but I don't think it is a plausible

project, since I think the biological account of function is the wrong account to use in a theory of content; that will be the topic of the next section. I do, however, think that if we are at all inclined to think that very simple creatures have representational states (if not bacteria then perhaps fish or gerbils) then the biological account of function is the right one.

I'll close this section with some remarks about Millikan's views about consciousness. She is emphatic about certain severe revisionist consequences of her biosemantic approach. I agree that revision is to be expected, but I think in this case Millikan has simply gotten trapped in just the picture of the mind she is so eager to overcome.

Millikan thinks that one of the reasons people are wary of an account like hers is that it makes facts about content depend on very complex relational facts, including facts about the deep past of a species, that are not available to consciousness. The wariness, Millikan thinks, stems from our deep conviction that consciousness is epistemically transparent, that our knowledge of our own mental states, in particular our knowledge of their content, is incorrigible: facts about our deep past are not present to consciousness, so they must be irrelevant to content. This conviction she labels "meaning rationalism"; one of her programmatic goals in LTOBC is to show the pernicious influence of this conviction and to show why it's wrong. Her solution is therefore to deny that consciousness is epistemically transparent. I take her to hold that we may have a thought, and know that we have it, yet fail to know what its content is; worse, we may think we have a thought but be wrong. "It turns out that we cannot know a priori either *that* we think or what we think *about*" (LTOBC, p.6).

Her response is hasty. It's true that incorrigibility is not the mark of the mental and that occasionally it's not true of our mental states, but this is a very rare and special sort of event. A somewhat easier route to pursue is to deny that there is anything epistemic about consciousness that isn't given the same sort of functional account, and then to go on to argue for the claim that when we have thoughts further thoughts about their content are available to consciousness. (Davidson makes this kind of claim about our authoritative beliefs about our own beliefs.⁶) Brian Smith makes the same sort of claim about consciousness.⁷ We certainly do not want to make consciousness epistemic in some impoverished way, as Millikan seems to want, holding that consciousness generally delivers something about

⁶"First Person Authority" (Davidson, 1984a).

⁷In unpublished work presented at the Cognitive Science Seminar at UC Berkeley, November 17, 1989.

thoughts but not facts about their content.

5.4 Problems for Millikan's Account of Content

In "Biosemantics" Millikan proposes her theory of content as a pair of conditions; in LTOBC the theory comes out as a more loosely numerable set of conditions on content. When I presented the conditions in Section 2 above I deferred to Millikan on the strength of the connective she intended. Presumably her account of content has the same logical character as her account of function. It is to be understood as a very high-level empirical hypothesis about content, to be treated as a guide for empirical work. It can also be approached as a piece of conceptual analysis, since it is meant to be true to what we currently believe about biology, function and content. I propose to treat her account as a piece of conceptual analysis, reading "provided" as "necessarily if and only if," and indicate in passing where I think Millikan can with some right ignore my complaints, and where I think my objections are telling even if the account is only an empirical hypothesis.

How does Millikan's account work, then, as a set of necessary and sufficient conditions? I will try to answer this question for the case of human belief, since that is by far the most interesting case, and the case Millikan pushes hardest in "Biosemantics"; I think the account works better, like most functional accounts of content, for very simple things.

I'm going to argue that the conditions don't work in either direction. They aren't sufficient because they let in some cases which by Millikan's own lights are not representations. The argument against their necessity is a bit more complex. It's very hard to hold that it is just false that a normal condition for the successful performance of people and their cognitive apparatus is the truth of their beliefs. The argument will be instead that the particular way Millikan needs to understand this claim (in terms of the proper functions of representation consumers) is unlikely to be true; even if it were true, it's not very clear what bearing it has on our confident ordinary content ascriptions.

Sufficiency first. Millikan is confident ("Biosemantics", p.282) that a red face is not a representation of what it was designed to respond to (perhaps being overheated). Yet it doesn't seem difficult to describe various aspects of the red face in such a way that it turns out to be a representation after all. The pieces we need are these: a range of particular kinds of red face, all united by the invariant aspect "red face"; a range of states of affairs in the world with which to correlate these

kinds; a consumer of these various kinds of red face; and some function that the consumer normally performs, given one of the range, only when the correlated state of affairs obtains.

Well, suppose the consumer is the skin. One of its functions is to transfer heat when internal temperature is incorrect and external conditions are suitable for correcting the imbalance. The degree of heat transfer is controlled by the degree of capillary dilation (the degree of redness): lots of dilation, lots of heat transfer. The function of the consumer is performed only when the degree of dilation corresponds well to the combination of the internal heat imbalance and the external temperature. It's dangerous to drink in the extreme cold since your capillaries will dilate and you will lose more heat than you should. (It should be noted that the red face still isn't really the representation, but rather the degree of capillary dilation; this is a rhetorical slide that Millikan makes, so I think it is fair to find something close enough that does count as a representation by her theory. It's important to respond to a case on which she gives a verdict, since she may well simply agree that many other internal regulative systems are representational systems, like the immune system or the digestive tract.)

Her definitions fail of sufficiency in a couple of other ways as well. Consider visual representations. Lacking a definition of what counts as their consumers we might just say, the agent. But Millikan's talk of devices and programs suggests that for complex creatures there are internal systems of representation transference and manipulation. Suppose it turns out that there is a little feedback loop in our optic system. Its job is to respond to certain features of the visual representations as they occur on the optic nerve. These features indicate various states of the retina. The feedback controls the state of the retina. There's no problem with holding that the loop has a proper function that it gets from a selectional account of function. Clearly it performs this function only when properties of the incoming representation do "get things right" about the retina, i.e., when the actual correlation between the representation and the retina is as it was designed to be. The trouble then is that the representation represents more than one state of affairs: whatever in the external world it is a perception of, and the current state of the retina.

This case illustrates another difficulty for Millikan's account, a version of the "proximity" problem, to show why a representation is about something in the world rather than about the surface of the person or some state of the brain. Using the notion of function is a very useful first step in solving this problem, since typically representational states have functions that depend on external conditions.

This case shows that even if the proximity problem gets a partial solution, Millikan's account is still committed to a lot of representations of internal states.

The proximity problem can be urged in a more general way. Consider visual representations again. One normal condition for the proper function of devices that consume visual representations is surely that the eyes are in good condition, the retina healthy, and so forth. The compositionality requirement on content rules this condition out as a candidate for the content of the representation, since it is constant across all visual representations. The trouble is that there must be some condition of, say, the retina, such that the performance of functions of visual representation consumers depends on this condition, and this condition varies with the particular structure of the representation. I see my cat in the window; my retina must be in the state, "satisfactory condition in producing a pattern of inputs to the optic nerve of such and such structural character", for my representation consumers to function correctly. Hence it appears my representation is also about that state of the retina.

I conclude that Millikan's conditions are not sufficient. The difficulties about proximity are not likely seriously to disturb Millikan, despite her inclination to raise them against other theorists ("Biosemantics", p.282-3). All she needs is an account that gets some central cases right, in particular, one that solves the proximity problem. After that she will simply accept the extra content attributions as a sort of theoretical discovery. Whether she can dismiss the objection about the red face is trickier. Additional refinements may rule out the worst cases like this; as the theory grows more sophisticated Millikan's inclination simply to accept a few odd consequences will also increase.

There is one other important question about sufficiency: does a theory like this have the resources to explain the intensionality of attributions of content, particularly the degree of intensionality of attributions of propositional attitudes? My own view is that it does well enough except for some problematic cases in human cognition involving necessarily coextensive predicates. I will discuss this kind of objection in more detail in Chapter 6 on Fodor.

I turn now to necessity. One clear way to show a lack of necessity is to hold that normal conditions for the performance of the functions of things that consume beliefs don't always include the truth of the beliefs. This might be true for local belief manipulators, if there are any, but it's not likely to be true for persons: in the long run, at any rate, you do well only when your beliefs reflect the truth. What makes this straightforward objection so difficult is that we just do not know what these devices and function are. One's suspicion is that whatever they are,

truth isn't *guaranteed* to be a normal condition for performance in *every* case. But it's hard to move beyond mere suspicion without more detail in the theory. I propose therefore to look elsewhere; I will return to this style of objection in the next section.

Jerry Fodor and Graeme Forbes urge related complaints:⁸ how can a normal condition for the performance of consumer functions be the truth of a belief, if the belief could not possibly be true? There just are no historical conditions in which anything interacting with such a belief has performed a function because the belief was true. Millikan has a ready answer: beliefs and desires have derived proper functions, ones that they inherit from the devices that produce them and consume them (see "Laws", especially p.55 and p.63). Granted this much the rest is easy: since the whole panoply of propositional attitudes that these devices generate and consume have a "compositional semantics", we can see how a belief that the circle can be squared has a content that depends on the way its components serve functions in other beliefs.

The real trouble for the necessity of her proposal comes with the need to make the functions of all these things proper functions, in some suitably biological way, and for there to be such a collection of mechanical items in the head to subserve belief and desire. There are three reasons I think all this is not necessary.

First, I do not think there must be internal representations to subserve the propositional attitudes. No convincing reasons have been given by Millikan or by anyone else to the contrary. My reason for thinking this comes originally from Churchland.⁹ Suppose we think of propositional attitude ascriptions as a kind of measurement. We look at another agent, see how she behaves, linguistically and otherwise, and we write down a theory that reports what she believes and what she wants and what her words mean. The theory captures aspects of her doings that are mirrored by the way aspects of my speech and behavior are related, just as a measurement of temperature captures aspects of a body that are mirrored by certain relations in the rational numbers. Yet certainly we wouldn't want to say that a body that has a temperature of 98 degrees contains somewhere within it the numerals "98". If the parallel is a good one, we should be equally cautious in

⁸Jerry Fodor, "Information and Representation" (Fodor, 1989); Graeme Forbes, "Biosemantics" (Forbes, 1989).

⁹Paul Churchland, *Scientific Realism and the Plasticity of Mind* (Churchland, 1979, 105). Donald Davidson urges this point with the additional suggestion that the relevant scheme of measurement is a unified theory of decision and interpretation; see, for instance, "A New Basis for Decision Theory" (Davidson, 1985a).

claiming, because we mirror aspects of someone's behavior with relations among our sentences, that the person must have somewhere inside little sentences that back up the ascriptions.

If there do not have to be mental representations in this sense, then for instance mapping relations in the strong sense I sketched in Section 2 above are not needed for there to be beliefs.

It should be noted that for someone's behavior to be measurable in this way there must be some aspects of them that can be stably correlated with recurring aspects of the representing system. These aspects of the system being measured may be internal representations, but equally they may be aspects of behavior, as for instance linguistic behavior. The reason Millikan is open to this kind of objection is that she aims to treat the content of thought and the content of language independently (LTOBC, pp. 89- 90).

Second, even if there are such mental representations and cooperative interpreter devices, it's unclear what relevance this has to our propositional attitudes. Suppose we undertake to find these devices and their proper functions, and we discover that in fact their proper functions determine rather different contents for our thoughts than we think they have, or none at all. (Imagine that we turned out to have been built by Martians to perform these elegant cognitive dances for some purpose, but the purpose can be fulfilled regardless of any mapping relation between beliefs and the world.) I do not think that would incline us to think that what we thought we thought was wrong. Or suppose, less fancifully, that we do the required empirical investigation and it turns out that our heads are filled with cooperating devices, but they don't have the right proper functions to determine the contents we think they should have. Again, I don't think that would incline us to change our minds about our thoughts.

Third, Millikan seems to think that there is a clear argument from the fact that our cognitive life does have a function to the conclusion that there must be devices in us that have proper functions; then the necessity of her conditions follows simply from the fact that we can come up with functions for our cognitive life. The arguments ("Biosemantics", p.293, "Laws", p.55) have the form, to think otherwise would be irresponsible, or, terrible consequences (epiphenomenalism) would follow. But there simply is no argument here. It could well be that all these inner devices are "mechanisms that evolution devised with other things in mind" ("Laws", p.55).

I conclude that Millikan's proper function conditions on the content of human

propositional attitudes are not necessary.

How important is this conclusion to Millikan's project? Accidental double cases could show that proper function isn't necessary to the content of very simple creatures. Millikan is on reasonably solid ground either if she holds that such cases do not yield stable, grounded conclusions about function, or if she simply stipulates that such cases are not relevant to her claims. Then proper function might be necessary for very simple creatures.

The problems I raise for the necessity of proper function for an account of human thought are more serious, no matter how we understand the theory. If I'm right that even if there are mental representations in some appropriately strong sense, and they have proper functions, this doesn't matter to content attribution, then it seems very unlikely that the theory is of any interest in understanding content.

Here's a variant on this objection: what reason would we have to have such a notion of content? Content attributions are made in the service of explaining and understanding particular human beings. Certainly their selectional history is relevant to understanding aspects of what they do and how they do it; but why think that reason explanations of particular actions somehow reach into the history of the species? Why should we care if content attributions based on ordinary interpretation of individuals diverge from attributions based on the species history?¹⁰

5.5 Millikan on Other Functional Accounts

In this section I assess Millikan's responses in "Biosemantics" to other writers who offer naturalistic accounts of content, and comment on one way her ideas might be applied outside the project of naturalizing content. I also describe what may turn out to be a fatal dilemma for all projects of naturalizing content.

Millikan points out what might be called the scope of function attributions. "Functions to ..." creates intensional contexts. As with "says that ..." there is no syntactic guarantee that every part of the sentence succeeding the phrase is governed by the phrase; various devices, both syntactic and pragmatic, are available to indicate which terms are in transparent position. Lack of caution in stating a functional theory of content then leads to the sort of problems Millikan points out: something might have the job of producing items that indicate, but the job

¹⁰I owe this variant to Alison Gopnik.

might not be to produce items to indicate.

Millikan's suggestion that content should be fixed relative to the normal conditions for the performance of the functions of representation consumers is, I think, a major contribution to the whole discussion. She claims several specific virtues it has over Dretske and Matthen's account, and I'll discuss these in a moment. It is also clearly superior to the sort of account that Fodor criticizes in his attacks on teleological accounts of content; I will return to this point in Chapter 6 below. Even more remarkable, and virtually unnoticed, the account is not a *causal* account of content at all: it is possible for a representational state type to be of a type of state of affairs although no instance of the representation type is ever caused by the state of affairs. This is true, on her theory, both in the case of complex theoretical beliefs, where we would expect such a result, and in the case of the simplest beliefs of experience: we can have true sensory beliefs that are never caused by what they are about.

This shift in focus gives a clean (and, I think, correct) answer to what the magnetosome in certain marine bacteria represents. One very strong intuition is that the magnetosome doesn't represent anything at all, bacteria just don't have enough of whatever it takes to represent. Dretske finally decides in "Misrepresentation" that it doesn't represent at all. His reasons for this decision are poor, however, as we saw in Chapter 4 above. Clearly a learning based account of function has more promise as a base for an account of human intentionality, but it would be arbitrary to deny that very simple things have contentful states even though they satisfy the general account of content. Millikan is happy to attribute content to them and her account gives what seems to be the right answer: the representation is about those conditions that obtain when the representation consumers function well given the representation. In the bacterium the representation is never caused by those conditions; it is only caused by the position of magnetic north.

Millikan accuses Matthen of giving up too easily on the function of beliefs and desires. The difficulty he sees is that the functions of particular beliefs and desires are far too idiosyncratic to support a function based account of content.¹¹ This is a critical point, since each function based account of function we have examined has this problem. Millikan thinks she can "cleanly bypass" this whole problem, by looking at normal conditions for performance of unspecified functions rather than at functions to indicate.

The particular form of Millikan's proposal—that the functions must be biolog-

¹¹Mohan Matthen, "Biological Functions and Perceptual Content" (Matthen, 1988).

ical proper functions—renders it unacceptable. Could Millikan’s idea be deployed in a more acceptable framework? There are two questions to answer here. First, if our project is naturalizing content, can we use some other naturalistic notion of function, like Dretske’s, to set content in Millikan’s way? Second, if we are interested in the structure of content and do not care about naturalizing it, can her suggestion be used in other accounts of content? The answer to the first question is, I think, no; this makes for a serious dilemma for any function based account of content. After describing the dilemma I will point to one way to use Millikan’s suggestion in a rather different theory of meaning, and leave the project of evaluating it for another occasion.

If we aim to put Dretske’s account together with Millikan’s, the suggestion would go something like this. We posit various representation producers and consumers. We attribute functions to them on the basis of a learning history: if a system can and does alter the characteristics of this collection of devices in order to serve needs better, then they have functions. The representations they exchange must, as in Millikan’s account, have a compositional structure, and the functions of the devices are normally performed only when the representation occurs along with its mapped state of affairs.

The suggestion may solve the problem about specifying the function of particular beliefs. It doesn’t settle the question whether needs are naturalistic. This is important since if biology is not the standard by which to judge whether the theory is naturalistic then we are faced again with the urgent problem whether value is a natural aspect of things. It makes the same unacceptable demand that there must be inner representations. Finally, it certainly would not explain why content is rational, since part of the account involves an explanation for certain structures and events that cannot be stated without the idea of an optimal balance of cost and benefit.

Millikan’s attempt to bypass the problem of the function of particular beliefs suggests a dilemma for *all* function based accounts of content. Dretske and Matthen attach function directly to particular types of belief in order to set their contents. The trouble then is that the notion of function is certain to be question-begging. Millikan uses a simpler notion of function to set content, the function of general purpose devices that exchange representations. But now since the functions involved are distinct from the functions of particular beliefs we have no guarantee that these functions will always line up with content, even though apparently they must in a very large range of cases. It’s hard to see how to rule out the possibility that for some restricted range of beliefs the consumer devices

function just as well regardless of the truth of the beliefs. Since these alternatives seem to exhaust the possibilities, it seems any careful deployment of function will have the same problem: either the notion of function is clearly question-begging, or it is not strong enough to guarantee that the actual content may diverge from the content set by the functional account, for some narrow range of cases.

How would Millikan's suggestion look in the general project of understanding content? It could provide an interesting corrective to a feature of Davidson's theory of meaning. Davidson interprets a speaker's sentences in two stages. (See "A New Basis for Decision Theory.") We collect a base of evidence about weak preference that one sentence rather than another be true. Using this evidence we can determine which parts of the sentences are truth-functional sentential connectives, and which are the apparatus of quantification, and we can determine the subjective probabilities and desirabilities of the sentences. The second stage is to get interpretations for predicates and singular terms; we do this by attending to what causes weak preference and to what objects and events are connected to indexical elements in speech. The second stage strongly suggests a directly causal account of meaning: in some basic range of cases sentences are about what causes them. But if Millikan is right we may need to supplement this. We consider when an agent holds a sentence true, and look for what must be true for desired outcomes to be realized through the actions of the agent when the reason that causes the action includes the belief that corresponds to holding the sentence true; the content set in this way may differ from the content set by considering what causes the sentence to be held true. (This alteration would have no consequence for the epistemological virtues of a causal externalist theory of meaning; it simply states more precisely how causal relations bear on meaning, and entails that meaning would still vary with a different distribution of external causes and valuable objects and properties.)

5.6 Millikan and Psychophysical Laws

In "Laws" Millikan presents a defense of cognitive science based on her theory of content. In this section I briefly discuss this defense, and the related question whether there can be psychophysical laws.

Millikan sees the issue between the Churchlands and Fodor over the question whether psychology is a science as an argument over what kinds of inquiry deserve the honorific "science". The parties to that debate (but not Millikan) as-

sume that for a proper science there must be laws of a certain kind, and for a certain class of items to be real items there must be a proper science of them. The Churchlands think that there is no predictive science of psychology, so there is no proper science of psychological states, so they think there aren't any; there is a proper science of neurophysiological states, so that's what people who were interested in cognition ought to do. Fodor takes pains to show that there is something psychology does that deserves the honorific "proper science," hence there are laws of a certain kind and there really are psychological states.

Millikan's response to all this is, first, to agree that there are no laws, but second, to offer a rather different diagnosis of why there are no laws. The Churchlands hold that psychology is an empirical theory that captures certain dispositions, and that it is a completely moribund theory. Millikan holds that psychology is an empirical theory that captures items in functional categories. There are no laws governing psychological states because things in functional categories may perform their functions only rather rarely, especially if performance depends on cooperation by the world. Then psychology and cognitive science have a clear project: to discover these items and their functions and to describe how they perform the functions when they do.

I have two remarks on these ideas.

First, I doubt whether Millikan provides a good argument that shows there cannot be psychological or psychophysical laws. Her argument is this. Some people hold that if you include normal conditions for the proper performance of some function in a *ceteris paribus* clause, you can get a strict law: given appropriate antecedent conditions, an item with a function will perform that function, provided conditions are normal. She points out that even under these conditions items with functions often fail to perform their functions. There are two things wrong here. First, there may be rough laws here anyway. Second, perhaps there aren't any strict laws linking antecedent conditions with *performance* of function, but there could be, for all she says, a strict law linking, for instance, some function predicate (having the function, not performing it) with some relatively tidy physical predicate. (See LTOBC, p.139, for the same mistake.) She is, I think, right in a slightly different point: the likelihood that there is such a law is very remote, and the consideration that perhaps psychological categories really are function categories sharpens our appreciation of how remote the likelihood is.

Millikan's argument about laws shows that in broadest outline she agrees with the overall argument I am pursuing. First, if we take a suitably general notion of rationality, a notion such that a system displays this kind of rationality pro-

vided explanation of its aspects and movements relies in some way on benefits these can yield, then Millikan clearly holds that if a system has content-bearing states then it is rational in this sense. Presumably then she would be happy to agree with the thesis of content rationalism, as it was described in Chapter 1. Second, she traces the difficulty for psychophysical laws to the functional nature of psychological states. Hence she also agrees that the special character of psychological states makes a difficulty for psychophysical laws. So despite our virtually complete disagreement on the details, Millikan agrees with my claim that content rationalism is true and that that shows why psychological explanation is different from explanation in the natural sciences.

Second remark: certainly some beautiful results in cognitive science have been precisely discoveries about how certain functions are performed. Saul Sternberg showed that the way we answer questions about whether a currently presented word occurs in a list that currently resides in short term memory is by doing a serial exhaustive search of the list.¹² Whether that is a biological function or a biological solution to a biological problem is altogether another question. In any event, I doubt all cognitive science is or should be inquiry into what biological functions psychological states have and how they perform them. Testing decision theory, on the face of it, is not inquiry into biological function. If anything it is work on whether Millikan's theory is correct. Suppose we have persuasive evidence that our decision making is systematically flawed. We might then seek a biological explanation for why this is so, something to do with strategies successful in local conditions. We might discover that there is none: we simply are wired in such a way that we don't live up to the norm of rationality provided by decision theory. It might turn out that there isn't even an explanation available of the form, but for other selectional pressures we would be more rational; it might turn out, in other words, that these failings are selectionally irrelevant.

5.7 Conclusions

I have argued two claims against Millikan's biosemantic approach. First, she has not shown that the notion of function she defends is naturalistic in the ways such a notion ought to be naturalistic. Second, supposing that her project has been only to put psychology on as naturalistic a ground as biology, she has not shown

¹²Saul Sternberg, "Memory-scanning: Mental processes Revealed by Reaction-time Experiments" (Sternberg, 1969).

that (human) psychological categories are biological proper function categories. I strongly doubt that they are; they might be proper function categories in relation to the functions we clearly understand pieces of our mental life to have, but that would not be a naturalistic theory. I do think that to the extent we are willing to attribute contents to very simple biological creatures, Millikan's account captures the way we do this as well or better than rival accounts.¹³

¹³I'd like to thank Kirk Ludwig and the participants of the Foundations in Cognitive Science seminar at UC Berkeley for comments on an earlier version of this chapter.

Chapter 6

Psychosemantics

6.1 Introduction

We have examined three attempts to give a more or less causal account of content. The moral I take from this examination is that these accounts must find some analogue of rationality in a causal structure if they are at all persuasive in the claim that the structure has semantic properties. The analogue to which each has appealed is a notion of function. It is an analogue because function is understood as an aspect of things whose existence and activities are guided by some benefit the system in which they are embedded may receive.

Jerry Fodor is in the midst of developing a theory of content for mental representations.¹ His account differs from the ones we've considered in that he rejects the appeal to function. One of his aims is to show why what he calls Meaning Holism is an incorrect view of the content of human thought, by developing a Semantic Atomist theory of content. I discussed Meaning Holism in Chapter 1 above, citing Fodor's as an extreme example of a theory that entails that what I called Content Rationalism is false.

In this chapter I discuss the details of Fodor's theory. My discussion is guided by two main concerns. First, Fodor claims that he can give a sufficient condition for one bit of the world to be about another bit of the world in purely non-semantic, non-intentional, that is to say, purely physical terms. I am concerned to show

¹I will be discussing three texts: *Psychosemantics* (Fodor, 1987), "Banish disContent" (Fodor, 1986), and "A Theory of Content", chapter 3 of a forthcoming book. I will call these PS, disContent, and TOC.

that this claim is incorrect. Fodor's discussion is vitiated by a systematic lack of attention to the sort of representational system to which the theory is supposed to apply, but I will show that the claim does not hold in the cases most favorable to the analysis, and fails even more badly for human intentionality.

Second, I want to evaluate Fodor's claim that Semantic Atomism is possible and that therefore something is wrong with Meaning Holism. I think that a very carefully circumscribed version of Semantic Atomism is possible; Fodor is right on this much. But this doesn't show that Meaning Holism is false; I think we can demonstrate that Semantic Atomism could not possibly be true for mental states like our own. The mistake is simple: from the fact that something could have one concept, rather like one of our own, in the way described by Semantic Atomism, it does not follow that something could have many concepts, and handle them the way we handle our concepts, where the content of all these concepts is determined atomistically. In any event one key part of Content Rationalism is true for either kind of creature: something has content only if the explanation of its existence and activities involves guidance by benefit to the creature.

Before I begin I think it is important to be clear on how different Fodor's overall project is from my own. My aim in seeking support for Content Rationalism has been to find grounds for the thesis that the mental is neither conceptually nor nomologically reducible to the physical, in particular for the thesis that there are no strict laws linking the mental and the physical.

Fodor's worry is something he calls "Irrealism," the thesis that there is something radically wrong with psychological explanation and hence that there are no such things as mental states. This thesis is alleged to follow from Meaning Holism. I see nothing radically wrong with psychological explanation, in particular nothing that follows from Content Rationalism, and I see no reason to hold that there are no such things as propositional attitudes. On the other hand Fodor and I agree that the mental is not nomologically reducible to the physical. We disagree on the reason; I assume he thinks that any roughly functionalist account of the mental entails irreducibility, and I do not think the reasons the functionalists offer are good ones.²

Here's how I will proceed. In section 2 I'll describe and motivate the theory of content in PS. Section 3 describes a distinction between what I call "perceptualist" contents and others, which serves to clarify the sort of theory of content Fodor could be offering. Section 4 shows why a theory of content that honors perceptual-

²See for instance (Block, 1980, 178).

ism does not “bring the mind into the world order”, and section 5 shows that there are non-perceptualist thoughts and why Fodor’s theory does not work for them either. In section 6 I will discuss Fodor’s responses to teleological accounts of content. Section 7 attempts to settle the general worry about intensionality raised by Fodor’s attack on teleological accounts. Finally in section 8 I describe what part of Fodor’s Semantic Atomism I accept, and argue that this is not enough to show that Meaning Holism is incorrect.

6.2 Fodor’s Theory of Content

Fodor believes that the only way to show that mental states are real is to show how they are part of the “world order”, that is, to show how they can be reduced to physical properties, or properties that are neither intentional nor semantic: “If aboutness is real, it must be really something else” (PS, p.97).

What is there, then, really? Fodor nowhere gives a list or a characterization of the allowable properties. His attack on the teleological accounts suggests that function properties are not allowed, nor are “optimality” properties, like conditions under which something performs optimally. I will assume that he also does not allow value properties, as for instance what is good for a system. He doesn’t actually say these are not allowed, but he does not make use of them, and it is not plausible that they will show up in the physicists’ “catalogue . . . of the ultimate and irreducible properties of things” (PS, p.97). Finally I will assume that we are not allowed to appeal to decision theory or to “inductive logic”, since either of those theories begs just the questions at issue.

Fodor does not aim to provide a complete physicalistic analysis of content. To show the Irrealist wrong it is enough to show how to describe a physicalistic sufficient condition for something to have semantic properties. The condition is not meant to be necessary.

The physicalistic notion central to Fodor’s account is causation. He is somewhat sensitive to the many difficulties with the notion of causation; he shrugs them off with the observation that it’s a notion that is central to the rest of science, so that if his sufficient condition is satisfactory on other grounds he will have shown that intentionality is at least “in the world order” as it is understood by the rest of science. I propose to allow him this out.

Fodor doesn’t attempt to justify the claim that content is essentially a matter of causation, beyond some schematic remarks to the effect that despite Chomsky’s

criticisms Skinner was right that meaning is a matter of causal control (TOC, pp.53-7). This is unfortunate, since while it is plausible that content has something to do with causation, the connection is clearly not as straightforward as Fodor thinks.

Fodor starts with what he calls the Crude Causal Theory: if a symbol is caused by all and only instances of a property, then it is about that property. If the symbol ‘horse’ is caused by all and only horses then it is about horses or expresses the property of being a horse.

The CCT is obviously unacceptable in both directions. Only horses: if a symbol is caused only by instances of the property it expresses then error is impossible. All horses: no symbol system produces a symbol for every instance of the property it expresses; we need a sufficient condition that can be realized by the things we know have semantic properties. Fodor’s energies are devoted to fixing these difficulties; the result is what he calls the Slightly Less Crude Causal Theory of Content, the SLCCTC. I will describe the two fixes at some length, and state the SLCCTC at the end of the Section.

Fodor calls the difficulty with error the “disjunction problem.” If error is possible then possibly a symbol is caused by something other than what it is about; yet if the symbol is about what causes it then it is about the disjunction of all the things that cause it. The way out of the problem, in Fodor’s view, is to distinguish between causal chains; the chain between the symbol and the property it expresses is different in some way from the chain between the symbol and other properties.

Fodor’s idea is to locate the difference in counterfactual properties of the causal connections. He notes a certain asymmetry that holds of concepts. Suppose someone misidentifies one thing as another: she takes a cow to be a horse.³ Doing this requires possession of the concept of a horse. But it is false that possessing the concept of a horse requires sometimes misidentifying cows as horses. Possessing a concept for horses requires the possibility of applying it incorrectly, but it does not require that it should be applied to one kind of thing incorrectly:

But for the fact that the word ‘horse’ expresses the property of *being a horse* ... it would not have been *that* word that taking a cow to

³What has misidentifying to do with symbol tokening? One of Fodor’s “operative assumptions” is that thoughts are relations to inner symbols in the Language of Thought (PS, p.97; pp.16-24; and the “Appendix: Why There Still Has to be a Language of Thought”, pp.135-154). The agent has a thought which involves the Mentalese symbol which means what ‘horse’ means in English. This inner syntactic string is caused by something other than a horse. I gave my reason for thinking that this “operative assumption” is not needed in Chapter 5, section 4.

be a horse would have caused me to utter. Whereas, by contrast, since 'horse' does mean *horse*, the fact that horses cause me to say 'horse' does not depend on there being a semantic—or, indeed, any—connection between 'horse' tokenings and cows. (PS, p.107-8)

Undoubtedly much work can be done to clean up and defend this asymmetry, but it seems to me there must be something correct about it. The challenge is to turn this obviously semantic asymmetry into a condition on the causation of symbols. Fodor does this by rewriting the asymmetry in terms of dependencies and causation. False tokenings of symbols depend on the existence of true tokenings, but not vice versa:

So, the causal connection between cows and 'horse' tokenings is, as I shall say, *asymmetrically dependent* upon the causal connection between horses and 'horse' tokenings. So now we have a necessary condition for a B-caused 'A' token to be wild [incorrect or false]: B-caused 'A' tokenings are wild only if they are asymmetrically dependent upon non-B-caused 'A' tokenings. (PS, p.108)

I will call this the AD account of error or wildness. Fodor doesn't justify the rewriting or introduce any new premises in his account; instead, he reconstructs the semantic asymmetry in causal terms and leaves the justification to hang on how well the result captures some semantic facts. The asymmetry is explicated first in terms of subjunctives and then in terms of possible worlds. B-caused 'A' tokens are asymmetrically dependent on A-caused 'A' tokens just in case if there were no A-caused 'A' tokens there would be no B-caused 'A' tokens, while if there were no B-caused 'A' tokens there would still be A-caused 'A' tokens; or just in case (PS, p.109):

1. A's cause 'A's.
2. 'A' tokens are *not* caused by B's in nearby worlds in which A's *don't* cause 'A's.
3. A's cause 'A's in nearby worlds in which B's don't cause 'A's.

Nearby possible worlds are just those in which "by and large" the natural laws that hold in the actual world hold there too.

I'd like to mention one objection simply to dismiss it. The AD account is rather complex; to some this is ground to object that they simply cannot make

sense of it.⁴ I see no special difficulties in understanding it. Causation, causal dependency, and nomological dependency are all notions we use in understanding the world, and notions any complete metaphysics must explicate. The difficulties are well known and enormous, but they are not difficulties special to Fodor's way of understanding content. Fodor's statement of the theory is paralyzingly vague but it's not obvious on that ground alone that the account cannot work.

Fodor is not careful in the way he phrases various points about the AD account of wildness. He claims it is both necessary and sufficient for wildness. Yet it is offered as one part of a *sufficiency* claim: 'A' means A if ('A's are caused by all and only A's). We are now working on the "only A's" in the right hand side. Since the overall claim is only one of sufficiency, there is no sense in which the current claim is even necessary for wildness. (Fodor seems to think it's necessary for any account of meaning, causal or otherwise, but this is demonstrably wrong; see Section 5 below.) If we consider only what is needed for the sufficiency claim, perhaps AD is necessary, but it is clearly not sufficient unless it entails the "all 'A's'" part of the sufficiency claim, and it does not. Fodor gives an argument for the necessity of the AD account, that there cannot be asymmetrical dependence when a symbol means the disjunction of the properties in the dependence relation, but I think he is wrong about this even in the context of the sufficiency claim. The difficulty hinges on whether the asymmetry is based in some facts about the representational system or facts about the world outside the representational system. If we suppose the former then Fodor appears to be correct. Suppose detecting horses is important to some creature, but the creature sometimes mistakes cows for horses. This creature might come to do better: leave off making this mistake, and hence more reliably detect horses. And if horses ceased to be important to the creature, it might well stop ever using the concept. But if we suppose the difference is external Fodor is incorrect. We have a disjunctive concept, jade, that applies to two different minerals. It could be true (although it's not) that there is an asymmetric dependence in the types of mineral, so, for instance, if there were no jade there wouldn't be any jadite, but not vice versa.

I turn now to the "All A's" part of the sufficiency claim. It's clearly false that all horses cause 'horse's in me or anyone else.⁵ Fodor calls this the "collateral information" problem. What distinguishes the horses that do cause 'horse's in me from the ones that don't is that I stand to them in an optimal epistemic relation:

⁴Millikan, for instance, in "Speaking Up for Darwin", in Georges Rey and Barry Loewer, eds., *Meaning and Mind: Fodor and His Critics* (forthcoming).

⁵The term 'horse' names a symbol in the Language of Thought.

they are within eyesight, I am paying attention, I know what horses look like, I don't have collateral information that what look like horses around here are really zebras, and so forth. We need some non-question-begging way of stating this condition. Fodor's solution has two parts. First, he appeals to the generalizations of psychophysics. These might hold that if any symbol system is placed in certain circumstances (eyes open, bright light, red wall at such and such a distance) it will produce a token of a symbol type that means 'red'; the guarantee that the circumstances can be described without begging any questions is provided by a presumption that psychophysics begs none of the relevant questions.

The second stage takes us from these "psychophysical" semantic properties to the rest. Suppose we want to know when protons cause symbols that mean proton. Clearly we don't simply sense protons, so it's unlikely that PROTON is a "psychophysical" concept. But we do detect protons using our senses. Protons will cause 'proton's in suitable test situations, ones where the presence of a proton will be manifested in some sensible display. Perceiving the display isn't enough for a thought of protons; the agent has to know that the display is connected to protons in a suitable way. Fodor's suggestion is that although this knowledge in us comes from our physical theories, all that matters for the meaning of the symbol 'proton' is that in fact its occurrences vary correctly with the presence of protons. So if someone who knows no physical theory is placed in a test situation in which the results are faked, and she still produces the symbol 'proton', it doesn't mean proton; but if she is somehow causally sensitive to the difference between the faked and the real situations, and produces the symbol only when the test is appropriately connected to protons, then symbol means proton.

The second stage is the theory of content with which Fodor supports his denial of Meaning Holism. For normal epistemic agents like ourselves the content of the term 'proton' will be determined by how that term appears in theoretically-mediated inferences. Fodor's claim is that what matters to the content is the causal connection, rather than the semantic properties of the inferential surroundings of the term. All that's needed is that there is some mechanism that takes the psychophysical traces of the evidence of protons and yields a token of 'proton'.

We can now state the SLCCTC. 'A' expresses the property A if (but not only if)

All instances of A's cause 'A's when (i) the A's are causally responsible for psychophysical traces to which (ii) the organism stands in a psychophysically optimal relation;

If non-A's cause 'A's, then their doing so is asymmetrically dependent upon A's causing 'A's. (p.126)

6.3 Perceptualism

Fodor is vague on the capacities of the systems to which his account is supposed to apply. He writes that what matters to him is a causal account of the semantic properties of human mental states, and that it doesn't much matter to him if it turns out that the intentionality of every other thing is derived from that of our mental states (PS, p.99).

The theory is intended to be atomistic: it gives a sufficient condition for a symbol to mean something, entirely independently of the meaning or presence of other symbols. The account might work for systems that have symbols that are in fact independent of other symbols in this way, for instance, extremely simple representational systems, ones that only, as it were, have one concept. It might also work for certain of our mental states, for instance ones that always stem directly from perception. (Whether there are any such states is a matter of some controversy, but present purposes do not demand a resolution.)

Let us define a kind of content that is atomistic in this way, and call it "perceptualist":

the content of a symbol is perceptualist iff the content of the symbol is not in any way dependent on the content of any other symbol in the representational system in which it occurs.

The point of this definition is that it allows us to ask whether Fodor's theory works for the kind of content it is built to work for. Fodor is eager to deny there are any non-perceptualist contents, on the basis of a coherent account of perceptualist content. But the strategy is flawed. It is one thing to show that there are perceptualist contents; it is quite another to show that non-perceptualist contents are impossible. I think we can demonstrate that most of our thought content must be non-perceptualist; that is the topic of Section 5 below.

The questions to ask now are, does Fodor's theory work, i.e., satisfy the reductive constraint stated at the start of the last section, for perceptualist states? and, does it work for non-perceptualist states? The answer to both these questions is 'no'; I'll examine the first in Section 4, and the second in Section 5.

6.4 Fodor's Theory for Perceptualist Contents

Fodor does not provide a non-question-begging sufficient condition for one bit of the world to be a perceptualist representation of another. The heart of the difficulty is that his condition entails no normative descriptions of the things to which it applies. There might be some type of state the instances of which are caused by a variety of things, where the causation by each of those things but one is asymmetrically dependent on the causation by the one remaining, yet it is false that there is any sense in which its occurrence is correct about the one or incorrect about the others.

Fodor will have a number of responses here; none avoids the difficulty. I will consider three.

A first response is that there is more to the account of content than the AD account of wildness: the state with content must derive from a "psychophysical trace" in a certain way. This is no help. Fodor seems to think the generalizations of psychophysics are an allowable resource in his project, but he is wrong. Psychophysics is not in the business of specifying the extension of the psychological predicates it investigates in purely non-intentional, non-semantic terms; rather it investigates physical conditions for states uncontroversially identified as psychological.

Psychophysics is also not, as far as I can tell, in the business of giving nomologically sufficient conditions for instantiation of thought types. Fodor seems to think it is, however. He asks whether perhaps psychophysics might be able to solve the collateral information problem for more complex concepts like horse (you will, after all, token 'horse' willy-nilly if brought into suitable contact with horses), and answers no, since psychophysics won't be able to guarantee that any particular organism has the concept of horse (PS, pp.116-7). But psychophysics also cannot guarantee that any particular organism will have any particular concept.

Fodor's second response: not just any state of the world is a candidate for bearing semantic properties. Once we settle which parts of the world count we will see where the normativity comes from. But how are we to do this? Fodor, as far as I can see, has no answer to this question that does not beg the question. He addresses it in two places, and gives the same inadequate answer. In *disContent* (p.14) he imagines a Quite Crude Causal Theory according to which being caused by water makes something a water-belief. He notes that the objection, that this makes mudslides and the growth of strawberries water-beliefs, is frivolous,

since nothing is a belief unless it has the functional role of a belief. Again, in articulating his conception of a teleological account of content (PS, p.105), Fodor uses the notion of mechanisms of belief fixation. This would be question-begging if we had no non-question-begging account of what a mechanism of belief fixation is. But, Fodor claims, we do: having a belief is simply a matter of having a certain causal role, the one given by a functional theory of believing. But this cannot help. Block's formulation of functionalism makes this clear. According to this formulation, something is a belief if and only if it satisfies a Ramsey sentence constructed from one of two types of psychological theory. The first type is the one constructed by taking as much of ordinary wisdom about psychology as we can manage. The second type gets its generalizations from a scientific psychology. But clearly there is no way to isolate this causal role except by starting with a psychological theory. Here's Fodor's own characterization of what the right functional role is:

So, suppose that a belief state is by definition one that causally interacts with desires and actions in the way that your favorite decision theory specifies; and that causally interacts with memories and percepts in the way that your favorite inductive logic specifies (PS, p.69)

The trouble is that functionalism is intended to settle a worry quite different from any Fodor now has in mind, the worry what relation mental states have with physical states. Functionalism says any physical state with a certain functional role is a mental state. It does not claim that functional role is reducible to physical terms; in fact functionalism claims it is not reducible.

Functional role would satisfy the need for normativity in the theory, since the relevant generalizations are normative. Even if it were otherwise acceptable it wouldn't help with perceptualist contents. Functional role includes the rich normative interactions our mental states have, and, as I'll argue below, perceptualism is false of contentful states with these interactions.

Functionalism is not a suitable candidate for this sort of job on a quite independent ground. Block and others, notably Putnam,⁶ argue that functionalism has a serious problem with "liberalism". Functional role is abstract causal role; the trouble is that the role is abstract enough that it is satisfied by anything: wheat fields, the Republic of China, the telephone on my desk.

Third (this is a response Millikan suggests on behalf of Fodor) these relations

⁶Hilary Putnam, "Appendix" to *Representation and Reality* (Putnam, 1988, 121-25).

suffice for content only when they obtain in intact organisms. Since the notion of an intact organism is a legitimate notion that biologists use, and biology is part of “the world order”, no questions are begged so far. Millikan will go on to say that the notion of an intact organism is a normative notion, and so it provides the requisite normativity for the account of content. Fodor simply does not use this strategy. He may be using the notion of an intact organism to partly locate the symbol side of the nomological dependencies, in the sense that the symbols must occur within the spatial bounds of organisms. But that doesn't solve the problem, since there will be interactions between the world and parts of organisms which lack semantic properties such that the interactions satisfy Fodor's theory of content.

Fodor attacks teleological accounts of content on a normative ground (p.106):

The teleology story perhaps strikes one as plausible in that it understands one normative notion—truth—in terms of another normative notion—optimality. But this appearance of fit is spurious; there is no guarantee that the kind of optimality that teleology reconstructs has much to do with the kind of optimality that the explication of ‘truth’ requires.

If Fodor is right then an account of content must somehow “reconstruct the optimality of truth,” or, less obscurely, provide a ground for thinking that its attributions have normative force. Teleological accounts do this with the notion of function. An appeal to decision theory and inductive logic does this only by failing to honor the central constraint on the problem at hand. There may be some other way to do it but Fodor does not describe it.

6.5 Fodor's Theory and Non-perceptualist Contents

The content of a mental state is non-perceptualist just in case something about its relations with other mental states helps determine what its content is. It wouldn't do to argue against Fodor that given the assumption that non-perceptualist contents are possible, we can show that his account is neither necessary nor sufficient, since his aim is to show how all content is perceptualist. But it's fairly simple to show that the assumption is true for things like ourselves. In this section I'll show this by way of a description of clear cases where Fodor's theory is not necessary and where it is not sufficient for the content of creatures capable of simple

inferences and for things like ourselves.

I begin with the argument against necessity. Suppose a creature is capable of a single inferential connection. It has a mental state whose content is that all B's are A's, as a matter of law, and mental states for A's, B's, and C's. Suppose that the AD account (and whatever else Fodor specifies) is satisfied by each of the pairs $\langle A, 'A' \rangle$, $\langle B, 'B' \rangle$, $\langle C, 'C' \rangle$. Suppose now that the creature makes inferences from B to A even when its 'A' token isn't directly controlled by an A's being A. The creature senses B-ness, does not sense A-ness, but the B is an A and the creature infers it is an A by its single inference rule. Furthermore, the inference rule is faulty: some B's are not A's. (There might be a selectional explanation for this of the form: the inference is reliable enough to give those that make it an advantage over those that don't.) Hence there are two ways for an erroneous 'A' to occur: caused by something else in the world (a C, perhaps), or caused by the incorrect inference.

Consider now the AD account of error. All three clauses are possibly false of this 'A' symbol type compatibly with it actually meaning A.

(i) A's should cause 'A's. Yet they may not. Suppose the causal connection between A's and 'A's is missing, because there no longer are any A's, but the inferential connection is still present. The creature at hand may be a descendant of those selected with this inference pattern. Fodor remarks,

To deny that [the AD account is] even *necessary*, you must accept the following as a possible case: We apply 'horse' to cows, and we would continue to do so even if we didn't apply 'horse' to horses; yet 'horse' means *horse*, and applications of 'horse' to cows are ipso facto false.

This doesn't look like a possible case to me. What on earth would make 'horse' mean *horse* in such a case? What would stop it from meaning *cow*? (PS, p.164, note 8.)

The answer is simple: content necessarily involves some kind of normativity, and the normative relations can sometimes determine content even when Fodor's causal relations are missing.

(ii) 'A' tokens should not be caused by B's in nearby worlds in which A's don't cause 'A's. The first case shows why this clause can be false; our creature is caused to produce 'A' tokens by inference only and never by A's.

(iii) 'A' tokens should be caused by A's in nearby worlds in which B's don't cause 'A's. If we have, for instance, selectional reason to think the creature is as

I've described it, then it has something that means A, even if 'A's aren't caused by A's. So there is no obstacle to holding that in a world where B's don't cause 'A's A's don't either, but 'A' still means A. Which tokens of 'A' mean A, if neither A's nor B's cause 'A's? The ones that are produced by inferences. The causal theory must at some point cope with the fact that very many of our thoughts are caused by other thoughts but are not about them.

Complicating the mental life of the creature so that it approximates our own just makes it easier to produce counterexamples to the necessity of Fodor's claim. Fodor has two kinds of answer. First, he can say that the account does work (given solutions to the problems of section 4) for uncomplicated creatures. Second, he can balk on the evaluation of the subjunctives or counterfactuals.⁷ I think this is a bad strategy. If the account is to work for anything at all there must be some way to evaluate the subjunctives or the counterfactuals that is clearly not question-begging. If there were a fruitful scientific theory that provided truths about dependencies of the sort the AD account requires, and whatever else is needed is in place, there would at least be hope for the account. If the evaluation of the subjunctives or counterfactuals depends on our intuitive sense of which worlds are closer than others, the account is hopeless. (Fodor seems curiously insensitive to this sort of concern. Millikan's idea that content is a biological category should be cashed out using other categories provided by biology. Any attempt to put intentionality into "the world order" will have to use some set of concepts legitimately employed in some theory of another part of "the world order".)

The difficulty for the necessity of Fodor's account is that causation and dependency are not the only factors relevant to content; for creatures like ourselves the normatively evaluable connections made between thoughts are relevant as well. This extra normative element shows Fodor's account isn't sufficient for content either.

We imagine someone who uses the words 'cow' and 'horse' as we do. She's an intelligent well-informed person who knows a wide variety of things about cows and horses: hamburger comes from cows, cows are now blamed for part of the greenhouse effect, and so on. What's odd about her is that Fodor's causal facts are exactly backwards for her sensory confrontations with cows and horses (but only her sensory confrontations). She calls the things at the dairy 'horse' and the things at the races 'cow.' Furthermore all the counterfactual details are in place:

⁷This was the substance of his response to Paul Boghossian in an "Author Meets Critics" session at the American Philosophical Association meeting in Washington, D.C., December 30, 1988.

e.g., if she were to leave off calling horses ‘cow’ (that is, tokening ‘cow’ when confronted with a horse) she wouldn’t ever call cows ‘cow’ either, but if she were to cease occasionally getting it right with cows she would persist in misnaming horses. Her sensory confrontations are odd, but every *other* use of the terms is just like our own use of the terms; for instance, she can *tell* us that she’s now having a visual perception as of a horse, although she knows and can tell us that it must be a cow that she perceives.

It seems to me that what matters to what her words and thoughts mean is what she knows about horses and cows; her extremely peculiar sensory incapacity does not show that her words and thoughts mean what they are related to by Fodor’s causal pattern.

Let me summarize the results so far. Fodor aims to give a sufficient condition that shows how some contents are part of “the world order.” I’ve argued that the condition is not sufficient for perceptualist contents. Further, there are non-perceptualist contents and the situation is appreciably worse for them. The main difficulty is with the normativity of content: the purely causal account provides no ground for claiming that “wild” tokens are incorrectly applied, and while non-perceptualist content has the requisite normativity, it also has the resources for generating counterexamples to Fodor’s theory.

6.6 Fodor’s Attacks on Teleological Accounts

In developing these objections to Fodor’s account I relied on two other theories of content, one roughly selectional, the other our intuitive understanding of what our thought is like and what it is possible to think. Obviously Fodor cannot object to the latter, since his project is to vindicate ordinary psychological explanation. But he is quite emphatic that what he calls⁸ teleological accounts of content are unacceptable. In this section I criticize his reasons.

The first “teleological” account Fodor works on in PS is Dretske’s learning account of belief as it is expressed in *Knowledge and the Flow of Information* (Dretske, 1981).⁹ Dretske’s idea, as Fodor reports it, is that if an internal structure

⁸Fodor’s terminology is unfortunate: he criticizes function based accounts of content and concludes that teleological accounts are unacceptable. But certainly on one way to understand the term ‘teleological’ our ordinary intuitive understanding of content is a teleological understanding.

⁹Dretske speaks of the function of beliefs in the book, but this idea is not stressed; so although Fodor doesn’t I shall call it one of the teleological accounts.

can be brought to carry information in the course of training, then that internal structure comes to mean what it carries information about. There can, by definition, never be misinformation. Mistake comes about through a combination of information and teaching. Fodor's response to this early Dretskean idea is, I think, essentially correct, but Fodor puts the point rather badly. The basic difficulty Fodor points to is that it's hard to see how teaching could possibly be a non-semantic, non-intentional notion. Fodor concentrates on the way the learning period is identified; he notes that it is hard to give a non-question-begging way of marking the end of the learning period. Fodor fits this sort of criticism into the procrustean bed of the disjunction problem. He claims that since the informational structure turns out to be capable of error after the training period has ended, it's clear that what it always carried information about was the disjunction of the property on which it was trained and all the other properties that could cause it, so it means the disjunction after all. The reason this is a procrustean fit is that Fodor loses sight of the distinction between decent teaching and no teaching at all. It's just false that someone has been taught to use the word 'apple' if, once they are let out of the classroom, they indiscriminately call apples and tomatoes 'apple'. Teaching someone a word must leave them able to make many discriminations between things, and in particular leave them making errors that are rationally explicable. So the difficulty is not simply the disjunction problem; it's more that Dretske's way of solving it is clearly question-begging.

Fodor's attacks on other teleological accounts are directed at a theory he articulated in "Semantics, Wisconsin Style." This gives a certain solace to some of his opponents, like Millikan, whose theories are significantly different. I think, however, that one of Fodor's ideas really is telling against the whole endeavor, the one captured in the remark I cited above in talking about normativity: for one of these accounts to succeed it must show how the normativity in function can precisely overlap truth in belief. Fodor doesn't make the point in precisely this form. He cites Stich, who makes the somewhat different point that simple representational systems are often designed to issue a lot of false positive representations (for instance cats are extremely skittish; recall also the discussion of the Venus Fly Trap in Chapter 3), since this is a better overall strategy than to wait to get the positives completely correct. This point is different because it seems that we *can* reconstruct *this* notion of correctness in teleological terms: those representations are correct that occur in the presence of some environmental property that contributes to an explanation of the presence and activity of the representation (type) via some benefit to the creature it occurs in.

In more recent work Fodor aims at a different, “principled,” argument for holding that teleological accounts cannot work. I want to show that this argument is incorrect.

A theory of intentionality is a second order theory that describes how to construct first order theories of the intentionality of particular organisms or groups. The first order theories ascribe particular semantic properties; they might say, for instance, that a certain alignment of a magnetosome in a magnetotactic bacterium means that anaerobic water lies in a certain direction from the bacterium. We should be able to determine from the second order theory the degree of intensionality that the first order ascriptions manifest. There might be no intensionality at all, if the ascriptions remain true under all substitutions of coreferring terms; or there might be the level of intensionality that propositional attitude ascriptions display, where no substitutions of coreferring terms are permitted. There is a lot of controversy about the importance of an understanding of intensionality. Some hold that if a theory of human intentionality is not intensional or fails to explain the intensionality of propositional attitude ascriptions, it is thereby shown incorrect. Others hold that intensionality is not important, and that perhaps we can give an account of thought that does not entail that ascriptions of thought are intensional.¹⁰ I agree, and Fodor agrees, with the former group.

Fodor’s principled argument against teleological accounts of content is that, while the resulting first order theories of content may be somewhat intensional, they are not intensional to the same degree as our ordinary ascriptions of propositional attitudes, and in fact the difference is major. The reason, he claims, is that function is indeterminate. Frogs have a mechanism that causes their tongues to flick in the direction of little ambient black things; if a spot of a certain angular displacement moving in a certain way registers on their retinas the tongue will flick. What is the function of this movement? Is to flick at little ambient black things, or is it to flick at flies? Fodor holds that natural selection cannot justify one answer over the other:

though you *can* describe the teleology of the frog’s snap-guidance mechanism [so the mechanism is for snapping at flies]—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

¹⁰John Searle is representative of the former group; I suspect that the vast majority of philosophers of language agree. Nathan Salmon is a representative of the latter group; see his book, *Frege’s Puzzle* (Salmon, 1986), for his reasons.

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." (TOC, pp. 71-2; Fodor is quoting from David Israel.)

There are, so far as I can see, two reasons Fodor has for this remark. First, abstracting from the remark, there are two plausible designs the insides of the frog might have, and nothing to determine which one to cite. The mechanism might be designed to get flies, or it might be designed to get little ambient black things, and the frogs have this mechanism because as it happens the little ambient black things around frogs are flies often enough for the frogs to survive and procreate.

A selectional account of content must rely on truths about what designs things have. These truths depend on explanations for the presence and persistence of the things which show how the creature containing them receives a selectional advantage through their presence. This puts a very strong constraint on claims about design. The explanation for why frogs now have this snap-guidance mechanism involves a benefit frogs receive when the snap yields a fly. The alternative Fodor offers could have been correct, but if it is correct then there is a much stronger story that needs to be told about this mechanism. A mechanism can be claimed to be designed for a less specific purpose only if there is some way to support the claim that the less specific task was relevant to its design. For instance, if the snap-guidance mechanism was designed for snapping at little ambient black things, then it must do this to other ends in the frogs, or in other relatives of frogs. Hence Fodor is wrong that the design attribution is indeterminate.

A related thought is that the claim that the mechanism is for snapping at little ambient black things does after all explain its persistence, since as it happens little ambient black things are flies in the normal world of frogs. This is like explaining a crater in a munitions factory by saying one of the red ones did it, since as it happens in the munitions factory red things are 16-inch shells. There is a perfectly good reason to prefer one of these explanations to the other. The selectional account appeals to explanations based on the aspects of things that matter to proliferation; being a fly matters to the frog's proliferation and being a little ambient black thing does not. The reason one matters and the other does not is shown by the fact that while all flies are little ambient black things, not all little ambient black things are flies, and many of the ones that aren't would be of no use

to the frogs.

Fodor's second reason for holding that there is no determinate answer starts with this response. There are many properties in the lives of biological creatures that are reliably instantiated together. In the frog's world all flies are little ambient black things and all little ambient black things (except the bee-bees we toss at it) are flies. Fodor concludes that no selectional account has the resources to choose one of the properties over the other, if they are correlated in this way, in attributing function or content:

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. (TOC, p.73)

His reason for drawing the conclusion is decidedly peculiar. The same sort of problem comes up in trying to decide what it is about a stimulus that an animal responds to in a conditioning experiment. We can solve this problem, with luck, by varying the stimulus. We try to figure out all the possible ways the stimulus could be affecting the animals, and then produce other stimuli that hold the various possibilities constant. Fodor says that the reason natural selection indifferently supports reliably equivalent content ascriptions is that Mother Nature can decide "only if she can perform a "split stimulus" experiment" (TOC, pp.86-7). The reason she cannot is just that the opportunity never arises for properties that are reliably coinstantiated in actual fact. But this is utterly specious. The way we *tell* what an animal is responding to is by performing split stimulus experiments; but the fact about what it is responding to is independent of whether split stimulus experiments are done or can be done by a certain sort of inquirer. The facts about what it is responding to depend on the right explanations of the way its sensory states are affected by environmental interactions. Similarly the facts about what function some system has depend on the right explanations of the way that system contributes to benefit for the creature. Perhaps, for instance, F and G are reliably coinstantiated; but the explanation for how the frog receives a benefit from things that are F and G goes through F rather than G.

Fodor isn't simply making the poor move from "what experiments Nature can perform" to facts about functions; he thinks there is a metaphysical ground for the claim that Nature does not care about the difference. The ground is that selectional theories are not permitted to use counterfactuals. (If they were the solution would be easy.) This is a baffling claim. Standard biological accounts

of the functions of things use counterfactuals all the time. Fodor seems to have run several issues together that have nothing to do with standard practice in biology: Millikan contrasts her account of the mind with purely dispositional accounts (but does not eschew counterfactuals); unrestrained reliance on counterfactuals is question-begging; and merely possible benefit has no consequences for fitness (TOC, pp.75-6).

6.7 Intensionality

How intensional are content ascriptions based on function or natural selection? If they are demonstrably not as intensional as propositional attitude ascriptions, that is a good reason to abandon the whole project, even if Fodor's demonstration fails.

What makes first-order function-based accounts of content intensional? There are three closely related features of these theories that are important. First, they all base content on explanation of some sort, and the explanations are based in turn on the notions of causation, of nomological connection, and of counterfactuals. Sentences reporting each of these connections are intensional.

Second, Dretske, Millikan and Fodor each iterate explanatory contexts. Fodor's theory is the most explicit about this: the notion of error is cashed out in terms of one nomic dependency depending on another. (It is interesting that iterated explanation is a hallmark of both these theories of content and many theories of function, as for instance Larry Wright's (Wright, 1976, 39). Still more abstractly, many feel that higher-order control is essential to thought; Davidson and Lehrer are two quite different representatives of this idea.¹¹) The point of iterating explanatory contexts is to distinguish between two things that come together in the actual world, as for instance the things to which a symbol is correctly applied and the things to which it is incorrectly applied.

Third, if I am right that any acceptable account of content along the lines of the theories of Dretske, Millikan or Fodor must make the explanations involve a notion of benefit to the organism, then there are further distinctions to be made. Many things do and could causally control the frog's snap-guidance mechanism, but what it is for depends on what causally controls it when it yields a benefit for the frog.

¹¹Davidson, in, for instance, "Thought and Talk" (Davidson, 1984c); Lehrer in "Metamental Ascent: Beyond Belief and Desire" (Lehrer, 1989).

These resources appear to suffice for ascriptions that are intensional within nomological necessity, i.e., that do not support substitution of actually coextensive but nomologically distinguishable predicates. Propositional attitude ascriptions are more intensional; for instance, ‘bachelor’ and ‘unmarried male’ are nomologically inseparable. This case does not, however, show the functional accounts cannot handle this degree of intensionality. These phrases are nomologically inseparable only considered as units. But our ground for thinking that a creature has a phrase like our phrase ‘unmarried male’ must be that the terms ‘unmarried’ and ‘male’ can occur independently. (I am assuming that the functional accounts of thought content will be able to make similar remarks about internal structures, if the creature is not a language user.) If they can occur independently then it is remotely possible that a creature has a belief about bachelors but fails to have the corresponding belief about unmarried males. Hence wherever two phrases are correlated with more than nomological necessity, but they have different semantically significant structure (where this fact is revealed by differences within nomological necessity), functional accounts can capture the difference. Similarly there are nomological differences even between phrases without semantic structure that mean the same; for instance terms with different linguistic origins can be known to have different linguistic origins.

I conclude that functional accounts do not have a special problem about intensionality. It is certain that they will not be able to handle intensionality without begging the questions we have been investigating; but they beg these questions long before the question about intensionality becomes important. It is also certain that a great deal of detailed work would need to be done to give a fully adequate analysis of intensionality using the resources sketched above.

6.8 Semantic Atomism and Content Rationalism

The main difficulty with Fodor’s account of content is that it lacks a clear description of the way content is involved with value. Suppose this lack could be remedied in some way. For the sake of definiteness suppose it is remedied in a very simple way: a state has the content that *p* provided it is causally controlled in the Fodor style by the fact that *p* and furthermore part of the explanation for its existence and activities involves a benefit to the creature. What does the possibility of such an account show about the thesis of Content Rationalism as sketched in Chapter 1?

In Chapter 1 I gave a discursive characterization of Content Rationalism. Here I will give more succinct characterizations of three versions of the thesis. I'll base them on one of Davidson's claims (from "Thought and Talk", p.157):

Even to wonder whether the gun is loaded . . . requires the belief, for example, that a gun is a weapon, that it is a more or less enduring physical object, and so on. There are good reasons for not insisting on any particular beliefs that are needed if a creature is to wonder whether a gun is loaded. Nevertheless, it is necessary that there be endless interlocked beliefs. The system of such beliefs identifies a thought by locating it in a logical and epistemic space. ("Thought and Talk", p.157)

I'll take the term "interlocked" to mean that content depends in some way on actual and potential interaction; it's hard to see why the mere presence of other beliefs would matter to the content of a thought. Interlocking in turn entails that there is some sort of process of rational change in contentful items (at least the sort of change that Dretske's learning account requires). A *minimal* Content Rationalism says that if something has the concept of a gun then thoughts involving this concept are involved in change through perception and action. Minimal Content Rationalism is compatible with there being no *other* interlocking. A *contingent constitutivist* claim is that if the creature is capable of making inferences involving this concept then precisely what concept it is depends in part on the pattern of inferences it makes. A *necessary constitutivist* claim is that if a creature has this concept then it is capable of making inferences to and from thoughts involving this concept.

I think we have already seen that the minimal claim must be true if Fodor's account is even to get off the ground. Further, the contingent claim is true as well; the arguments of section 5 show this to be so. Davidson holds the necessary constitutivist claim. I think Fodor is correct that his Semantic Atomism shows this claim is false, but its falsity leaves the weaker kinds of Content Rationalism untouched.

Semantic Atomism is the claim that what concept a creature has does not necessarily depend on what other concepts a creature has. I see no reason why, given the resources I sketched in the last section and at the beginning of this section, a creature cannot have a perceptualist concept with the same content as certain concepts we possess. All our concepts have various conditions of application such that in favorable cases we can tell whether they obtain. As with Fodor's discus-

sion of the proton/‘proton’ case what is needed for a perceptualist content is just a mechanism that reliably detects those conditions of application. ‘Reliably’ is a normative, not a statistical, notion: the mechanism reacts to instances of the concept enough for the creature to get on, and what explains its getting on is that the instance is an instance of that concept.

I concede this is an extremely implausible claim. Let me show why I accept it by showing why I reject four reasons for not accepting it. If in the end we decide against accepting it, then the *necessary constitutivist* claim is, after all, true of most of our concepts, and Fodor’s position is even weaker with respect to showing that Content Rationalism is false.

First, the requirement that the reliability of the connection should only be normative meets the objection that many of our concepts are such that reliably detecting their instances requires our full judgmental capacities, and so therefore a mechanism could not possibly be able to detect instances of that concept. (Our concepts of psychology are like this.) But a normative reliability doesn’t require that the mechanism is very good; only that the mechanism benefits only when it responds when the concept is instantiated.

Second, probably most of our ordinary concepts are such that the requisite normative reliability is just not nomologically possible. I cannot imagine how a creature could live off reliably detecting incumbents, and nothing else. This doesn’t show that atomistic conditions for content are impossible; it just shows that for a variety of essentially contingent reasons things cannot have certain contents atomistically.

Third, it may be urged that wherever there are analytic connections between concepts a creature cannot have one without the other. I do not see why this should be so even in the most favorable cases. Suppose someone reliably assents to ‘knowledge?’ in the face of instances of knowledge, yet her assents to ‘justified?’, ‘true?’, ‘belief?’, are normatively and statistically random in the face of instances. (It does not matter whether she applies these terms correctly to the things to which she applies the term ‘knowledge.’) It is extremely difficult to imagine how such a case is possible; but I do not see that it is impossible. If it is not, it appears to be a case in which someone has a concept but lacks the concepts analytically connected with it.

Fourth, it may be urged that no creature has *our* concept if it doesn’t make many or most of the connections that we make with the concept. If it has the concept of knowledge it must be able to infer that there is a true belief involved.

If it has the concept of a gun it must be able to infer that a gun is a physical object. This is essentially the same objection as the last: there are things we do with our concepts such that if something didn't do those things with its concepts it just wouldn't have the same concepts. The response is the same: if the extension is the same, fixed using the explanatory and normative resources above, why should we hold that the concept is different?

Last, a related difficulty that is harder to dismiss. My concept for guns is closely connected to my term 'gun.' My term is a noun; it contributes in various ways to the meaning of sentences that contain it in ways that connect with its being a noun. A "one concept" creature cannot be said to have thoughts with any structure at all; as a result there is no sense to the notion that its so-called concept is a concept at all.

What we need is a criterion by which to decide when two creatures share a concept. Many such criteria are possible; I shall describe a very weak one that is insensitive to issues about structure.

It seems to me that we can correctly say that "one concept" creatures do have semantically evaluable states, mostly because I think we can see when they go wrong. The marine bacterium moved to the opposite hemisphere swims in the wrong direction and dies; it is, I think, literally true that it swims in the wrong direction. How should we report what they are wrong about? We generate a theory based on what properties explain benefit. We state what they are wrong about using our language; we can tell from the second-order theory what about our report must be preserved if we capture content correctly, and what may vary. We know the bacterium has no syntax, and so syntactic distinctions in our report are irrelevant to what is reported. Now suppose that regardless of the way we report it one property is reported (the direction of anaerobic water). In this case I hold the theory shows we and the subject share a concept—even if we are otherwise as dissimilar as we and bacteria.

I have given some reason to think that Fodor's atomism is acceptable for a limited range of concepts when realized in perceptualist contents. His atomism cannot be accepted for things which have thoughts like our own. Fodor seems to think that if an atomistic semantics for the terms of a language of thought can be given, it is essentially trivial to give the semantics for the entire language: we derive the truth conditions for the sentences from the content of the terms using a recursive truth theory. This idea makes no sense. It makes no sense for creatures that it was designed for, and it makes no sense for us.

We can imagine a creature which has a syntax but no inference and no practical reasoning: a creature which can generate structured representations, but which cannot make any rational connections. There is a simple and perfectly formal reason why Semantic Atomism cannot be right for such a creature. If the meaning of the syntactic compounds is derived from the meaning of the terms, then the meaning of the terms must be suitable for generating complex meanings; otherwise the structures would be no more than groups of independently meaningful entities.

The reasoning of Section 5 shows why the idea makes no sense for us. The meanings of our terms simply do not derive solely from their causal relations with things in the world. Once a creature is capable of inference and practical reasoning the meanings of its terms will depend on how they show up in that reasoning, as well as on their relations to the world.

Fodor has apparently made a simple blunder. He argues that for many of our concepts possibly a creature could have that concept and lack the other concepts we normally connect with it in thought. From that it does not follow that possibly a creature could have many of our concepts in this way; and it certainly does not follow that possibly our thought derives its content from the world in this way.

Hence the minimal thesis of Content Rationalism is true, and the contingent constitutivist thesis is as well. I conclude that Fodor has not shown how to understand content without using a notion of rationality.

6.9 Conclusions

Fodor has attempted, and continues to attempt, to give an account of intentionality that puts it “in the world order,” i.e., reduces it to non-intentional, non-semantic properties. In particular he aims for a non-teleological reduction as well. I think he is right that the functional accounts of content that have been offered fail in various ways, centrally in not demonstrating that there is any way performance of function requires truth. I think he is wrong in his “principled” attack on functional accounts, and I think his own theory fails because he does not appreciate what the functional accounts have right: some way for the theory to entail that a particular bit of the world is incorrect about another.

That concludes my investigation into accounts of content that aim to show why Content Rationalism is false or that aim to explain why Content Rationalism is true in terms that are demonstrably different from the notions used in articulating Content Rationalism. There are many other such accounts and the activity in

the field suggests there will be more. But the difficulties with the ones I have considered suggest a general conclusion. Purely causal accounts of content are impossible; causal relations alone are never right or wrong. Supplementing purely causal accounts always involves some notion of benefit or value or optimality, such that things have content only if there is some explanation for their presence and activities that connects these with a benefit for a creature which contains them. This explanation is a close relative of the full-blown conception of rationality we find in decision theory; so close that such a theory of content can only be seen as vindicating an analytic connection between representation and rationality, and not as a discovery of what content really is that shows why we associate the concepts of representation and rationality as closely as we do.¹²

¹²I have benefited from discussions on aspects of Fodor's work with Elisabeth Lloyd, Kirk Ludwig and Dugald Owen.

Chapter 7

Psychophysical Laws

7.1 Introduction

We now have two theses about content and meaning in hand. Consider the class of theories of meaning that includes Davidson's and the theories of Stampe, Dretske, Millikan and Fodor. Each is a secondorder theory that shows how to construct a firstorder theory, a theory of a particular organism or person, that specifies the content of certain items (utterances or inner representational states). The secondorder theories can be ranked according to the complexity and nature of the systems to which they apply, and according to the power of certain explanatory concepts they use. Theories for the simplest representational systems, like the magnetotactic bacterium, lie at one end of the ranking; theories of meaning for natural languages lie at the other. The first thesis is that every theory in the class must use value concepts. The second thesis is a more particular version of the first, for theories that apply to representational states like our own: such theories must use the concepts of rationality.

Both theses are theses about the concepts of content and meaning: necessarily, if something is a person with propositional attitudes then it is rational; necessarily, if something has representational states of any kind (simple representations to language) then these states stand in some explanatory relation to a good for their possessor. A different way of saying the same thing is this: content terms are not conceptually reducible to a set of terms that does not include any for value properties.

Either of these two theses marks psychological explanation as different from

many other kinds of explanation, provided we are assured that the concepts used in the other kinds of explanation do not similarly involve the concepts of value or rationality.

Davidson argues that conceptual irreducibility entails an even sharper difference in kinds of explanation.¹ The conceptual irreducibility entails nomological irreducibility: there are no strict laws linking terms for the propositional attitudes with terms from, say, physics, or some other basic science into which we might hope to reduce the rest. This is the thesis of the anomalism of the mental, the thesis that there are no strict psychophysical laws.

In this Chapter I evaluate Davidson's claim. The considerations Davidson advances, and some of their implications, suggest very strongly that there are no strict psychophysical laws, but I believe Davidson provides no conclusive argument that there cannot be. The reason is simple: except in cases where the concepts more or less explicitly rule out the possibility of a law (e.g., there cannot be a law linking being red all over with being green all over), whether a universal generalization is a law is up to the world.

In the next section I review Davidson's argument for Anomalous Monism. In section 3 I describe the structure of the argument for anomalism. I show that the crucial premise (that mixing predicates "not made for one another" bars lawlikeness) is not correct. In section 4 I consider whether Davidson's support for psychophysical anomalism might nevertheless provide us reason to think the mix for psychophysical generalizations does bar lawlikeness; I conclude it does not. Section 5 describes and evaluates three further arguments based on the idea that if there were strict psychophysical laws then physics could contain a theory of rationality. I believe this idea is not correct, but these three arguments show how peculiar the laws would have to be if there were any; I take this peculiarity to show that we have very good reason to think there are no psychophysical laws.

¹The argument I will discuss occurs in "Mental Events" (Davidson, 1970). I will refer also to "Psychology as Philosophy," "The Material Mind," and "Hempel on Explaining Action" (Davidson, 1980g,f,d); I'll call these 'ME', 'PP', 'MM', and 'HEA'. Davidson has at least two other arguments in print. One, relying on the notion of a "causal concept," occurs in "Freedom to Act" (Davidson, 1980c); the other, based on the fact that there is no causal *analysis* of action, occurs in PP. I will not discuss these arguments here.

7.2 The Argument for Anomalous Monism

Anomalous Monism is the thesis that every mental event is identical to a physical event, but that there are no laws linking mental predicates and physical predicates. The argument has three premises (ME, p.208).

- (1) The Principle of Causal Interaction: at least some mental events interact causally with physical events.
- (2) The Principle of the Nomological Character of Causation: if two particular events *c* and *e* are related by causation, then *c* and *e* are related by a strict deterministic law.
- (3) The Principle of the Anomalism of the Mental: there are no strict deterministic laws on the basis of which mental events can be predicted and explained.

These premises can seem to be inconsistent: for instance, if mental events interact causally with physical events, and causation requires strict laws, then there must be strict laws on the basis of which mental events can be predicted and explained. The solution to the seeming inconsistency is the observation that events, like all particulars, can be described in many ways. True lawlike statements use only some predicates satisfied by the events they cover. In light of this observation we rewrite the second premise slightly. Say an event *e* instantiates a law under predicate *d* if *e* satisfies *d* and the law uses the predicate *d*.² Then if *c* causes *e* then there is a strict deterministic law which *c* and *e* instantiate under predicates *d*₁ and *d*₂. Clearly *c* and *e* satisfy many other predicates as well under which they instantiate no law. The third premise then can be rephrased to claim that mental predicates are not ones under which events instantiate strict deterministic laws.

The argument for monism then runs as follows. An event is a physical event just in case it has a physical description. A law is a strict deterministic law only if it is phrased only in physical descriptions. If a mental event causes a physical event, the mental event must instantiate a strict deterministic law. Hence there must be a physical description of the mental event. Hence the mental event is the same event as some physical event.

²I follow Davidson in holding that laws are sentences, rather than parts of the world. A sentence is a law only if the world is a certain way. Sentences have predicates as components; in this sense a law relates predicates. We may also say a law relates properties, if a lawlike sentence is true that contains predicates that express the properties.

Various reasons for thinking the conclusion of this argument is false have been given.³ The most likely suspect in the argument is premise 2, the claim that causation requires strict laws.⁴

If the argument as stated does not work, there is a variety of arguments that share its general form. Suppose we hold on to premise 1. Then we need variations on premise 2 and premise 3 such that premise 2 supplies some plausible condition on causation while premise 3 claims that true statements of empirical regularities involving psychological predicates do not satisfy this condition. For instance, a necessary condition on causation might be that if c causes e then c and e have aspects linked by some regularity, where the regularity is not phrased in any terms whose analysis reveals a connection with value properties. We have already established that psychological regularities would not satisfy this condition.

Whether any of these arguments succeed, premise 3 in its original form is independently interesting. I think many people believe it (Searle and Grice, for instance, and Ned Block holds that functionalism entails it), but few produce arguments in its behalf.

7.3 The Argument for Anomalism

The structure of the argument of ME is this:

1. a necessary condition for a sentence to be lawlike is that the vocabulary the sentence uses should be drawn from a theory that contains strong constitutive elements or synthetic *a priori* laws.
2. the synthetic *a priori* laws of physics differ from the synthetic *a priori* laws of psychological theory.
3. sentences that draw their vocabulary from theories with differing synthetic *a priori* laws are not lawlike.
4. Therefore psychophysical generalizations are not lawlike.

³John Haugeland, “Weak Supervenience” (Haugeland, 1982); Jennifer Hornsby, “Review of *Essays on Actions and Events*” (Hornsby, 1982, 84) (see also the doctoral dissertation of Olav Gjelsvik for a detailed elaboration of a similar idea); Tyler Burge, “Individualism and the Mental” (Burge, 1979).

⁴G.E.M. Anscombe, “Causality and Determination” (Anscombe, 1971); Noa Latham, “Singular Causal Statements and Strict Deterministic Laws” (Latham, 1987).

Premise 1 is found in this suggestion of Davidson's,

the existence of lawlike statements in physical science depends upon the existence of constitutive (or synthetic) *a priori* laws like those of the measurement of length within the same conceptual domain. (ME, p.221)

Earlier in ME Davidson lists three conditions ordinarily thought to be necessary if a generalization is lawlike:

Lawlike statements are general statements that support counterfactual and subjunctive claims, and are supported by their instances. (ME, p.216)

In premise 1 Davidson is offering the first of two additional necessary conditions on lawlikeness. (Premise 3 offers the second). This condition is correct for any genuinely quantitative science. If a scientific theory contains laws that relate numeric measures of quantities, then there must be a representation theorem which shows how the behavior of the empirical system can be mirrored in certain relations in the numbers. The representation theorem establishes the ways that numbers may be applied to the empirical situations the theory is supposed to describe. If the theory is *genuinely* quantitative then certain more or less natural mathematical relations in the numbers can be used to infer new facts about the empirical system.⁵ (By contrast, we use numbers to keep track of bus lines, but there is no natural mathematical relation between, say, the numbers 53 and 21 that reflects the differences between those two routes.) These relations, read as logical relations among the predicates of the science, make up the strong constitutive elements of the science. Davidson discusses the relation *longer than*. This relation is transitive, asymmetric, and nonreflexive. These features are required if we are to give numeric measures of length that have the features of the measurements we make.

I do not think premise 1 is a strictly necessary condition on lawlikeness. Denying the condition requires that it is possible for there to be a law relating two predicates, such that the predicates have no such relations to other, similar predicates. I doubt whether any interesting science could be based on such a law.

Support for premise 2 comes from a characterization of the synthetic *a priori* laws of psychology and their differences from the synthetic *a priori* laws of

⁵See Patrick Suppes and Joseph Zinnes, "Basic Measurement Theory" (Suppes, Patrick and Zinnes, Joseph, 1963, 48).

physics. Davidson thinks of a theory of meaning for an agent as a theory that offers measurements of the agent's characteristics in the form of statements of what her words mean; these entail statements of what she believes or wants or intends (i.e., a characterization of her psychology) when supplemented with statements about her sentential attitudes. The synthetic *a priori* laws of this measurement theory are the laws of rationality, since the relations in the system of objects in which we represent the empirical relations of the agent are the rational relations among our own sentences.

The analogy with measurement in physical theory is not perfect. Psychology, thus conceived, differs from other sciences in that the system of objects in which we represent empirical occurrences does not have the clarity of structure that the systems used by other sciences have: we can prove representation theorems for the measurement of temperature, but there is no such proof forthcoming for the measurement of propositional attitudes.

In Chapter 1 I indicated my reasons for skepticism about Davidson's claims that the propositional attitudes must be rational, including this claim that psychological attributions are a kind of measurement. In Chapters 3 through 6 I gave an independent reason for thinking the attitudes must be rational. Two reservations must be noted. First, we have not mitigated the conditional character of the conclusion of Chapter 1: if we are inclined to think of content as a *sui generis* property, there is nothing to stop us. But we will have no interesting characterization of the synthetic *a priori* laws of psychology. Second, the arguments of chapters 3 through 6 do not yield a conclusion quite as strong as Davidson's arguments that, for instance, the axioms of decision theory are among the synthetic *a priori* laws of psychology. They only show that the explanatory relations that suffice for content can only be described in terms of various rational relations, like the relation of one belief being a good reason for another. I did not specify precisely how these notions would have to be used. So for instance we might allow that an agent's preferences are literally not transitive, provided they are close enough, where we must use the ideal of perfect transitivity in an explanation of the deviation.

I think it is clear that the sciences into which we might seek a reduction of psychology are not governed by synthetic *a priori* laws like these. Being an electron, for instance, does not require standing in inferential relations to any other kind of particle or force.

Premise 3 is Davidson's second necessary condition on a sentence's being lawlike. He gives a single example to show why it is a necessary condition (ME,

p.218). Nelson Goodman urges that the sentence “all emeralds are grue” is not lawlike (Goodman, 1954, Section III). Davidson agrees, but notes that we can construct lawlike sentences using predicates like ‘grue’ so long as we do not mix them in this way. “All emerires are grue” is both lawlike and true.⁶

Davidson says that the terms ‘emerald’ and ‘green’ are “made for each other”, as are ‘emerire’ and ‘grue’, whereas we know *a priori* that ‘emerald’ and ‘grue’ are not made for each other. Can we cast this remark in terms of ‘strong constitutive elements’? Suppose we collect predicates by our conceptions of causation and change. Take a single predicate; then add to the set any predicate P such that if any from the set applies to a thing, then a change in whether P applies to it counts as a change that requires an explanation involving causation. It’s hard to capture what lies in the resulting set with any simple principle (like the transitivity of length), but the members of the set are “made for each other” in the relevant way. Then if emerald were grue then it would change color i.e., undergo a change in the applicability of predicates from the set of predicates that define the notion of change along with the term ‘emerald’ at a certain point in time, but there would be no cause for the change. This would violate an even more fundamental synthetic *a priori* law, that objects do not undergo changes without a cause. When an emerire goes from green to blue it remains grue throughout; so the more fundamental condition on causation is not violated.

Davidson’s reading of Goodman’s case does not demonstrate that Premise 3 is a necessary condition on lawlikeness. Suppose there is an empirical system with various relations. Two of the relations are denoted by the twoplace predicates F and G. Suppose each predicate supports quantitative measures, as the two place predicate ‘is longer than’ supports the measurement of length by numbers. Suppose the synthetic *a priori* laws governing these predicates are different: one is transitive and the other is not. Suppose finally there is a law linking F and G.

These suppositions are consistent. Consider three cases. First, suppose F is transitive and G is not, and the law is $\forall x\forall y(Fxy \rightarrow Gxy)$. If the only pairs of objects in the extension of G are F as well, then the transitivity of F would be “transmitted” through the law.⁷ That is, given any three objects x, y, and z, if Fxy and Fyz then Fxz, by transitivity; then Gxy, Gyz, and Gxz, by the law; hence G would be transitive. But the transitivity of F need not be “transmitted”, if there are some pairs of objects which are G and not F, for which the transitivity of G

⁶“Emeroses by Other Names” (Davidson, 1980a).

⁷See Jaegwon Kim’s “Psychophysical Laws” (Kim, 1985, 375) for the idea that laws can “transmit” constitutive properties.

fails. Second case: suppose F is not transitive but G is, and the same conditional is a law. Then given any three objects x, y, and z, if Gxy and Gyz, then Gxz, by transitivity; but one or another of these pairs might not lie in the extension of F. Finally, suppose the law is a biconditional. Neither of these strategies work to show that transitivity wouldn't be "transmitted" by the law. Still, there could be a law that does not violate the synthetic *a priori* laws. Synthetic *a priori* laws hold with greater modal strength than the nomic biconditional. So perhaps the extensions of F and G differ in some counterlegal situation.

Synthetic *a priori* principles need not be transmitted even if we add monism to the body of suppositions. Take the case where F is not transitive and G is. Every pair of objects that is G also has a description in terms of the theory from which F is drawn. We can construct a new predicate F' using these descriptions that covers just the G things. The sentences

$$\forall x\forall y(F'xy \rightarrow Gxy)$$

and

$$\forall x\forall y(F'xy \leftrightarrow Gxy)$$

are true, but they may or may not be laws. So while F' is, as we might say, materially transitive, it may not be nomically transitive. If it is nomically transitive then the transitivity of G has an "echo" in the theory from which F was drawn. The strength of the echo depends on how many laws are required to get a nomically necessary condition for being a G. If there are two or three laws then the echo is strong; if there are thousands or endlessly many then the echo is very or vanishingly weak. Whether the Fsystem would be taken to reduce the property of being G by practicing scientists depends on the strength of the echo; if it is very weak, perhaps it would not, even though there is an exhaustive nomic biconditional.

Notice finally that there is more room for synthetic *a priori* principles to fail to be transmitted in the case where they are known to have exceptions, or where their consequences cannot be described with precision. This is the case with the synthetic *a priori* laws of psychology. Consider beliefs whose contents have the form, x is longer than y. The transitivity of *longer than* supplies a norm of reasoning: if someone believes that x is longer than y and believes that y is longer than z, then she ought to believe that x is longer than z; and if she *fails* to have the latter belief along with the former beliefs often enough then there is reason to doubt she does have the former beliefs. If there were a simple law running from physical conditions to beliefs of this form, the norm of reasoning need not show up in the physical predicates, simply because the inference is not always made.

Even if Premise 3 is not strictly necessary for lawlikeness, it may be that certain mixes of synthetic *a priori* laws preclude lawlikeness, while others do not. How would we determine whether a psychophysical universal generalization is lawlike, if we cannot tell straight off from the quasilogical principles that relate the predicates involved?

The reason “all emeralds are grue” is not lawlike, according to Goodman, is that it is not supported by its instances. A generalization is not supported by its instances not confirmed if its observed instances do not incline us to “project” the generalization to unobserved instances: actual but unobserved instances and hypothetical instances. Every green thing is also grue. We are not willing to project the generalization because we think any emerald we inspect would be green even if it were not observed, and that it will remain green and not turn blue.

Goodman thinks that the difference between projectible generalizations and the others stems from the “entrenchment” of their predicates in explanatory practice. Entrenchment is not an independent criterion of lawlikeness: there could not be a universal generalization that was confirmed by its instances and which supported counterfactual claims that failed to be lawlike because its predicates were not entrenched. Hence whatever consequences entrenchment has for lawlikeness it has through these other aspects of lawlike claims. Davidson concurs: if a true universal generalization is supported by its instances and supports counterfactuals, then it is a law. If a universal generalization is not lawlike, because its predicates are drawn from theories with nomically incompatible synthetic *a priori* laws, then it must also fail to be supported by its instances, or fail to support counterfactuals. Hence if psychophysical laws are not possible then we should be able to demonstrate this by showing how psychophysical generalizations do not support counterfactuals or do not project to unobserved cases.

It would not suffice to show that psychophysical generalizations are not lawlike to point to possible failures in the generalizations. Every natural law is such that possibly it is false, since nomic necessity is weaker than logical necessity.⁸ Nor would it suffice to demonstrate that one or several psychophysical generalizations are not lawlike; the task is to show that all of them have a problem.

The situation won't be as clearcut as with the generalizations that demonstrate the lawlike/nonlawlike distinction. The generalization, “everyone in this room has \$.60 in their pocket,” fails to support counterfactuals because we know that if

⁸George Myro based a general criticism of Davidson's argument for anomalism on this observation. I think Davidson's argument is *prima facie* good enough that this cannot be the central difficulty with it.

someone with \$.61 in their pocket were to come into the room they wouldn't lose a penny. A psychological generalization could not fail to support counterfactuals in such an obvious way, since it is almost always true that we can adjust other attributions to make room for the attribution of a certain belief to someone. So the best we can expect is that psychophysical generalizations would generate too much *ad hoc* adjustment for comfort. The flavor of the kind of failure is given in Davidson's remarks about the nature of constitutive laws. He asks what we should say if we observe a triad of objects that violates the transitivity of the 'longer than' relation. In a very strong sense we do not know what we should say. We could give up the constitutive law, or we could give up the empirical tests that determine of a pair of objects which is the longer, or we could give up the assumption that the objects are rigid. We are not forced by the facts to do any of these.

I will assume the generalizations we are investigating are universal conditionals with only physical predicates in their antecedents and only psychological predicates in their consequents. Anything we can conclude about these conditionals we can most likely conclude about conditionals that run from psychological to physical properties.

7.4 Support for the Failure of Lawlikeness

After describing synthetic *a priori* laws and the way they bear on lawlikeness, Davidson sums up his argument that mental and physical predicates are not "made for one another", and why we cannot hope for psychophysical laws, in the following central passage in ME (pp.222-3):

[The irreducibility is not due] to the possibility of many equally eligible schemes, for this is compatible with an arbitrary choice of one scheme relative to which assignments of mental traits are made. The point is rather that when we use the concepts of belief, desire and the rest, we must stand prepared, as the evidence accumulates, to adjust our theory in the light of considerations of overall cogency: the constitutive ideal of rationality partly controls each phase in the evolution of what must be an evolving theory. An arbitrary choice of translation scheme would preclude such opportunistic tempering of theory; put differently, a right arbitrary choice of translation manual would be a manual acceptable in the light of all possible evidence, and this is a choice we cannot make.

Any psychophysical generalization would interfere with this opportunistic tempering. It would, perhaps, work up to a certain point in time. At that point, the evolution of the psychological theory would lead to divergence from what the psychophysical generalization says. This temporal claim has a modal analogue: since changes in the life of an agent force changes in psychological theory, it is likely that any small difference between actual circumstances and a counterfactual circumstance will be reflected in the applicability of a psychological theory.

There are two ways to understand the suggestion that a psychological theory is necessarily an *evolving* theory. Neither shows that there cannot be psychophysical laws.

Perhaps psychological theory is such that a theory for an individual is never complete. New data always lead us to revise old attributions. For Davidson, this thought should be understood as follows. At time t_1 an interpreter generates a truththeory T_1 for the language of an agent (call him Sandor). The interpreter also generates a list of Sandor's sentential attitudes; the theory and the list together entail a set of particular propositional attitude ascriptions. At t_2 the interpreter has a new collection of utterances and behaviors to interpret. (Sometimes new data lead us to revise our list of sentential attitudes; considerations of overall sense can lead us to reinterpret old observations.) The combined body of evidence (what the interpreter had at T_1) plus new evidence until T_2 plus revisions generates a new truththeory T_2 and a mental economy such that some of the earlier propositional attitude ascriptions are revised.

For instance, at t_1 Sandor says, "I'd like a Coke," to his colleague going out to the store. A week later the interpreter learns Sandor has long been a rabid nutritionalist and he firmly holds Coke is good only for taking paint off cars. Arbitrarily choosing the earlier translation manual and holding it fixed would yield a decision that Sandor is quite irrational. But now we have enough of a story about him to make it come out that he *cannot* have wanted a Coke. A psychophysical generalizationone that claims the physical state Sandor was in at t_1 is sufficient for the desire for a Coke has broken down.

Perhaps this sort of theory revision happens more often in psychological theorizing than in other kinds, but it certainly does happen in other branches of science. In fact this pattern is simply the pattern of improving a theory; the earlier versions were false, and the later ones were improvements.

The other way to understand why psychological theory is necessarily evolving theory is based on the possibility of change in meaning. What a particular agent

means by her words often changes through time. At t_1 we use an interpretive truth theory T_1 to interpret her. By t_2 her patterns of usage have changed. Perhaps even her ways of generating complex utterances from simpler parts have changed. We need a new theory T_2 to interpret what she now says and thinks. Both theories are true provided they are indexed to the periods during which they serve to interpret. Presumably it is not necessary that every agent changes her language in this way, but it is likely enough.

The problem, however, is not whether a single interpretive theory could be acceptable in light of all possible evidence about the entire life of an agent, but whether a set of psychological laws could track the changes, if they occur.

Let us revise Davidson's claim in the following way. We imagine a "physical scheme" of interpretation, a set of strict laws which generate propositional attitude attributions to a person from physical descriptions of her and her surroundings. The "ordinary scheme" derives propositional attitude attributions the way the radical interpreter generates them: beginning with a base of evidence about lawlike correlations between occasions of holding sentences true and things true on those occasions, the interpreter confirms a truth theory for the person's language, and in so doing manages to show the person as rational as possible, by showing what she believes and wants and intends. Suppose the "physical scheme" agrees with the "ordinary scheme" until a certain time t_1 ; afterwards the ascriptions disagree. We can, perhaps, still understand the ascriptions made by the "physical scheme," but they require us to posit more (or less) irrationality on the part of the agent than we think is necessary.

Davidson claims that we must always be prepared for such a divergence: not simply for the possibility that we are wrong, but for the fact that no generalization will hold up.

Is this claim correct? Suppose we set out to confirm a universal generalization that is a candidate for one of the laws in the "physical scheme." As Davidson writes,

rationality is a trait that comes and goes. . . even the most rational person will often do things for poor reasons, while, as I have said, the looniest action has its reason. (HEA, p.2667)

Imagine that we begin the construction of a "physical scheme" for an agent whose rationality stays on an even keel. It is intelligible to suppose that the physical scheme generates interpretations that have *constant* rationality. That is, were an

agent in the mental states the generalizations predict the agent's rationality would in that situation diverge from the ideal by a constant amount. Imagine plotting rationality on the y-axis and time on the x-axis. The "physical scheme" is consistent with a large number of evolving systems of psychological states: the plot for all these systems is a straight line parallel to the x-axis. (Obviously there is no such graph, since there is no clear sense to attach to the idea of a quantitative measure of rationality. But we need something like this metaphor to make sense of the idea of that holding something fixed would violate the synthetic *a priori* laws of psychology.) Now imagine our agent becomes more normal: her rationality begins to wax and wane. Plotting her rationality yields a curve that overlaps this straight line. Clearly our "physical scheme" is no longer acceptable.

If we can understand the idea that a "physical scheme" holds a person's degree of rationality constant, we can certainly understand the idea of a "physical scheme" displaying an agent as rational. But now there doesn't seem to be anything contradictory in the supposition that such a "physical scheme" might simply follow the "ordinary scheme": delivering just the same attributions as those we would make given perfect information of the sort we usually use in interpreting. Nor is there anything obviously contradictory in the supposition that the attributions of the physical scheme support counterfactual claims, or the supposition that we might count a psychophysical universal generalization confirmed when confronted with enough instances. Hence it doesn't seem impossible for a fixed set of psychophysical laws to deliver just the interpretations we would make.⁹

7.5 Rationality in Physics

In this section I consider three arguments that stem from the central argument of ME. The first is a reformulation that Davidson offers in other papers. The second is a version of this reformulation that John McDowell urges against Brian Loar. The third is suggested by other things Davidson says about meaning. All three fail in essentially the same way: they hold, I believe incorrectly, that if there are strict psychophysical laws then rationality is a synthetic *a priori* law of physics, or that rationality can be exhaustively and precisely characterized in physical terms. Even though these arguments are not compelling they show how peculiar psychophysical laws would be if there were any.

⁹Brian McLaughlin comes to a similar conclusion in "Anomalous Monism and The Irreducibility of the Mental" (McLaughlin, 1985).

In *MM* Davidson considers whether knowing everything there is to know about an artificial man would gain us an understanding of his psychology that we do not presently have about our own. He notes that arriving at the psychology of the artificial man would involve precisely the same process as we currently use to understand each other:

Our standards for accepting a system of interpretation would also have to be the same: we would have to make allowance for intelligible error; we would have to impute a large degree of consistency, on pain of not making sense of what was said or done; we would have to assume a pattern of beliefs and motives which agreed with our own to a degree sufficient to build a base for understanding and interpreting disagreements. These conditions, which include criteria of consistency and rationality, can no doubt be sharpened and made more objective. But I see no reason to think that they can be stated in a purely physical vocabulary. (p.259)

A similar formulation occurs in *PP* (p.231):

in inferring this system from the evidence, we necessarily impose conditions of coherence, rationality, and consistency. These conditions have no echo in physical theory, which is why we can look for no more than rough correlations between psychological and physical phenomena.

Physical theory does not use notions of rationality or coherence. Further it is as certain as anything we know that these notions will not be reconstructed within physical theory. If strict psychophysical laws required that rationality has an echo in physical theory, then there could be no psychophysical laws.

I do not think that if there were strict psychophysical laws then there would have to be an echo in physical theory of the constraints of rationality we use in interpretation. Suppose there were a set of strict psychophysical laws for a particular person through a stretch of time. Suppose they took the form of conditionals whose antecedents specified conditions physically, and whose consequents specified the thoughts of the person. Using these laws and evidence about the physical facts we arrive at a description of the psychology of the agent. The description (together with various statements about the world) shows the agent to be rational, consistent, a believer of truths. Where the agent makes mistakes the description

shows why the mistakes are intelligible. The laws do not contain a statement of what is rational, nor do they use a statement of what is rational in generating the description of the agent's psychology. All they do is state physical conditions under which an agent has a set of propositional attitudes that make just the kind of sense we see him as making in interpreting him. In section 3 I noted that natural laws need not "transmit" synthetic *a priori* laws. This response to Davidson's claim is an application of that point. The physical realizations of mental states might support laws in certain circumstances, but the cases that block the "transmission" of the synthetic *a priori* laws can constitute a "nomic miscellany."

Even if there are laws governing all the exceptions, so that there is a predicate in physics that is *nominally* coextensive with a predicate in psychology the echo may be exceedingly faint. Consider the weakest possible kind of law: the antecedents of the laws contain state descriptions of the entire expanse of space within informational range of a particular propositional attitude change for a certain span of time. (If supervenience is taken to have modal force, then there are such laws. If citing these as a counterexample to the thesis of anomalism is the best we can do, I think we should concede that anomalism is correct.) These laws provide a physical predicate that tracks actual rationality, but it is as uninformative as it could possibly be. It does not provide a useable, or even a knowable, reconstruction of the conditions of rationality in physical terms.

John McDowell (McDowell, 1985) articulates an even stronger version of this claim. Loar argues (Loar, 1981, 20-25) that Davidson is wrong that the patterns required by rationality have no echo in physical theory. On the contrary: the syntactic patterns we capture in a formal theory of deductive validity are patterns we can describe in physical terms. McDowell responds as follows:

The Davidsonian claim, now, is that this structure... cannot be abstracted away from relations between contents, or forms of content, in such a way that we might hope to find the abstracted structure exemplified in the relations among a system of items described in non-intentional terms. And in this case the claim is actually susceptible of something like proof. Someone who denied the claim would find it hard to explain how his position was consistent with the fact that there is no mechanical test for logical validity in general.

I think McDowell overstates his case. Loar claims that if someone has a logically true thought, or makes a logically valid inference, then the pattern her thought makes has a physical description. This claim does not entail that the "physical

scheme” includes a theory of inference that contains a mechanical test for logical validity. Even if there is an echo of rationality in physics in the very weak sense I just sketched, this echo provides a theory of rationality in physics in the way a list of logically valid arguments provides a mechanical test of validity.¹⁰

I now describe an argument suggested by Davidson’s general stance on meaning and interpretation. The argument begins with a remark about our possible epistemic relations to meaning.

Meaning is public and available. We use language to communicate with one another, and what there is to meaning is what is available to the ordinary interpreter that will aid the process of communication. Undoubtedly there is much we do not know about language and communication. For instance, we do not know the mechanical process by which vibrations of the mastoid bone control speech production. But that process cannot be essential to meaning. We speak and hear and understand what there is to be understood in the utterances of others without knowing the first thing about this and other bodily processes.

Normally when one theory reduces another there is a net epistemic gain. The reduction of Mendelian genetics to molecular genetics taught us a great deal about genes. We were finally able, for instance, to describe a mechanism by which alterations of genetic material from generation to generation were possible.

If a reduction of psychology to physics via strict psychophysical bridge laws were possible, then we would stand to learn a great deal about meaning. But since we already know what there is to know about meaning and the content of propositional attitudes, there could not be such laws.

This argument provides a good reason to think there are no strict psychophysical laws. But it is not a compelling reason. Reduction normally provides this sort of epistemic gain. But it is not obvious that reduction through laws has to teach us anything beyond the laws. (Philosophers inclined to think there are strict psychophysical laws will also be inclined to deny that meaning is public and available

¹⁰Loar makes a point similar to the one I pursue in this section: “Nothing like the constraints of rationality shows up in physical *theory*; but counterfactual constraints that are isomorphic to rationality constraints may show up in the physical *facts*.” (p.23) Loar holds that the possibility of a functionalist reduction shows that anomalism is incorrect. If by “functionalist reduction” we mean nomological reduction, and we know that such a reduction is possible, then Loar is correct. I do not think we know whether such a reduction is possible; my argument has been that we do not know *a priori* that it is not possible. Loar is clearly wrong if “functionalist reduction” means that we can describe the causal relations among “inputs,” “outputs,” and other functional states without using notions like rationality, and be confident that the result describes a psychology.

in the sense described; I think this inclination should be resisted.)

The difficulty is the same as the difficulty with Davidson's "no echo" claim and McDowell's much stronger version of it. We might know that a set of laws is a set of strict psychophysical laws, yet fail to know a variety of other things about them. We could well fail to know anything about the antecedents of these laws that rationally compelled acceptance of their status as strict psychophysical laws, because there simply isn't anything about them that would do this. A set of strict psychophysical laws might correctly predict the propositional attitudes of an agent but not entail anything about the nature or structure of meaning.

This argument does provide a good reason to think there are no strict psychophysical laws if we think that psychophysical reduction requires a net epistemic gain. In this respect what shows this argument invalid resembles the claim that nomic supervenience shows that laws are possible: since the situation that would show that anomalism is false is so completely unlike what we might have had in mind in thinking about psychophysical laws and reduction, we should concede that no laws of any philosophically interesting sort are possible.

7.6 Conclusions

Davidson's argument for monism relies on a claim about causation and a claim about psychophysical laws. Causation requires strict laws; there are some strict laws, presumably those of physical theory, but there can be no strict psychophysical laws; hence when a mental event causes a physical event the mental event must instantiate a strict law of physical theory, hence the mental event is a physical event.

I have concentrated on the argument that there can be no strict psychophysical laws. Davidson holds that the strong conceptual connection between the concepts of psychological theory and the concepts of rationality demands that there could be no strong nomic connection. I think the fundamental difficulty is that unless the concepts explicitly preclude the possibility of laws it is up to the world whether there are laws.

I do not think this difficulty shows that Davidson is wrong in what he claims. He notes that he cannot demonstrate that there are no strict psychophysical laws:

In fact, however, nothing I can say about the irreducibility of the mental deserves to be called a proof (ME, p.215)

Davidson maintains that lawlikeness is “much like analyticity” and that there are no nonquestionbegging criteria of when a sentence is a law (or when a sentence is analytic), that we might use to generate a proof. The best we can do is inspect the commitments of the theories from which the terms derive, and arrive at a judgment of the likelihood that laws are possible.

For all I have said we still can be reasonably confident that there are no strict psychophysical laws. I described two more or less concrete situations that showed that strict psychophysical laws might be possible. One supposes the laws entail oracular deliverances, like complete ascriptions of propositional attitudes that turn out to be just as sensible as the ones we would confidently make if we possessed complete information about the person. The other derives the possibility of laws from supervenience. If this is the best we can hope for we might as well concede there are no strict psychophysical laws in any sense worth talking about.

We have found a number of differences between psychophysical generalizations and others. Content ascribing predicates have an analytic connection with value ascribing predicates; in the case of propositional attitude psychology, content ascribing predicates have an analytic connection with rationality ascribing predicates. These differences mark important differences in the kinds of explanations that psychology and the nonintentional sciences offer. They do not, of course, show that psychology is not a science. Many have taken Davidson’s claims as an attack on psychology. The only explanation for this mistake must be that they have overlooked the fact that Davidson does not object to the possibility of psychophysical laws.¹¹

¹¹Thanks to Jerry Fodor and Bruce Vermazen for comments on an earlier version of this chapter.

Chapter 8

Causal Relevance

We might put the conclusion of the last chapter as follows: anomalism is not necessarily true, but it is overwhelmingly likely to be actually true, and anomalous monism provides a simple and elegant model for the way psychological states are related to physical states.

Many philosophers have felt that anomalous monism does not do justice to our conviction that what we think makes things happen. There are two versions of this argument. One holds that if Davidson is right then only certain aspects of mental events are causally relevant. The causally relevant aspects are the ones captured in strict laws. Hence mental aspects of mental events are not causally relevant.

The second version concentrates on semantic externalism, a view held by Davidson. According to externalism, semantic properties of things are relational properties. If something that has the same internal description as something with semantic properties had had relations to different things external to it, its semantic properties would have been different. Yet two things with the same internal properties have the same causal powers. Hence, it is held, the semantic properties do not contribute to the causal powers of the thing.

I believe both these objections to anomalous monism are completely misguided. Causal relevance need not be connected with strict law in the way suggested. And relational properties of things can be causally relevant.

I begin this Chapter with a description of some aspects of things that are *not* causally relevant. This extended argument lays the metaphysical groundwork for the responses I will make to these two objections. In Part II I consider a couple of formulations of the argument about causal relevance and strict laws. A careful ex-

position of the notion of supervenience shows why the more subtle version fails. In Part III I then consider three formulations of the argument about externalism. The last involves probably the most articulate conception of causal relevance in the recent literature, Nancy Cartwright's probabilistic account of general causal claims. I show that there are some simple structural reasons why relational properties can be counted as causally relevant within Cartwright's theory.

8.1 Causal Relevance and Conceptual Connections

8.1.1 Introduction

Singular causal claims are statements of the form '*a* caused *b*,' where '*a*' and '*b*' are names or descriptions of particular events. A true singular causal claim may fail to provide a causal explanation: for instance, the event that occurred 6 feet due north of me last Tuesday caused the event that worried Sue on Thursday. Causal explanation requires a certain kind of description of events. What kind of description is it?

Part of the answer is that the right descriptions pick out causally relevant aspects of events.¹ Sandor died last week; his death was caused by his eating sole with sauce béarnaise for the first time in his life. The causal explanation for his death is that the sauce was tainted with salmonella. The causally relevant aspect of the cause of his death is the fact that the sauce was tainted with salmonella.

There are several ways to understand the causal relevance relation. Recently much interest has been focussed on probabilistic accounts of general causal claims (like "aspirin relieves headache"); other accounts connect causal relevance with counterfactuals or laws.² All accounts need some sort of independence condition on causal relevance. Otherwise, it seems, we might have to count drawing an ace of spades as causally relevant to drawing a spade, since the right probabilistic or counterfactual or nomic relations are satisfied.³

¹John R. Searle holds that the explanatory power of a singular causal claim depends on whether it describes events under causally relevant aspects; see his *Intentionality* (Searle, 1983, 116-7).

²For probabilistic accounts see Patrick Suppes, *A Probabilistic Theory of Causation* (Suppes, 1970) and Nancy Cartwright, "Causal Laws and Effective Strategies" (Cartwright, 1983). For a probabilistic account involving counterfactuals, see Ronald Giere, "Causal Systems and Statistical Hypotheses" (Giere, 1980). For an account connecting causal relevance with counterfactuals and natural law, see Ernest Lepore and Barry Loewer, "Mind Matters" (Lepore and Loewer, 1987).

³Wayne Davis discusses the need for an independence condition in a probabilistic account of

What sort of independence condition should it be? Complete logical independence is both too strong and too weak. So long as the events are distinct drawing an ace of spades might well cause drawing a spade. Kim objects to Lewis' counterfactual account of causation with several non-causal non-logical dependencies: "I opened the window" is logically independent of "I turned the knob," but my turning the knob does not cause my opening the window, since I open the window by turning the knob.⁴ LePore and Loewer formulate a notion of metaphysical independence that comes closer but which also rules out both too much and too little.⁵

I will concentrate here on one aspect of the problem of specifying an adequate set of independence conditions: cases where one property logically entails that its instances are caused by instances of another, specified, property. (Sunburn, for instance, must be caused by exposure to the sun.) I describe five ways to formulate a condition that rules out pairs of properties like this as causally relevant, and accept three. The conditions I accept mark a distinction in kinds of causal explanations: some causal explanations cite causally relevant properties, while others, for instance causal explanations of sunburn, do not. (Despite this distinction I think these formulations of the independence condition are part of the analysis of an ordinary notion familiar from ordinary contexts of causal explanation.)

This account of causal relevance sheds some light on the question whether reasons are causes. Philosophers have argued that the conceptual connections between reasons and actions show that reasons do not cause actions. I think the conceptual connections do show that being a reason is not causally relevant to being an action; but they do not show that reasons do not cause actions.

8.1.2 Metaphysical Background

In this section I state some assumptions that make up the conception of events and causation with which I will be working, and give a more precise statement of the thesis I'll be arguing for in later sections.

Events are particulars; they have places and times and cannot be repeated. Several events of the same type may occur. Each event is of countless types; some types are interesting for explanatory purposes, many are not. The types to

causation, in "Probabilistic Theories of Causation" (Davis, 1988).

⁴Jaegwon Kim, "Causes and Counterfactuals" (Kim, 1973).

⁵LePore and Loewer, *op.cit.*, footnote 13, p.635. See below, notes 13 and 18, for details on why their definition isn't quite right.

which events belong, and the nature of events, are closely connected with changes in the properties of objects. For instance the event of Sandor eating sole is a change in Sandor: first he was not eating, then he was eating. Types and properties correspond exactly: an event is of a certain type if and only if it has a certain property.

Properties of events correspond to descriptions and their meanings. On this conception of properties, the property of turning red is distinct from turning the color of the back cover of “Twelve Picturesque Passions,” even though the back cover of that work is red. Properties can be described in several ways as well, however; there must, therefore, be some way to decide when two different phrases express the same property. Describing how this is done is, of course, a difficult task.

We have relatively few terms that clearly denote properties of events only. We know how to generate such terms, since any event is a change in what properties are instantiated by an object. Since it is cumbersome to denote properties of events in this way I will speak loosely of properties of objects as a shorthand for the events of their coming to have those properties.

I will be discussing three quite different relations involving causation. *Causal relations* are relations between particular events. Causal relations are often reported with singular causal claims. One distinctive feature of these reports is that they are extensional. (“Causation” is another name for this relation.)

Causal relevance is a relation between properties.⁶ A report of causal relevance is also extensional: “taking aspirin is causally relevant to the waning of headaches”, if true, remains true regardless of the way the properties are described. (Clearly causal relevance is by definition not a causal relation. We call it ‘causal’ since many, if not all, instances of the causal relation involve causally relevant properties.) A pair of properties is in the causal relevance relation only if there is some explanatory link between them, such as the ones given by the probabilistic, counterfactual, or nomic accounts of general causal claims.

Causal explanation is a relation among two causally related particulars and two properties. Sentences reporting causal explanations are intensional, since changing the descriptions used to pick out the particular events involved may alter which property is picked out. If “taking aspirin causally explained the waning of the headache” is true, most likely “swallowing the unique small white thing in the

⁶This is something of a terminological stipulation, since causal relevance is sometimes taken to be a relation between particular events, regardless of their properties.

house causally explained the waning of the headache” is not.

The term ‘causes’ in causal claims is ambiguous: it can express either causal relations, causal relevance relations, or causal explanations. Consider a singular causal claim: “the short circuit caused the fire.” This sentence may either simply report a causal relation between two particulars, hence be extensional, or it may give a causal explanation, the truth of which depends on getting the descriptions right. General causal claims (like “smoking causes cancer” or “explosions cause bridge collapses”) exhibit the same ambiguity, although the extensional reading is somewhat more forced. General causal claims can in addition report a causal relevance relation between properties.

The claim I defend below is this. One condition on causal relevance is a certain kind of independence, defined in terms of the causal relations instances of the properties can have. Causal explanation usually involves citing causally relevant properties, but many (quite ordinary) causal explanations cite properties that are not independent in the required way, and hence are not in the causal relevance relation.

8.1.3 Notation and Terminology

I will state the condition on causal relevance in a variant on standard notation for the predicate calculus. ‘*x*’, ‘*y*’, ‘*z*’ are variables ranging over events. ‘(Ex)’ and ‘(x)’ are quantifiers over events. ‘*a*’, ‘*b*’, ‘*c*’, ‘*d*’, ‘*e*’, and ‘*f*’ are names of events. I’ll also need names and variables and quantifiers for properties. ‘*p*’ followed by a numeral is a name for a property; ‘*p*’ followed by a capital letter is a property variable. ‘(EpX)’ and ‘(pX)’ are property quantifiers.

Since I am both talking about properties and asserting that events have them, I need one special relation between properties and events. ‘*a* has *p*1’ means the event *a* has the property *p*1; ‘(EpX)(x) x has pX’ means there is a property such that every event has that property. I will sometimes use different locutions to express the same relation: e.g., ‘*e* is *p*1’ or ‘*e* is an instance of *p*1’ or ‘*p*1’s instances.’

Causal relevance is one relation between properties; ‘CR(*p*1,*p*2)’ means the property *p*1 is causally relevant to the property *p*2. Metaphysical independence is another; it is the relation I will be describing. ‘mi(*p*1,*p*2)’ means the property *p*1 is metaphysically independent of the property *p*2. I’ll also use ‘wmi’, ‘smi’ and ‘mmi’ for weak, strong, and middling metaphysical independence, respectively.

8.1.4 A Weak Condition

In this section I state a weak independence condition on causal relevance, then give some examples to show where it is and where it is not satisfied.

If one property is causally relevant to another then the second is independent of the first:

$$(pX)(pY) \text{ if } CR(pX,pY) \text{ then } mi(pY,pX)$$

In words, for all pairs of properties pX and pY , if pX is causally relevant to pY then pY is metaphysically independent of pX .

Here is the weakest version of mi I'll be defining:

$$(pX)(pY) \text{ wmi}(pY,pX) \text{ iff}$$

possibly $(\exists x)[x \text{ has } pY \text{ and}$

$\sim(\exists y)(y \text{ has } pX \text{ and } y \text{ caused } x)]$

In words, for all pairs of properties pX and pY , pY is weakly metaphysically independent of pX if and only if it is possible for there to be an event that has pY that is not caused by an event that has pX .⁷ (Hence causal relevance is partly defined in terms of causal relations.) There are two ways it is possible for an event to have pY and not be caused by an event that has pX : either it is caused by no event that has pX , or it is not caused at all. The second way will be important to the acceptability of the stronger conditions: miracles, i.e., uncaused events, are possible. The conceptual grounds offered in support of the claim that every event has a cause are strong but not, I think, conclusive.

By 'possibly' I mean logically or conceptually possible. I will have in mind any connection between concepts such that denying that the connection holds presents an unintelligible possibility. I do not assume that anything clear can be said about what is and what is not conceptually possible. I do assume that there are some circumstances that are conceptually impossible.

I now describe three cases where causal relevance is absent because the properties involved are not weakly metaphysically independent.

⁷Notice the temporal asymmetry in the notation for these relations: if $CR(p1,p2)$ then events that have $p1$ cause, and hence occur prior to, events that have $p2$. If $\sim mi(p1,p2)$ then events that have $p1$ must be caused by, and hence occur after, events that have $p2$.

i. Suppose being color-caused is a property things have in virtue of being caused to have their color by something that is colored (rather than transparent and colorless) and the color of the cause explains the color of the effect. Developing a print from 35-mm color film exposed in the usual way is an event of something coming to be colored and color-caused. Developing a print from 35-mm infrared color film exposed to radiation emitted by transparent colorless jellyfish is an event of something coming to be colored but not color-caused.

A change in the color of an object is weakly metaphysically independent of the color of its cause. An event of an object changing color could occur uncaused (a miracle) or caused by something transparent (a jellyfish).

The property of being color-caused is not metaphysically independent of the cause being colored. Hence the fact that the cause is colored is not causally relevant to an object's becoming color-caused. What is relevant to the object's becoming color-caused is that the demands set by the concept of being color-caused are met; we might call this logical relevance rather than causal relevance. (A *causal explanation* of an event of something becoming color-caused may, however, cite the color of the cause.)

ii. Exposure to the sun burns the skin. Sunburn is an inflammation of the skin caused by exposure to the sun. Consider the event of my coming to be sunburned. The sun shining on my skin is its cause. The sun's shining on my skin is causally relevant to the burn. It is not causally relevant to the sunburn, since the property of coming to be sunburned is not weakly metaphysically independent of the property of being exposed to sunlight. (I don't think it is possible to get sunburned by a sunlamp. If you do, consider snowblink.) The sun's shining on my skin may, however, be cited in a causal explanation of my sunburn, although this would be a pretty thin explanation.

iii. The minting of a US dime is not metaphysically independent of certain properties of its cause: the press must be operated according to the regulations that govern the minting of coins. Hence those properties of the cause of the minting of a dime are not causally relevant to the property of coming to be a US dime. They are, however, legally relevant. Legal relevance is a sort of conditional logical relevance: given the laws as they are, something is a dime only if it was produced in a certain way.

8.1.5 Justifying the Condition

These examples manifest some distinction, but it may be thought that the crucial difference has nothing to do with causal relevance. The objection might run as follows: conceptual connections like these merely select from the range of causally relevant features, and the features thus selected remain causally relevant to the causally committed property. Many things cause, and are causally relevant to, the kind of burn that sunburn is; sunburn is different from that kind of burn only in being more particular about its causes. Exposure to the sun then is causally relevant both to sunburn and to the kind of burn that sunburn is.

What follow are five strategies for justifying the weak condition. I think all fail in the same way: they beg the question at issue. Despite this problem I think causal relevance should be contingent, and so I accept wmi, even though I cannot see how to persuade someone who thinks otherwise.

i. There is a sort of conceptual parsimony that recommends wmi. All the examples of causally committed properties are examples of properties whose instances have *other* properties that stand in fully contingent relations. Sunburn is a kind of burn that is produced by exposure to the sun, and there is a contingent explanatory relation between exposure to the sun and this kind of burn. This fact about all the examples might persuade us that exposure to the sun is causally relevant to these other properties and not to sunburn. The trouble is that we could as well arrange our definitions the other way around: there are two kinds of causal relevance, the kind that does, and the kind that does not, satisfy wmi.

ii. Perhaps epistemic considerations justify the condition. Causal relations, it might be said, are discovered by observing contingent connections. This is an essential aspect of causal relations. Then causally relevant properties must be ones that can only be discovered by discovering contingent connections.

There are two difficulties. First, it may be that causal relations must be contingent in some way, but from that it does not follow that one property is not causally relevant to another if the properties are not metaphysically independent.

Second, we can and do discover causal relations through conceptual connections. I read in the newspaper that my congressperson voted 'no' on AB2020. Knowing this I know a great deal about the causes of the event. It may be objected that this knowledge is justified by knowledge of various facts about newspapers, reporters and agents, and that somewhere in that justification there must be a premise about contingent relations of the events surrounding the vote. Whether or not this is so it does not entail that the conceptually connected properties are

not in the causal relevance relation.

iii. One very plausible condition on explanation is that explanation requires information.⁸ Suppose I know of an event e under one of its aspects p_2 , and I wish to explain it. Mentioning a property p_1 of its cause such that the sentence ' e has p_2 ' entails 'the cause of e has p_1 ' does not explain the event, since from the entailment it is already known that there is something around that is p_1 that caused e .

This requirement on explanation does not justify wmi , since it does not show why causal relevance requires information. Causal explanation and causal relevance are different things, and so it is unclear why a requirement on one must be a requirement on the other.

iv. Perhaps there is some connection between explanation and causation that justifies the condition. One necessary condition for the truth of the statement that c caused e might be that c and e have properties p_X and p_Y such that mentioning that c has p_X explains the fact that e has p_Y .⁹ If $\sim wmi(p_Y, p_X)$ and explanation requires information then p_X and p_Y do not meet this requirement. Then if an event has p_Y there must be something else about that event which, together with some aspect of its cause (p_X will do), does meet the requirement. Finally we assert that only properties that meet this requirement are causally relevant.

Another way to articulate this strategy is to claim that whenever a singular causal claim is true there is something about the events that *accounts for* this fact. c 's having p_X and e 's having p_Y *account for* the causal relation between c and e only if c 's having p_X explains e 's having p_Y . Since if $\sim wmi(p_Y, p_X)$ this condition is not satisfied, p_X and p_Y do not account for the causal relation. Tying causal relevance to this idea of accounting for a causal relation, we get the desired conclusion that if $\sim wmi(p_Y, p_X)$ then p_X is not causally relevant to p_Y .¹⁰

Again there are two difficulties. First, it is unclear why causation requires

⁸Davidson's "Actions, Reasons and Causes" (Davidson, 1963, 14-15) appeals to this condition on explanation. Some accounts of explanation do not include this condition; for instance we can distinguish what information a statement actually carries and what information it carries for some particular inquirer.

⁹I phrase this condition in terms of explanation (a rather broad and vague notion) because many more plausible conditions, for instance ones that require instantiation of contingently related properties, would be question-begging. The condition I give doesn't beg the question and licenses the use of a condition on explanation I've accepted (which apparently does not beg the question) in an argument to the desired conclusion.

¹⁰For the idea that causally relevant properties are ones that account for causal relations, see Mark Johnston, "Why Having a Mind Matters" (Johnston, 1985, 423).

anything of events other than that one causes the other. Davidson's claim that whenever two events are related as cause and effect they have descriptions that instantiate a strict law is one strong version of this thesis. Anscombe has persuaded many that Davidson's claim is not conceptually necessary.¹¹ The claim that the causal relation must be *accounted for* is even stronger than Davidson's; why they should be thus *accounted for* is completely obscure.

Second, even if we accept these conditions on causal relations it is still unclear why they should rule that only wmi properties are candidates for causal relevance; all they would say, for instance, is that among the causally relevant properties are ones that *account for* the causal relation.

v. Finally, one may feel, as I do, a reluctance to think that nature contains any logical or conceptual connections. Not accepting wmi leaves open the possibility of a case in which the *only* explanatory relation between properties of two causally related events is one involving causally committed properties. Such a case appears to be one in which one event occurs because it is logically required by another. Again I doubt this possibility forces us to accept wmi. For one thing, we could insist, consistently with rejecting wmi, that causally committed properties are always accompanied by causally uncommitted properties. For another, we could simply accept the possibility, despite its oddness. I do not claim to understand it but I also do not see anything incoherent about it.

Despite the difficulties in the way of justifying wmi I accept it; I also accept some of the ideas that were offered on its behalf, construed now as guides to what properties are causally relevant. The parsimony consideration is specially important. Whenever some property pX requires that some other property pY stands in a contingent explanatory relation with another aspect pZ of pX's instances, I count pY and pZ as the causally relevant pair of properties.

8.1.6 Details and Elaboration

In this section I give some of the logical properties of the CR and wmi relations. I'll also draw a couple of the consequences of the definitions and sketch one natural extension.

¹¹Anscombe, G.E.M., "Causality and Determination" (Anscombe, 1971). Dropping any such requirement on causal relations would leave a very puzzling account. It would not, however, require admitting the possibility that one event caused another but there is no explanatory relation between them at all. The causal relation remains an explanatory relation even if there is no further explanatory relation between properties of the events.

wmi is not a symmetrical relation: possibly $\sim\text{wmi}(p1,p2)$ and $\text{wmi}(p2,p1)$. For instance sunburn is not wmi of exposure to sunlight, but exposure to sunlight is wmi of sunburn. wmi is not reflexive, since there can be properties whose instances require a cause that is another instance of that property. If nothing is a horse unless its cause is another horse, then $\sim\text{wmi}(\text{coming to be a horse}, \text{coming to be a horse})$.¹² wmi is not a transitive relation. Suppose sunburn can be caused both by direct exposure to the sun and by light reflected by mirrors, windows, etc., provided it comes ultimately from the sun. Sunburn is wmi of light reflected by mirrors, windows, etc., and light reflecting from those sources is wmi of exposure to sunlight; yet sunburn is not wmi of being exposed to sunlight.

CR is not symmetrical or reflexive because the explanatory link is not symmetrical or reflexive. Symmetry first: exposure to the sun is causally relevant to a burn, but burns like this do not typically or regularly lead to exposure to the sun. Reflexivity: getting a burn like this does not typically or regularly lead directly to another burn. The trouble for transitivity comes both from the independence condition and the condition that the link should be explanatory. The same example shows how the independence condition interferes with transitivity. Light emitted from the sun can be causally relevant to light reflecting from a mirror, and light reflecting from a mirror might in turn be causally relevant to someone getting burned, but the sunlight is not causally relevant to the sunburn. Explanatory relations are not generally transitive, nor are the various relations that have been proposed to explicate causal relevance.

Property entailment has various consequences for wmi and CR. (A property pX entails a property pY just in case necessarily, $(x)(x \text{ has } pX \rightarrow x \text{ has } pY)$.) Suppose $\sim\text{wmi}(p2,p1)$. Then for any property pX such that p1 entails pX, $\sim\text{wmi}(p2,pX)$. This follows from the fact that if p2 requires its cause to be p1 equally it requires its cause to be anything that p1 requires. More generally, *lack of wmi is transitive*: if $\sim\text{wmi}(p3,p2)$ and $\sim\text{wmi}(p2,p1)$ then $\sim\text{wmi}(p3,p1)$.

I turn now to a possible complaint about the way wmi is defined. wmi and CR are relations among properties, and hence fully general. But the reference to causal relations in the definition of wmi introduces an element of particularity. Couldn't we dispense with it to get a stronger notion of metaphysical independence, as follows:

¹²Hence becoming a horse is not causally relevant to anything becoming a horse. This may be thought to be a *reductio ad absurdum* of the weak condition, since it entails that no species property is causally relevant to itself. But surely there is much else about a foal to which being a horse may be causally relevant.

$(pX)(pY) \text{ wmi}'(pY,pX) \text{ iff}$

possibly $(\text{Ex})(x \text{ has } pY \text{ and } \sim(\text{Ey}) y \text{ has } pX)$

My reason for preferring *wmi* to *wmi'* is that *wmi'* seems too strong. It makes it impossible for one property to be causally relevant to another if the second requires only that somewhere in space and time there is an instance of the first. Suppose that writing a check is an event that requires that banks have opened, but it is not necessary that an event of writing a check must be caused by a bank's opening. I should think that being a bank opening could be causally relevant to a check's being written.¹³

Notice that *wmi* is quite strong as it stands. The definition entails, for instance, that if being an event of a check's being written did require a cause that is an opening of a bank, then being an opening of a bank is never causally relevant to any check's being written, even if it is an aspect of some event other than the cause of the check's being written.¹⁴

¹³LePore and Loewer's notion of metaphysical independence is defined in terms of particular events:

c's being *F* and *e*'s being *G* are metaphysically independent, iff there is a possible world in which *c* (or a counterpart of *c*) is *F* but *e* (or a counterpart of *e*) fails to occur or fails to be *G* and vice versa. (op.cit., p.635, note 13)

Clearly this definition doesn't entail that a particular bank opening is not causally relevant to a particular check's being written. It is still too strong in the following respect. Suppose the definition is not satisfied because there is no world at which *e* is *G* where *c* does not occur or fails to be *F*, but among the worlds at which *e* is *G* there are some where *c* doesn't cause *e*, and furthermore some of these worlds are ones at which no cause of *e* is *F*. (I omit the qualification 'or a counterpart' for clarity.) Then it is possible for *e* to be caused to be *G* by something which is not *F*, even though *e*'s being *G* (somehow) requires that *c* being *F* must be on the scene somewhere. *wmi* permits *c*'s being *F* to be causally relevant to *e*'s being *G* in this case, while LePore and Loewer's definition does not.

¹⁴It is possible for an aspect of one event to be causally relevant to an aspect of another even where we would not count the first as the cause of the second. A good explanation of some event might cite one event as its cause, and an aspect of another as the causally relevant factor. The cause might be a necessary enabling factor for the operation of some other mechanism. For instance, the cause of the crash of the airliner was the failure of the controller to command the de-icing of the wings. The causally relevant factor was the thickness of the ice; the change in the shape of the airfoil reduced its lift.

Causally committed properties are not committed as to whether there are objective facts about the unique causes of events. If some event is an event of getting sunburned then among the events that cause it at least one must be an exposure to the sun.

My definition of *wmi* only mentions requirements on properties of causes, but there is a natural extension to cases where the requirement is on properties of effects. Suppose I pull the trigger of a pistol in a competition; this bodily movement is an action of hitting the bullseye only if what I do causes the bullet to hit the bullseye. My hitting the bullseye is surely not causally relevant to the bullet's hitting the bullseye.¹⁵

8.1.7 A Stronger Condition

Suppose $\sim wmi(p2,p1)$. Accepting *wmi* as a condition on causal relevance means denying that *p1* is causally relevant to *p2*. In the last section I showed that properties that *p1* entails are also not causally relevant to *p2*. Are there further acceptable independence conditions that rule other properties as not causally relevant to *p2*? In this section I consider properties that entail *p1*. In the next section I'll consider the rest.

Consider again the property of coming to be color-caused. Consider some particular event *e* that has this property; we know that its cause *c* must be an event that involves something that is colored. *c*'s involving something colored is not causally relevant to *e*'s involving a color-caused object. Suppose the color is blue. Is the blueness causally relevant to *e*'s involving a color-caused object?

Explanation requires information. Citing the fact that its cause involves a colored object gives no information about an event involving a color-caused object, but citing the fact that the cause involves a blue object does give information. Is this extra informational content enough for causal relevance?

I think the fact that *c* involves a blue object is no more causally relevant to *e*'s involving a color-caused object than the fact that *c* involves a colored object. The property of being color-caused requires that its instances are colored and that their coloration is explained by the coloration of their causes. What *makes* something color-caused is the fact that this requirement is satisfied; the requirement specifies something that is logically, not causally, relevant to being color-caused. Being blue satisfies the requirement by way of entailing being colored.

This is the point at which causal explanation and causal relevance definitely

¹⁵Note that if we think of causal relevance as a relation between particular events (rather than a relation between properties), and we think that the phrase "my hitting the bullseye" denotes a bodily movement of mine, then my hitting the bullseye undoubtedly is causally relevant to the bullet's hitting the bullseye.

part ways. Earlier I claimed that exposure to the sun might causally explain sunburn, although exposure to the sun is not causally relevant to sunburn. This claim would be incorrect if we hold that explanation requires information; then exposure to the sun is not causally relevant to sunburn and does not causally explain it. But we now have cases where the cause property clearly does provide information not available given the effect property, but where the effect property is not metaphysically independent of the cause property. Certainly a causal explanation of an event of something becoming color-caused can refer to something's acquiring a particular color, even though the cause property is not causally relevant to the effect property. Citing an exposure to the sun that spanned four hours gives information about a case of sunburn, specifies something causally relevant to the burn (the length of the exposure), and can causally explain the sunburn, but it does not specify anything causally relevant to sunburn. (Causal explanations that do not pick out causally relevant properties do, however, usually *involve* causally relevant properties, in the sense in which an explanation of a case of sunburn *involves* a burn of a certain kind.)

The definition of middling metaphysical independence, *mmi*, reads like this:

$$(pX)(pY) \text{ mmi}(pY,pX) \text{ iff}$$

$$(pZ) \text{ if } pX \text{ entails } pZ \text{ then } \text{wmi}(pY,pZ)$$

In words, a property *pY* is middling metaphysically independent of another property *pX* if and only if *pY* is weakly metaphysically independent of every property *pZ* that *pX* entails; or, equivalently, there is no property that *pX* entails of which *pY* is not weakly metaphysically independent.¹⁶

The causal account of acting on a reason holds that reason explanations of action are causal explanations. Accepting *mmi* as a condition on causal relevance

¹⁶Another *reductio* threatens. Any property entails many things; among these, that its instances have properties and are self-identical. Any property that requires its instances have a cause requires that its instances are caused by things that have properties and are self-identical. So any property that requires that its instances have a cause is not *mmi* of any property. There are two responses. First, some properties of events do not demand that their instances have causes; since I think miracles are possible I think there is no general ground on which to deny this. Second, perhaps there is some way to segregate “categorical” properties like self-identity from “empirical” properties like acquiring a certain temperature; then *mmi* would require a property to be *wmi* of every “empirical” property the candidate property requires. (I owe these responses to Kirk Ludwig.) The second response promises to be difficult to formulate, but has the advantage that it is silent on the possibility of miracles; their impossibility would not render the notion of causal relevance absurd.

entails that being a reason is not, however, causally relevant to being an action. I'll use a sketch of Davidson's causal account of action to develop this point (Davidson, 1980b, essays 1-6).

A bodily movement is an action only if it has an aspect that makes true a sentence that says it was something done with an intention. An intention is an unconditional judgment that actions of a certain kind are preferable to all others. For an action to be performed with an intention is for it to be caused by an intention. In turn an intention must be caused by a reason: a pair comprising a *prima facie* valuation of actions of a certain kind and an instrumental belief that by means of a certain bodily movement an action of the valued kind will be performed.

Being an action is not wmi of coming to have an intention, so being an intention is not causally relevant to being an action. wmi also bars coming to have a reason from being causally relevant to coming to have an intention. Since being an intention entails causation by a reason, coming to have a reason isn't causally relevant to being an action.

Being an instrumental belief entails being a belief, so \sim wmi(being an action, being a belief); being a special kind of valuation of actions entails being a valuation, so \sim wmi(being an action, being a valuation). Beliefs, intentions and valuations are the only kinds of propositional attitudes, so being a propositional attitude is not causally relevant to being an action. mmi then bars being any particular kind of propositional attitude from causal relevance to being an action. These claims are all consistent with the claims that reasons cause actions, that reasons cause intentions, that intentions cause actions, and that reason explanation is a species of causal explanation.¹⁷

8.1.8 The Strongest Condition

Many properties remain as candidates for causal relevance to causally committed properties. Various properties of events which cause events that involve color-caused objects are causally relevant to the color of those objects. The intensity of the light affects the color of a print, and a particular intensity doesn't entail any color: it might be UV light. Could the intensity of the light be causally relevant to the fact that the print is color-caused? Or, returning to the theory of action, could

¹⁷LePore and Loewer's definition of metaphysical independence (see note 13 above) is in this respect too weak: it permits being a particular propositional attitude to be causally relevant to being a particular action, since that action could have occurred without the reason that actually caused it.

some other aspect of our mental lives be causally relevant to being an action, perhaps something to do with consciousness, or brain properties?

I think again the answer is no. The reason is just a more general application of the justification for mmi. Suppose the candidate properties are p2, which requires causes of its instances to have p1; the conjunctive property (p1 and p3); and some other property p4 of which p2 is both wmi and mmi. (p1 and p3) is not causally relevant to p2 because it includes what is conceptually required by p1. Instances of (p1 and p3) are “richer” than (some) instances of p1, in respect of having p3 as well; but what is relevant to something’s being p2 is only that its cause is p1. p3 doesn’t matter to something’s being p2. But the same reasoning should go for p4 as well: what does p4 contribute to an effect’s being p2 that isn’t already contributed by p1? What could it contribute if p1 were absent?

I think that if some property is causally committed in the way specified by wmi, then no property is causally relevant to it. Here is the definition of strong metaphysical independence:

$$(pX)(pY) (\text{smi}(pY,pX) \text{ iff } (pZ) \text{wmi}(pY,pZ))$$

In words, a property pY is strongly metaphysically independent of another property pX (any other, in fact) if and only if pY is weakly metaphysically independent of every property; or, equivalently, if and only if there is no property of which pY is not metaphysically independent.

As with wmi there are a couple of natural extensions to smi that involve effects rather than causes. Consider first the case of a property that demands something about the effects of its instances: an event is an act of hitting the bullseye only if it causes a bullseye to be hit. It might be said that whatever *else* is causally relevant to the bullet’s hitting the bullseye (e.g., the gust of wind) is also causally relevant to my hitting the bullseye. No doubt we would not want to say this about aspects of events that occur in time between what I do and the bullet hitting the bullseye, so the interesting question concerns aspects of events that occur before what I do occurs.

Again, I think these aspects are not causally relevant to my hitting the bullseye. What is relevant to my hitting the bullseye is that my act causes a bullseye to be hit. This is logical relevance rather than causal relevance. Hence I think nothing is causally relevant to a property that is causally committed in the future direction.

Are causally committed properties themselves causally relevant to anything? Sunburn is not wmi of exposure to the sun. Sunburn, like other burns, raises the

temperature of the skin, and hence heats the air a bit more than usual. The event of the air being heated to, say, 99°F is wmi of sunburn. Is the sunburn causally relevant to this rise in temperature?

I think it is not. (Notice the question is not, does the sunburn causally explain the rise in temperature; the answer to that question is clearly, yes.) Certainly the burned condition of my skin is causally relevant to the heating of the air. If my sunburn heats the air, it is because sunburn is a kind of burn, and that burn is causally relevant to the heating of the air.

We can imagine cases in which what matters to an aspect of some event is the fact that it is caused by sunburn rather than some otherwise indistinguishable burn. You, for instance, might advise me to stay out of the sun. But even here I think we should not say that sunburn is causally relevant to the advice. Consider trying to build a device that is sensitive to sunburn rather than similar burns that are not caused by the sun. It must, I think, work by way of detecting the presence of a burn and by detecting features of the etiology of the burn. By detecting both the device would detect instances of a burn caused by exposure to the sun, that is, instances of sunburn; but to detect sunburn reliably it must detect these other two properties. I conclude that wmi bars all causal relevance in the future direction as well:

$$(pY)(\exists pX) \sim wmi(pY,pX) \rightarrow (pZ) \sim CR(pY,pZ)$$

Is *any* logical commitment fatal to causal relevance? (If it were we might have reason to doubt the coherence of the notion of causal relevance.) One thing common to wmi, mmi, and smi is that metaphysical independence is independence of certain causal relations. If logical relevance were all that mattered in these definitions, we should be able to eliminate the restriction to those causal relations, and formulate a non-causal sort of independence. Many ordinary properties are relational, where the relations are not causal. Being a planet is being a ball of rock that circles a star; something can be a planet that wasn't caused to be a planet by a star. Complete metaphysical independence, cmi, would be a condition on causal relevance that barred such relational properties from any causal relevance:

$$(pX)(pY) [cmi(pY,pX) \text{ iff} \\ \text{possibly } (\exists x)(pZ) (x \text{ has } pY \text{ and } \sim(\exists y) y \text{ has } pZ)]$$

Notice that cmi generalizes from wmi' the same way smi generalizes from wmi: once the property that is required is barred from causal relevance, we bar everything else as well. The problem with cmi is the same as the problem with wmi':

it is too strong. We cite relational properties in causal explanations all the time: being a planet, being a gene, being a citizen. *cmi* does pick out a notion of causal relevance, but it is not an ordinary notion familiar from ordinary contexts of explanation.¹⁸

8.1.9 Conclusions

Causal relevance is a relation between properties; the relation is explanatory in something like the sense given by probabilistic, counterfactual or nomic accounts of general causal claims. The relation must also satisfy certain independence conditions. I have proposed three; I suggest that the two stronger ones, *mmi* and *smi*, do a better job than alternatives proposed by others.

Understanding causal relevance is one key to understanding causal explanation, since many causal explanations work by picking out their objects under specially related descriptions, ones that express properties that are in the causal relevance relation. But some causal explanations pick out properties of events that are not independent in the right way; this shows one way causal relevance and causal explanation are different relations. (Another difference is that some (very few) causal explanations explain simply by pointing out a causal relation between events.)

The difference between causal relevance and causal explanation can be used to explain why philosophers have argued, on the basis of conceptual connections between reasons and actions, that reasons cannot cause actions. These conceptual connections do show that being a reason is not causally relevant to being an action. They do not show that a reason cannot cause an action, or that a reason cannot causally explain an action.

¹⁸There is another important difference in the strength of *smi* and *cmi*. Externalist theories of representational content hold that persons have semantic properties in virtue of relations between their bodies and things outside their bodies. *cmi* and its natural extensions entail that nothing is causally relevant to being a propositional attitude (construed externalistically) and that being a propositional attitude is not causally relevant to anything. *mmi* and *smi* have this consequence only if being a propositional attitude (or being some particular kind of propositional attitude) requires having a certain kind of cause. One can accept *mmi* and *smi* and be an externalist either by saying that semantic properties are relational and not causal, or by saying that they are relational and causal, but no particular semantic property entails causation by any particular property.

8.2 Causal Relevance and Strict Law

We now have a metaphysical base from which to consider the claim that that Anomalous Monism entails the causal irrelevance of the mental.¹⁹ I believe this view is mistaken, and that it rests on a very simple mistake. In section 2.1 I present one form of the argument that anomalism leads to causal irrelevance, and point out the simple mistake. In Section 2.2 I consider a refinement of the basic argument proposed by Ernest Sosa, and develop a supervenience thesis that shows why the refinement does not make the basic argument any better.

8.2.1 The Basic Problem, and the Solution

John Searle presents the skeleton of the argument that anomalism entails causal irrelevance as follows.²⁰ The first three premises are intended to describe features of Davidson's views about causation, laws and psychological descriptions.

- (1) Events, if causally related, are causally related no matter how described.
- (2) Events instantiate causal laws only under certain descriptions.
- (3) Events are logically related only under descriptions other than those mentioned in (2).

Searle takes the relation between a perception and the objects of perception to be a logical relation. If we understand relations involving propositional attitudes as a species of "logical" relation, then (3) makes the claim of anomalism: psychological aspects of events are not those under which they instantiate strict causal laws.

Searle adds two premises:

¹⁹See LePore and Loewer, *op.cit.*, for a good discussion of the problem and references to other articles in which the supposed difficulty has been urged. I argued in Part I that their notion of metaphysical independence was not quite the right notion to use in barring certain claims of causal relevance. I do not fully agree with their solution to the problem they see Sosa presenting. They seem to accept the sort of condition on causal relevance that Sosa suggests. I do not think we should accept the conditions. The solution that I propose is, however, structurally quite like theirs; see below, Part III, sections 3.5 and 3.6.

²⁰in "Intentional Causation and Practical Reason", unpublished.

- (4) For any two events related as cause and effect, only certain aspects of the events are causally relevant.

Recall Sandor from the beginning of Part I: suppose Sandor eats Dover sole with sauce béarnaise for the first time in his life. He dies, because the fish was infected with salmonella. The fact that this was the first time Sandor ever ate sole with sauce béarnaise is not causally relevant to his demise. If the sole had not been infected with salmonella, he would not have died (not from food-poisoning, at any rate), so the salmonella are causally relevant to the death.

The last premise is this:

- (5) The aspects of events which are causally relevant are the aspects under which they instantiate causal laws.

The argument for (5) is that the whole point of formulating exact causal laws is to capture what makes a difference in the world. We don't expect to find laws relating first occasions of eating sole with sauce béarnaise because we don't think that makes any difference; we do think the salmonella make some difference; and we think that the ultimate causal laws will specify exactly what is important in any causal transaction.

The aspects of events under which they subsume strict causal laws do not include mental aspects, by (2) and (3). By (4) and (5) mental aspects of events are causally irrelevant. Their presence or absence makes as much difference to what's going on in the world, as the fact, that this was the *first* time, made to Sandor's death.

Certainly there is much that is obscure about this argument, and certainly there are many ways it might be improved or modified. The central difficulty is premise (5). There is no good reason to connect causal relevance with strict law in this way.

If we understand causal relevance in the way I suggested in Part I, causal relevance is a relation on properties or aspects of events. It is extremely difficult to state an adequate set of necessary and sufficient conditions for two properties to stand in this relation. But the idea is reasonably clear: the properties should be independent in a certain way, and citing one should constitute a certain kind of explanation of the other. The difficulties come in saying what kind of explanation: perhaps we can get away with a sophisticated probabilistic treatment, perhaps we will need certain modal resources, or perhaps we will conclude that this relation, like the causal relation on two particular events, is undefinable. No matter how

we articulate these theories, though, we must make room for the plain fact that aspects of things may stand in the causal relevance relation even if the connection between the properties is far from being a strict law.

It should be clear that the basic argument does not make the mistake of holding that *mental events* are causally irrelevant. There is a relation of causal relevance on particular events which is a different relation from the one I investigated above. Causal relevance in this sense is closer to the relation of event causation than to the relation of explanatory relevance. Hence reports of causal relevance in this sense are not sensitive to the ways the events are described. Even if the basic argument were successful in showing that mental *aspects* of events are not causally relevant, it could not show that mental *events* do not do anything, since it is the very mental events that have the aspects claimed to be causally relevant.

There is another, related, and somewhat more subtle error that this version of the basic argument avoids. There exists, apparently, a strong temptation to hold that there is no causal relation between events unless it is a causal relation *in virtue of* one or another of the properties of one of the events. Consider the following characterization of Davidson's views offered by Jaegwon Kim (Kim, 1983, 267):

According to [Davidson's] account, there are no type-type correlations between the mental and the physical; however, each individual mental event is in fact a physical event in the following sense: any event that has a mental description has also a physical description. Further, it is only under its physical description that a mental event can be seen to enter into a causal relation with a physical event (or any other event) by being subsumed under a causal law.

The sentence beginning "Further" may be read as the innocuous logical truth that the only way to describe an event as governed by a law, and to state the law, is to state the law; but Kim has something far stronger in mind:

whether or not a given event has a mental description... seems entirely irrelevant to what causal relations it enters into. Its causal powers are wholly determined by the physical description or characteristic that holds for it; for it is under its physical description that it may be subsumed under a causal law.

Two connections are being made here: the first, between causal relations and descriptions, the second, between descriptions and powers. Both are confused.

Davidson's principle of the nomological character of causation does not claim that if two particular events stand in the causal relation they do so in virtue of certain properties rather than others, or that they do so because they are subsumed under a causal law; it claims only that if they stand in the causal relation then there is a causal law. Whatever causal powers are, there seems to be no need to connect having causal powers with strict causal laws.

Kim and Davidson agree that some thesis of supervenience or other is true. There is a very weak sense in which what Kim writes is correct, given supervenience. The heart of his claim is that all the causal powers of an event are determined by its physical characteristics. Suppose all the causal powers of events are based on properties the events have that supervene on physical properties. Then the causal powers of an event are determined by its physical properties. Kim wants to draw the further conclusion that if a causal power is determined by a physical property, then it is a physical power. But surely this does not follow.

8.2.2 Sosa's Principle and Supervenience

Some strong intuition lies behind the claim that if one property is causally relevant to another then there is a strict law linking them. Ernest Sosa holds that Anomalous Monism has a problem about causal relevance (Sosa, 1984). He formulates a somewhat different test for causal relevance, based on a more specific diagnosis of where he thinks Davidson's view goes wrong.

The core of Sosa's idea is that physical properties are such that they operate on their own. If mental events lacked their mental properties, the effects of the events would be the same:

assuming the anomalism of the mental, though my extending my hand is, in a certain sense, caused by my sudden desire to quench my thirst, it is not caused by my desire *qua* desire but only by my desire *qua neurological* event of a certain sort... the being a desire of my desire has no causal relevance to my extending my hand (if the mental is indeed anomalous): if the event that is in fact my desire had not been my desire but had remained a neurological event of a certain sort, then it would have caused my extending my hand just the same. (p.278)

The locution "caused ___ *qua* ___" is essentially the same locution as "caused ___ in virtue of ___" and embodies essentially the same confusions as I pointed out

in section 2.1. Sosa contributes a characterization of what “caused ___ qua ___” means, however, in terms of a counterfactual comparison.²¹ LePore and Loewer suggest one way to understand this idea (LePore and Loewer, 1987, 638). Let F and F^* be properties of an event c and G be a property of e , an effect of c . Then F^* “screens off” F relative to events c and e and property G iff if c were to lack F but still be F^* it would still cause e and e would still be G . A property of an event is causally irrelevant to a property of an effect if there is some other property of the event that screens it off.²² LePore and Loewer put it this way:

Sosa’s Principle: c ’s being F is causally irrelevant₂ to e ’s being G if there is a property F^* of c such that $(F^*c \ \& \ \neg Fc \ > \ Ge)$ holds non-vacuously.

(Irrelevance₁ is the kind that comes with associating causal relevance with strict law. The connective ‘>’ is the counterfactual conditional “if . . . were the case then . . . would be.” ‘ $(F^*c \ \& \ \neg Fc \ > \ Ge)$ ’ would hold vacuously in the cases where if c were not F then c wouldn’t occur at all.²³)

There are several responses available to those who believe that Anomalous Monism has no serious difficulty with causal relevance. They divide into two groups: responses that deny the condition on causal relevance, and those that accept the condition but deny that it applies to the case of the mental.

LePore and Loewer take the former route. They point out an unhappy consequence of Sosa’s Principle. Suppose c ’s being F^* screens off c ’s being F with respect to e ’s being G . If F and G are linked by some rough law, and F and G can

²¹There is one absolutely fundamental difficulty in the way of understanding Sosa’s remarks. We must understand what it means to say some actual event would be different in some ways if it were different in others. The difficulty comes with the claim that some actual event is the same event as some other that has different properties.

In what follows I will write as though there simply is no difficulty. I assume that either the difficulty can be overcome, or that what I say can be rewritten dropping reference to identical individuals whose properties differ.

²²Notice that the causal relevance of properties is relativised to a property of effects. A property might be *strictly* causally irrelevant if there is *no* property to which it is causally relevant. Sosa’s Principle does not, even according to Sosa, entail that mental properties, understood as Davidson understands them, are strictly causally irrelevant.

²³Notice that both Sosa and LePore and Loewer hold that causal relevance is a four-place relation, on two particulars and two properties. I think this is a mistake. First, it encourages the confusion between particulars and properties. Second, more seriously, I think it gives the wrong results; see above, notes 13 and 18. I prefer to make causal relevance a relation between properties only.

be realized by a variety of physical properties, F also screens off F*. If Hurricane Donald lacked the physical properties it has (the precise evolution of distributions of water molecules, for instance) it would still devastate the Keys; hence the physical properties of hurricane Donald are causally irrelevant₂ to its effects.

I agree that Sosa's Principle should be rejected, but I do not think this is a good reason to reject it. Sosa might respond to LePore and Loewer that perhaps it is true that the detailed physical properties of Hurricane Donald are not causally relevant to this effect, but that does not save the causal relevance of the hurricane. Surely there is some other, more general, physical property of the hurricane that is preserved just so long as the hurricane produces this effect. This more general property is not screened off by the property of being a hurricane.

In section 3.5 below I shall indicate my reasons for rejecting Sosa's Principle. But for now it is worth pointing out that even if Sosa's principle were correct, it would not rule the mental as causally irrelevant.

To show this, let us return to the guiding intuition behind Sosa's Principle, the intuition that things would be just as they are if they were to lack the mental properties they actually have. My desire to quench my thirst is causally irrelevant, it is claimed, because if the relevant bits of my neurophysiology were to lack the property of being a desire with that content, they would be just the same; they would still cause the bodily movement that is my action of reaching for a glass of water.

We know that in some respects things could not possibly be the same if my desire were to lack the property of being a desire yet be neurophysiologically just as it is. For instance, the resulting bodily movement would not be an action. But we also know, from Part I, that nothing is causally relevant to my bodily movement's being an action, so we do not yet have a sufficient response to Sosa.

One additional claim, one already made by Davidson, is enough to guarantee that the mental is not shown causally irrelevant by Sosa's principle. There could be a weak relation between instances of mental properties and instances of physical properties such that if a given mental event were to lack the mental properties it has, things would be different physically, and moreover different with respect to the causation of effects with properties we are interested in.

The weaker relation is supervenience. Of course there are many accounts of what supervenience is, and some would not provide a response to Sosa. One that can is *modal supervenience*. Modal supervenience can be developed in a fairly natural way out of what Davidson says about supervenience.

Davidson remarks that AM is consistent with supervenience:

Although the position I describe denies there are psychophysical laws, it is consistent with the view that mental characteristics are in some sense dependent, or supervenient, on physical characteristics. Such supervenience might be taken to mean that there cannot be two events alike in all physical respects but differing in some mental respect, or that an object cannot alter in some mental respect without differing in some physical respect. (“Mental Events”, p.214)

Supervenience is a modal claim: it says something about how things must be, or about what is necessarily so. There are two distinct ways to understand what modal claim Davidson is making. To see this, consider first a *non*-modal formulation of the first claim, i.e., a generalization only about actual objects. I’ll use the language of Part I to phrase the claim, supplemented with a pair of world quantifiers, ‘(w)’ and ‘(Ew)’, for ‘for all worlds w’ and ‘there is a world w’:

(S₁) (x)(y) if

(pX) (physical property(pX) → (x has pX iff y has pX))

then

(pY) (mental property(pY) → (x has pY iff y has pY))

The quantifiers ‘(x)’ and ‘(y)’ range over events, and the property quantifiers range over properties of events. In words the formula reads, for all pairs of events, if they share all physical properties they share all mental properties. If there are no other properties than physical properties and mental properties, and indiscernible events are identical, then if two events share all physical properties then they are the same event. S₁ is true if all actual events that share all physical properties also share all mental properties.

The first way of making S₁ a modal claim is “extensional supervenience”. It says that every world is such that S₁ holds in that world; nothing is said about how any event that occurs in any world would be in any other world. Mental properties extensionally supervene on physical properties if and only if

S₂ (w) S₁

or, more explicitly,

$S_2(w)(x)(y)$ if (pX) (physical property(pX) \rightarrow

(x has pX at w iff y has pX at w))

then

(pY) (mental property(pY) \rightarrow

(x has pY at w iff y has pY at w))

The second way of making S_1 a modal claim is “modal supervenience”. It makes the somewhat different claim that if any event in any world is physically just the same as an event in another world then the events are mentally the same. (Again, on reasonable assumptions they are the same event; the extra reasonable assumption is that events in different worlds can be the same event, i.e., that the cross-world identity relation is defined and not empty.)

$S_3(w)(x)(w')(y)$ if (pX) (physical property(pX) \rightarrow

(x has pX at w iff y has pX at w')

then

(pY) (mental property(pY) \rightarrow

(x has pY at w iff y has pY at w'))

How are S_2 and S_3 related? S_2 is weaker than S_3 : it is entailed by S_3 but does not entail it.²⁴ First, S_2 doesn't entail S_3 . We can describe a model in which S_2 is true but S_3 is false. Let S_1 be true in all possible worlds; that makes S_2 true. Let there be at least one pair of worlds $\{w_1, w_2\}$ such that e in w_1 and f in w_2 are physically indiscernible, but there is some mental difference between them. S_3 is false in this model.

S_3 entails S_2 , since if S_3 is true then any world is such that if two events in it are physically indiscernible then they are mentally indiscernible. Sosa's Principle involves a subjunctive claim about events, so I will work with S_3 .

²⁴There is one interesting difference between them as explications of Davidson's idea. He suggests that supervenience and dependence are essentially the same relation: “mental characteristics are in some sense dependent, or supervenient” I believe that dependence is a modal notion; if A depends on B in some respect then if B were relevantly different then A would be different as well. S_2 does not capture this idea of dependency, and S_3 does.

We need one further step before we can compare modal supervenience and Sosa's Principle. Sosa's Principle makes a claim about what an event would be like if it lacked a certain property. To get a claim about lacking properties from our supervenience theses, we need the contrapositive of S_3 .²⁵

$S_{3c}(w)(x)(w')(y)$
 if not (pY) (mental property(pY) →
 (x has pY at w iff y has pY at w'))

then

not (pX) (physical property(pX) →
 (x has pX at w iff y has pX at w'))

or, driving the negation in,

$S_{3c}(w)(x)(w')(y)$
 if (EpY) (mental property(pY) and
 (x has pY at w and y does not have pY at w'))
 then (EpX) (physical property(pX) and
 (x has pX at w and y does not have pX at w'))

Or, in words, if two events in two worlds differ in some mental respect they differ in some physical respect as well. If we hold that there is such a thing as cross-world identity, S_{3c} entails that if some event e that is M were to fail to be M, then it would differ in some physical respect.

What physical differences attend on mental differences, if S_3 is true? Just because an object changes in some physical respect doesn't mean that everything that happens around it differs; it may be "for all intents and purposes" just the same. Elijah Millgram suggested this analogy in a different context: if what's important to you about your computer is that it types your papers, changing the color

²⁵Often the contrapositive of a natural law is not a natural law; contraposition does not preserve truth for all modal claims. Supervenience has more modal force than a law, however, since it requires a sentence to be true at every world, nomologically possible or not. In this case contraposition does preserve truth.

of the insulation on the wires isn't going to make much difference. Recall that Sosa's Principle is a condition on whether one property of one event is causally relevant to a property of its effect. Hence what is important for our concerns is the effect property.

All we need to respond to Sosa's principle is modal supervenience, and the claim that, for interesting properties of effects, the physical properties that would differ if a given mental event were different would be these interesting properties. Whether this conjunction is actually true is, of course, an empirical matter; but since it is an empirical matter and a contingent matter, we can conclude that Sosa's principle does not show that Anomalous Monism entails that the mental is causally irrelevant.

There is one small problem remaining. Modal supervenience demands a modal connection between mental and physical properties. This modal connection leads some philosophers to the conclusion that modal supervenience requires psychophysical laws, and hence to the conclusion that supervenience is not compatible with anomalism.²⁶ It seems to me that this conclusion is either false outright, or, if it is true, it is based on an exceedingly peculiar conception of law. Let's go back to S_3 : it says that if two events share every physical property then they share every mental property. Hence there is a law (of sorts) that says any event that has just the physical properties of some mental event has the mental properties of that event as well. But S_3 quantifies over all physical properties; the law says if the complete physical description of some mental event is satisfied again then the complete mental description will be satisfied again as well. It is certain that the complete physical description of any event cannot be satisfied by any other actual event, since among any particular event's properties are its time and place and physical relations to every other event that ever occurs.

²⁶See Kim, op.cit., p.269: "I think the two questions, whether intentional psychological states are supervenient on the physical and whether they enter into *law-based* causal relations with physical processes, are arguably equivalent questions."

8.3 Causal Relevance and Externalism

A suitably formulated supervenience thesis provides an answer to Sosa's claim that mental properties are screened off by physical properties. The causal relevance of mental properties is not thereby vindicated, however. This supervenience thesis is no help in answering the charge that if externalism in semantics and the philosophy of mind is true, then mental properties are screened off by intrinsic properties of bodies. I turn now to showing how to answer this difficulty.

My solution depends on a structural feature of the properties involved. Normally in testing whether some aspect of events is causally relevant we hold other aspects of the events fixed and determine what differences ensue from varying the aspect under consideration. The trouble is that we cannot properly do this in comparing relational and non-relational properties of events, since a relational property always requires the other elements of the relation. We then have a quandary: we *cannot* hold all other aspects of events fixed, and if we *do* not hold them fixed then it is easy to show that these other aspects make differences to what properties effects instantiate.

I will trace this quandary through two different proposals for tests for causal relevance that are designed to show why relational mental properties are not causally relevant (in contrast to intrinsic properties of bodies). In each case I'll exploit other differences in situations where relational properties are instantiated in order to show that the tests fail to show that relational properties are causally irrelevant.

In the last section of this Chapter I show how this problem looks in the framework of Nancy Cartwright's probabilistic account of general causal claims. Her account is in a sense designed to ensure that *ceteris is paribus*, since it is designed to handle cases of common causes and spurious correlations. It turns out that there is a difficulty in determining whether a non-relational property is causally relevant if a relational property is one of the possible causally relevant factors. A condition can be added to Cartwright's framework that avoids this difficulty and permits the causal relevance of both relational and non-relational properties.

8.3.1 Externalism

Externalism is the thesis that what meaning a term or a sentence has, and what content a thought has, is not an intrinsic feature of the agent who uses the term or sentence or has the thought. Instead, meaning and content are fixed partly by

intrinsic features of agents and partly by their relations to things around them.

Hilary Putnam argued for externalism about the meaning of natural kind terms in “The Meaning of ‘Meaning’” (Putnam, 1975). He asks us to imagine that there is a world just like our world, called Twin Earth, populated with Twins: people who are physical duplicates of ourselves with respect to properties bodies can have regardless of their relations to other things. This world is so like our world that if you were put in your Twin’s place you would be utterly unable to distinguish that situation from your current situation. There is one difference: the liquid they call ‘water’ is not water but some substance distinguishable from water only by the subtlest chemical investigation. Call this stuff ‘twater’. Putnam urges that their term ‘water’ means twater, not water. He suggests that the meaning of natural kind terms is set indexically: my term ‘water’ means the stuff, the liquid in the world around me, which is essentially the same kind of stuff as the liquid to which I learned to apply my term.

There are many ways to develop Putnam’s ideas, and there are many ways to argue for one or another kind of externalism. All the accounts of content we considered in Chapters 3 through 6 are externalist, since they claim that the content of a representational state is determined by the causal and normative relations the state has with something outside the owner of the representational state.

Davidson has recently described a situation that makes his own sort of externalism particularly vivid.²⁷ In order to arrive at an interpretation of another’s sentences, Davidson suggests, we should start with information about what causes an agent’s preferences between sentences. Hence he treats the actual causal history of an agent as essential to the fact that the agent means or thinks anything at all.

So now suppose Donald is out for a walk in Tilden Park. He’s walking through a swamp at the end of Lake Anza. A storm comes up. He is struck by lightning and vaporized. At the same time a fallen tree is also struck by lightning—this is a thoroughgoing accident—and its molecules are miraculously assembled into a duplicate of Donald. This creature, the Swampman, continues Donald’s walk, appears to chat about radical interpretation, and so forth.

Davidson holds that the Swampman’s words, at least at first, mean nothing, and that the Swampman has no thoughts. I will not discuss here the question whether Davidson is right in this claim. If the claim shows that our mental life is causally irrelevant in the sense under discussion, then there must be something

²⁷In “Knowing One’s Own Mind” (Davidson, 1987).

wrong with the views that lead to it. My question then is whether, if we hold that the Swampman has no thoughts, we are thereby committed to the conclusion that the mental properties Donald has and the Swampman lacks are not causally relevant.

The argument that the case of the Swampman so commits us can be represented in the following table:

	mental event		bodily movement	
Donald	P	R	R''/R'''	A
Swampman		R	R''	

Suppose P is some propositional attitude property that Donald has and the Swampman lacks. R is a physical property of the event of Donald's coming to have this propositional attitude. By hypothesis, this is a property that both Donald and the Swampman can instantiate. Suppose the event of Donald's coming to have this propositional attitude causes a bodily movement. This attitude, together with others, rationalizes the bodily movement; the resulting bodily movement is an action of Donald's, A. The bodily movement has a variety of physical properties. R'' is a physical property that can be instantiated both by Donald and the Swampman. R''' is a relational physical property of the bodily movement, one that Donald can instantiate, but the Swampman cannot, since it lacks the relevant relations to things.

The argument that the propositional attitude property P is causally irrelevant then runs as follows. Consider the effect property R''. Events in both the Swampman and Donald would cause an event with this property. The causes both have the property R. Apparently the fact that a cause event has, in addition, property P, is not relevant to its causing an event with property R''. Hence property P, the propositional attitude property, is not causally relevant to property R''.

In Part I above I showed that one possible answer to this argument is inadequate. Property P clearly makes a difference to whether the effect event has property A: Donald's bodily movements are actions, and the Swampman's are not, and the difference comes with the presence or absence of the propositional attitude. The trouble is that propositional attitude properties are not causally relevant to action properties, since actions are too closely conceptually connected with reasons.

Property P might, however, be causally relevant to R'''. Both are relational

properties, neither can be instantiated by something like the Swampman. Perhaps two relational properties may stand in the causal relevance relation.²⁸

8.3.2 Fodor's Argument

Jerry Fodor thinks causal relevance is a problem for externalism (Fodor, 1987, Chapter 2). He argues that relational taxonomies are not causal taxonomies, on the basis of a test for sameness of causal powers. I will show that his test does *not* show that me and my Twin, or my Swamp-Doppelgänger, have the same causal powers. This demonstration shows two things. First, relational properties can, indeed, be causally relevant to other relational properties. Second, there is a real difficulty, for tests like Fodor's, that stems from the need to ensure that *ceteris is paribus* when we investigate the causal powers of things.

Fodor suggests the following analogy. I have a dime that I toss now and again; it comes up heads or it comes up tails. It's easy enough to produce a new taxonomy of everything in the universe by associating it with the state of my dime: for instance, *x* is an H-muon just in case *x* is a muon and my dime comes up heads. That's a relational taxonomy if ever there was one. It's also obviously not a causal taxonomy, since H-muons and T-muons do just the same things: what muons do.

Now consider my Twin and I: here I am, staring dully at a glass of water, thinking that it's a glass of water, and there my Twin is, thinking that there's a glass of ZYX in front of him. Fodor claims that the difference in the semantic properties of these two mental states is just like the difference between an H-muon and a T-muon. The difference between the "property of being a mental state of a person who lives in a world where there is ZYX rather than H₂O in the puddles" and the property of being a mental state of a person around here is "irrelevant to their causal powers" (p.34). If that's right, that would be a good reason to suppose that this semantic difference is causally irrelevant.

But *is* it right? Not all relational taxonomies are thereby non-causal taxonomies. Fodor considers a relational causal taxonomy: 'planet' categorizes things relationally, since for a ball of rock to be a planet it must rotate around a star. Fodor has two ideas about what makes a relational taxonomy causal. About planets he writes,

The property of being a planet is [causally] taxonomic because there

²⁸Jennifer Hornsby proposes a very similar solution to a slightly different problem in "Physicalist Thinking and Conceptions of Behaviour", draft, 1985.

are causal laws that things satisfy in virtue of being planets. By contrast the property of living in a world in which there is ZYX in the puddles is *not* taxonomic because there are *no* causal laws that things satisfy in virtue of having *that* property. (p.43)

Perhaps; but it is a causal law of some sort or other that Twin Earthers who intend to pour a glass of ZYX do pour a glass of ZYX. It's rough but it's a law.

His other idea is more substantive: a test for sameness of causal powers. He imagines someone responding that there is indeed a difference in causal powers: When I say, "Bring water!" water arrives, when my Twin says, "Bring water!" ZYX arrives. Fodor rightly says this won't do. To compare causal powers we have to see whether the items behave the same way *in the same contexts*:

What *is* relevant to the question of identity of causal powers is the following pair of counterfactuals: (a) If his utterance (/thought) had occurred in my context it *would have had* the effects that my utterance (/thought) did have; and (b) if my utterance (/thought) had occurred in his context, it *would have had* the effects that his utterance (/thought) did have. For our utterances (/thoughts) to have the same causal powers, both of those counterfactuals have to be true. But both of those counterfactuals *are* true (p.35)

I suggest, on the contrary, that they are not.

First a simple case. I say "Bring water!" The interlocutor brings water. My twin is put in my place. He says "Bring water!" The interlocutor may well not bring water. Why not?

Even though my twin is put in my place he remains importantly different from me. If he is to count as ordering that ZYX is to be brought (and to belong to a different category in a relational taxonomization than I) he must actually have a history of the sort that makes his word mean ZYX. Perhaps he was brought to this world. He is to be as like me as two neckties or as two peas are alike; he is to be like me down to the very distribution of atoms in his body. But his actual history must be his, not mine.

But suppose the interlocutor knows about his history. My Twin's order might lead him to bring ZYX if he can, or to say something about why he cannot, or to say, "You know, Tony, things aren't as they seem ...". Of course he need not do any of these things, but he might.

I think Fodor is tempted to say that we can't count among the causal powers of mental states those that depend on knowledge of their semantic properties. But that begs the question. We interpret others and are led to do things by our ideas about what they think and mean. I think our best reason for thinking that all the semantic properties of mental states are causal powers is precisely that they have effects in virtue of interactions in communication. We're free enough to ignore the obvious mistake my twin makes; we can treat him as simply meaning that we should bring water. But in some circumstances it's important to find out what the actual cause of an agent's current state is. It's important, for instance, if there is a legal contract involved. It's important if memory is an issue.

In any event, since the difference in their semantic properties traces to a difference in their physical histories, we can build some sort of device which will do the interlocutor's job. It has to be sensitive to more than the *intrinsic* physical state of the two agents, since by hypothesis that is exactly the same. It must be sensitive to their history. It would suffice if the device has some way of keeping track of the current agents from their birth.

In fact the situation seems to be exactly analogous to the planet case. Consider a meteor on a certain trajectory in space. A ball of rock that is not a planet is moving along a path that intersects that of the meteor. To see whether a planet has the same causal powers as the ball of rock, we need to bring in the whole system that gives the planet its property of being a planet. Since it's a planet its path differs, so it doesn't have the same causal powers: it doesn't smash into the meteor.

But if we can "leave the situation the same" in introducing a sun, we can also leave it the same by introducing not only the required historical difference between the two agents but also certain of its effects, among them the interlocutor's knowledge. The historical difference plays the role of the sun in the planet case: it exerts a certain "gravitational pull" on current events (mediated by the interlocutor's knowledge) which yields a "swerve" in the effects like the swerve of the planet.

The trouble is that Fodor's test for sameness of causal powers simply doesn't work for relational properties. (This seems to be the point of Martin Davies' objection, Fodor, p.158, note 9.) There are some circumstances in which putting a planet in place of a ball of rock *must* change something else on the scene. Similarly, it is strictly impossible to put my Twin in the same situation as I am in, since my situation necessarily differs from his in endless ways. Normally when we test for sameness of properties we try to ensure that all else is equal: that *ceteris*

is *paribus*. But in testing relational properties against non-relational properties *ceteris is never paribus*.

Why does Fodor think the two agents will produce exactly the same effects? The trouble comes from an unmotivated restriction on the notion of a situation.

The point of taxonomizing relationally is an interest in relations rather than (exclusively) causal explanation. (Fodor makes a nice related point about attitude ascriptions for the sake of information transfer, p. 157, footnote 4.) My case exploits this difference; it says, let someone know what the relations are, then some property of some effect will be different. We might suppose there's a way for the agents themselves to produce different effects, simply in virtue of being two different people. Removing the interlocutor then requires finding some situation or context which "presents the truth" to the twins. Present sufficient information (the *same* information) to both of them that each can figure out which one *he* is from that information, and you can then generate any difference you like.

Unfortunately this is no recipe for generating cases. Suppose to start with that situations are individuated in some way that allows me and my twin to be placed in the same situation. The idea is something like this: a situation for a person is a temporally extended set of local environmental states and changes such that the person does or could alter her epistemic state about them. More tidily we picture a person confronted with various objects and events which she can see or sense and form beliefs about, perhaps through causally interacting with them.

By hypothesis the agents are molecule for molecule the same, and they are to be tested in one single situation. The trick would be to produce a situation in which one of them would get the information that she is not a local, while the other would not get this information. As far as I can see this cannot be done. To produce that sort of difference in what the agents come to think we would have to have some causal difference "coming in," i.e., we would have to have two different situations, different in some way that would let one of them in on the facts. (Notice that anything that would suffice to make one of them come to believe she wasn't a local would also make the other come to believe she isn't a local. *Confronting* each with the other might seem to help, but for either to tell *which* she is still requires some bit of evidence different for one.)

I think it's this last sort of quandary that makes Fodor's argument seem persuasive. What's happened is that a certain notion of causal powers got *built in* to the notion of a situation. If you can arrange to keep the actual differences between two things from producing different kinds of effects in the situations, then it's trivial

that those two things are indistinguishable with reference to those situations.

8.3.3 Screening Off Revisited

Although Fodor is wrong that relational taxonomies are not causal taxonomies, I do not think this is much help for the externalist. The sort of effect property to which relational propositional attitude properties were shown causally relevant were relational properties of effects, ones, for instance, that bodily movements have in virtue of being caused by things with a certain history. But what of less relational properties of bodily movements, as for instance, the property of being a certain sort of movement of a body, regardless of the history of the body? I think that if an account of the nature of content entails that our mental life is not causally relevant to aspects of our actions like these, then there is something wrong with the account.

In this section I describe another “screening off” criterion for causal relevance, essentially a generalization of Sosa’s Principle. It appears to follow from this principle that relational mental properties are not causally relevant to ordinary non-relational bodily movement properties. In the next section I’ll show it does not follow. The underlying problem is much the same as the problem for Fodor’s condition: if *ceteris* is not *paribus* then the Principle does not entail the causal irrelevance of mental properties; but in comparing relational and non-relational properties *ceteris* cannot be *paribus*.

Here is the Screening Off Principle:

Property P is causally relevant to property Q only if there is no further property R such that P is screened off from Q by R.

The idea behind the Screening Off Principle is that we shouldn’t say that one property is causally relevant to another if instances of the first *would cause* instances of the second even if they lacked the first property:

P is screened off from Q by R just in case all P events have R, P events cause Q events, and if the events that are P were to lack P and continue to be R, they would still cause Q events.

Suppose Q is some relatively non-relational property of bodily movements which many actions of a certain kind possess, that P is a propositional attitude property that the causes of these actions possess, and that R is some physical property

of agents who perform these actions, a property instantiated by all agents with property P and not instantiated by normal agents who do not possess property P. By hypothesis, events in the Swampman have property R but not property P, and these events cause movements of the Swampman's body that are Q. Hence by the Screening Off Principle the property P is not causally relevant.

8.3.4 The Mental Isn't Screened Off

In this section I consider whether the Screening Off Principle really does entail that propositional attitude properties are not causally relevant to properties of bodily movements that both agents and the Swampman instantiate. I'll develop an analogy with a minimally relational property, the property of being a gene, and show that being a gene would also turn out to be causally irrelevant to properties of events of phenotypic expression. The trouble can be traced to a difficulty in the modal force of the Screening Off Requirement.

Genes are things in the lives of biological organisms which are heritable, mutable, and produce phenotypic traits. The notion of a gene is somewhat externalistic, since it applies only to aspects of biological creatures. Most genes are made of DNA, but strands of DNA base-for-base just like actual genes may not be genes. Such strands can occur in cells and not function in the ways I've just specified; clearly they could also come into existence in completely different environments. So imagine a collection of strands of DNA formed in a warm gas cloud in deep space. These are not genes. Now imagine an accident whereby your genes are completely replaced by a collection of strands of DNA from deep space which *happen* to be base-for-base just the same as your genes. Your life goes on, the same proteins are manufactured, the same phenotypic traits are maintained. Hence things which are not genes (at first, at any rate) but which are physically just like genes produce effects much like the effects of genes. Hence there is a property shared by all things which are genes, such that things with that property which are not genes produce the same kinds of effects as genes do. According to the Screening Off Principle, being a gene is causally irrelevant to standard effects of genes: effects which are practically definitive of being a gene.

Something has clearly gone wrong. To see what it is, let's look harder at the screening off requirement.

The requirement is a modal requirement; it says something about how events would be in non-actual circumstances. The modality is clearly not unrestricted logical possibility. Mental properties are logically distinct from physical proper-

ties and hence there are logically possible worlds where all events are physically just as they are in the actual world, but lack all mental properties.²⁹ If this were the appropriate modality, then no properties, mental or otherwise, would be causally relevant according to the Screening Off Principle.

Perhaps the Screening Off Principle should hold only in all nomologically possible worlds. The Principle should be amended as follows:

if P is causally relevant to Q, then there is no property R that all P events possess such that in all nomologically possible worlds where the actual P events lack P but retain R they cause Q events anyway.

This Principle is better, but its modal force is still too great.

I claimed that if something which is a gene and has a certain molecular structure were to fail to be a gene but keep its molecular structure, it would still produce effects that share some interesting properties with the effects of actual genes. (Of course, there would be differences, too.) But this is not quite right. I suggested that strands of DNA just like genes can occur in cells and fail to be genes, and that perhaps strands of DNA can occur in deep space. But those strands of DNA would *not* do what genes do. Strands of DNA just floating around in the cells don't control phenotypic expression. They may have some effect on traits, but not the central causal responsibility we assign to genes. The strands of DNA in deep space aren't inherited, don't control phenotypic traits at all, and there's no reason to call changes in their structure mutations. Therefore being a gene is *not* screened off from phenotypic property Q by the molecular property of being a strand of DNA, since if something which is a gene and produces an effect of type Q were to fail to be a gene and to keep its molecular properties, it *would not* produce a Q effect. It would be floating around in the cell or in a Petri dish or maybe in deep space.

Another way to put this point is to say that in considering causal relevance we are asking whether instances of a given property *would cause* effects of a certain type, if they lacked the given property, rather than whether they *could* cause such effects. Presumably the answer to the latter question is almost always 'yes', since some mechanism or other can be arranged that yields instances of the effect property given events just like instances of the given property except in respect of that property.

²⁹This claim is true only if either the modal supervenience claims of section 2.5 are simply false, or the modal force of those claims is something less than logical necessity.

What's needed is a measure of how close worlds are to the actual world, or a measure of "similarity", as David Lewis puts it. The requirement on causal relevance then reads,

if P is causally relevant to Q, then there is no property R that all P events possess such that in the nomologically possible worlds most similar to our own where the actual P events lack P but retain R they cause Q events anyway.

It is not likely that there is any precise way to characterize the similarity of worlds; how similar one thing is to another depends, as always, on what interests us about the things. But the comparison is one made all the time in scientific investigations. If we are interested in whether things of type A yield things of type B, we do not consider elaborate contrivances whereby things of type A are made to yield things of type B. The interest-relativity of such comparisons shouldn't provide a reason for thinking that there is nothing to the distinction between what would happen and what could happen. In this case it is clear that if a gene were to keep its molecular properties and fail to be a gene, it wouldn't be doing what genes do.

Let's turn back to Donald, the Swampman, and Twin Donald. The argument that externalistically understood mental properties are not causally relevant to certain properties of bodily movements was that, with respect to a property Q of bodily movement events, if mental events were to lack their mental descriptions but retain their physical descriptions, they would do the same, i.e., still cause bodily movement events that are Q. (Of course the effect events must differ in some ways; the question is whether they would differ with respect to Q.) I suggest we were incautious in accepting this claim. Certainly they *could* do the same. But there are possibilities much closer to home than Twin Earth or the world of Swampman where mental events lack their mental properties but *do not* do the same.

One of the victories of molecular genetics is that it gives us a clear sense of what an individual gene is off by itself: it is a strand of DNA with certain characteristics. We've even got photographs of individual genes. As a result, we have little difficulty thinking about strands of DNA that fail to be genes. We don't have electron micrographs of the physical aspects of single thoughts. But there is reason to think that events of coming to have particular thoughts must have distinctive physical properties.

In "Actions, Reasons, and Causes" Davidson argued that reasons can be causes, and furthermore that there is one positive reason to think that they *are* causes. A person may have many reasons for an action, but act on less than the entire set.

Davidson proposes that the reason that explains the action is the one that causes it.

This account of reason explanation requires that prior to the action there are several events, one corresponding to each reason the person has for performing the action. Each is the event of coming to have that reason. The event of coming to have the reason that explains the action has physical properties linked by a strict causal law to physical properties of the action. The physical properties of the event that explains the action must be different in some way from the physical properties of the other events. Otherwise there would be no difference between the events from the perspective of causal laws relating properties of events.

Consider some particular mental event: it has its distinctive physical properties and mental properties. What would this event be like if it lacked its mental properties? Perhaps we can imagine a skein of brain tissue which is just the portion which is proximally responsible for my arm going up. Or perhaps we can think of a person so damaged that her behavior no longer warrants attribution of propositional attitudes, but such that some event that would have caused her arm to go up retains its physical properties. Here holism is a resource, since whether some particular event warrants a mental description depends on how it is situated among countless other events. Disturb enough of them while leaving the particular event alone and it loses its mental description.

It is hard, as I said, to know anything definite about how similar worlds are one to another. Suppose there is a set of non-relational physical properties. Physical duplicates are things that share all properties from this set. Complex physical objects are objects that have a very complex physical description. Persons are complex physical objects, as are fruits, while, say, viruses are not complex physical objects. I suppose that worlds that contain physical duplicates of complex physical objects are somewhat less similar to the actual world than worlds in which complex physical objects are physically similar but are not duplicates. I suppose that the more complex the physical object, the more dissimilar a world containing it and a duplicate of it will be from the actual world. So it seems to me that the world in which Donald is struck by lightning and the Swampman appears in his place is far more dissimilar from the actual world than a world in which a single mental event of Donald's is physically just as it is but lacks its mental properties. The world that contains both Twin Earth and the Earth is at least as dissimilar from the actual world as the world that contains the Swampman.

I suggest that in asking whether the Screening Off Principle shows that mental properties are causally irrelevant, we should not be distracted by Twin Earth

or the fable of the Swampman. We should concentrate on these more prosaic possibilities.

The question to ask now is this. Suppose I go to the refrigerator to get a beer. My reason includes the desire for a beer. The event of my coming to desire a beer causes the bodily movement of going to the refrigerator. The cause has property P, coming to have a desire for a beer; the effect has property Q, being an event of going to the refrigerator. Consider the nearest nomologically possible worlds in which this actual mental event lacks property P, but is physically just the same, with respect to the set of intrinsic physical properties. This is not the world in which I have a Twin; it is rather a world in which something less drastic is the case. Does this event produce an event which has Q?

The question is clearly an empirical question: against the background of the kinds of possibilities that ought to be considered relevant, do events like my mental event cause events like the bodily movement I undergo? The answer depends heavily on empirical facts, on the empirical details of how mental events are realized and what can be done to them. What's important here is that the Screening Off Principle does not rule *a priori* that the answer must be 'yes'.³⁰

8.3.5 Rejecting the Principle

I believe the argument of the last section shows that the Screening Off Principle does not entail that mental properties are causally irrelevant to properties of bodily movements to which we think they clearly are causally relevant. But the argument raises deeper worries about the status of the Principle itself.

The difficulty is connected with the difficulty we had in evaluating Fodor's claim that relational taxonomies are not causal taxonomies. If one thing has a relational property and another does not, and the two things are exactly similar according to some standard of similarity (e.g., the set of intrinsic physical properties I mentioned in the last section) then there is something about the situations of the two things that is different: whatever makes for the instantiation of the relational property. One way to find out whether an aspect of an event is causally

³⁰It is important to note that the Screening Off Principle certainly does not favor a more internalistic construal of mental properties. We are asking about events and their local properties. We envisage duplicating these local properties and at the same time disturbing the local or internal relational properties of these events (for instance, their relations to other mental events that determine their characterizations as mental events) enough to disturb their mental properties. This procedure is insensitive to the difference between internalism and externalism.

relevant to an aspect of an effect is to ask whether things with that aspect cause, *ceteris paribus*, things like the effect. The trouble with relational properties is that *ceteris* is not *paribus*, since something *else* must differ if the relational property is instantiated.

The argument of the last section raises this difficulty in an extreme way. There are various ways the relations bodies have with things around them that make for semantic properties can be disturbed. They can be disturbed in the relatively pure way that the case of the Swampman illustrates, or they can be disturbed in the messier ways the cases I described illustrate. My argument hinged on the claim that the messier ways are far more likely to be realized than the pure ways, and that in evaluating causal relevance with a counterfactual test like the Screening Off Principle we should consider only the more likely possibilities.

But it might fairly be objected that this is the wrong way to formulate a test for causal relevance. Consider genes again. It is not enough, the objection runs, to ask what some actual gene would do in reasonably different circumstances. We need to compare being a gene with not being a gene both in normal circumstances and in peculiar circumstances: i.e., as far as possible, we should consider whether being a gene is causally relevant to some effect property holding other things as constant as we can. So we would need to consider 4 kinds of circumstances. We are considering strands of DNA. We need to consider strands in the nucleus, where genes normally do their work, and then investigate the difference that being a gene makes to the interesting effect property. Further, we need to consider strands outside of this context, either in other places in the cell, in a Petri dish, or in deep space, and again investigate the difference that being a gene makes to the interesting property.

At first glance this improvement on the condition for causal relevance appears to rule being a gene *not* causally relevant to interesting effect properties, since a molecule that is base-for-base just like an actual gene but (somehow) fails to be a gene will do pretty much exactly the same things as genes do, in the same contexts.

In the next section I will show that there is still a difficulty in the way of a straightforward claim that relational properties are not causally relevant.

8.3.6 Probabilistic Causality

Nancy Cartwright's theory of general causal claims was constructed in response to the need to ensure that *ceteris is paribus*. Probabilistic theories of causation are based on the idea that a cause should increase the probability of an effect; a theory of general causal claims like Cartwright's holds that the probability of events of the effect type should be higher conditioned on events of the cause type. One central difficulty for any probabilistic account of causation is that such probabilistic relations may stem from common causes, rather than a causal relation between events of the classes under consideration. R.A. Fisher objected to the claim that smoking causes cancer on the ground that, although the probability of getting cancer given that one smokes is higher than the probability of getting cancer in general, the reason might be that there is a common cause of both smoking and cancer.

Cartwright's solution to this problem is to control for all causally relevant factors. Then a general causal claim 'A causes B' is true if and only if the probability of B events on A events is higher than the probability of B events in general, in all test situations homogeneous with respect to causally relevant factors for B. More precisely,³¹

CC: C causes E iff $\Pr(E/C.K_j) > \Pr(E/K_j)$ for all state descriptions K_j over the set $\{C_i\}$, where $\{C_i\}$ satisfies

- (i) If $C_i \in \{C_i\}$ then C_i causes (suppresses) E
- (ii) $C \notin \{C_i\}$
- (iii) for all D , if D causes (suppresses) E then either $D = C$ or $D \in \{C_i\}$
- (iv) If $C_i \in \{C_i\}$ then it is not the case that C causes C_i

CC is an impredicative condition on causation: the term "causes" is used on both sides of the condition. CC does not, therefore, define the notion of causal relevance; it rather imposes probabilistic conditions on causation. Cartwright holds, and I concur, that we must invoke the notion of causation somewhere in giving

³¹(Cartwright, 1983, 26) I have altered Cartwright's syntax slightly: where she has a hooked arrow I write "causes"; where she has a hooked arrow followed by '+_' I write "causes (suppresses)"; where she has an arrow with two horizontal bars I use the material conditional "if ... then ..."; and for her quantifier and set membership symbols I use "for all" and "is (not) in".

a probabilistic account of causation; equivalently, there is no purely probabilistic account of causation.

Let P be some property had by events of coming to have a propositional attitude of a certain content; Q a property of bodily movements, one to which we think our mental life should be causally relevant; and R a physical property had by all events that have P , one that events occurring within the Swampman could have. Assume finally there is a set of factors D that are also causally relevant to Q according to CC. Our question is whether CC licenses P as causally relevant to Q , and whether CC licenses R as causally relevant to Q .

Let us consider first whether P is causally relevant, assuming that R is causally relevant. There is a set of test situations K built from the set of properties D and the property R . Then P is causally relevant to Q just in case for each test situation K_i ,

$$\text{Prob}(Q/P.K_i) > \text{Prob}(Q/K_i).$$

Suppose K_s is a partition that holds R fixed; it contains Donald and an action of his which is Q and the Swampman and a bodily movement of the Swampman's that is Q . In that partition the probabilities are equal.

The argument so far is incomplete. The Swampman is a particular, while CC is a claim about classes. Nothing follows from the particular case about the class. Put differently, the Cartwright claim about causal relevance does not *entail* that $\text{Prob}(Q/P.K_i) \leq \text{Prob}(Q/K_i)$. To get the entailment we need something additional, as, for instance, that the likelihood of getting Q with or without P is equal.

The needed additional premise is not a serious difficulty, however. By hypothesis, there are many interesting bodily movement properties such that both actual agents and all Swampman-type physical duplicates of actual agents would undergo bodily movements with just those properties. Hence there is nothing unreasonable in the additional premise, that the probability of getting effects of those types is equal, whether or not the causes are P .

Hence so far we have the conclusion that P is not causally relevant with respect to Q ; the reason is roughly that P is probabilistically screened off by R .

But only half the battle has been won. So far we have simply assumed that R is among the causally relevant factors for Q . We need now to see whether this assumption is certified by CC, i.e., we need to consider whether R is causally relevant to Q . Suppose now the K_j is the set of state descriptions generated with P and properties from D . We need to include P since we have no ground for holding

that P is not causally relevant to Q that does not simply assume that R is causally relevant, and we cannot leave it out unless we simply assume that P is causally irrelevant. The partitions where P is not present are no difficulty: we count the events that are both R and Q, and the events that are Q, and determine whether $\text{Prob}(Q/R.K_j)$ exceeds $\text{Prob}(Q/K_j)$. The trouble comes when the K_j includes P. By hypothesis all P events are R events. Hence $\text{Prob}(Q/R.K_j.P) = \text{Prob}(Q/K_j.P)$. Hence the probabilistic relation required for R to be causally relevant to Q is not satisfied. So neither P nor R is causally relevant to R, according to CC.

Cartwright's account requires a judgment of which factors are causally relevant before we may judge of a particular causal claim, unless the probabilistic information alone determines which factors are causally relevant. I assume that the probabilistic facts alone do not determine which of P or Q is causally relevant. I believe that R is causally relevant, on the ground that there must be good causal explanations of bodily movements in terms of body mechanics, and these explanations rely on causally relevant properties. P, however, makes problems for CC confirming the causal relevance of R to Q. I think we should modify CC to avoid this "pathology." The set of background conditions against which we evaluate a claim that C causes E should not include a property C' such that either if something is C it is C' or if something is C' it is C. More formally, we may add a new condition to CC:

(v) if C_i is in $\{C_i\}$ then it is not the case that if an event has C then it has C_i or if an event has C_i then it has C.

The conditionals within the scope of "it is not the case" are not the material conditional; the condition is designed to exclude properties that are very strongly correlated, as for instance by a supervenience relation.

Cartwright's condition (iv) is quite similar: if C's cause D's and D's also cause E's, we should not include D as one of the background factors:

The test situations should not hold fixed factors in the causal chain from C to E. If it [sic] did so, the probabilities in the populations where all the necessary intermediate steps occur would be misleadingly high; and where they do not occur, misleadingly low. (p.30)

Finally, the proposed additional condition also rules supervenient properties as causally relevant. A similar "pathology" in the probabilistic relations occurs with supervenient properties. The difficulty is quasi-logical, since supervenience is not

an empirical relation. Hence we have good philosophical reason to arrange CC to allow for causal relevance of supervenient properties.

I conclude that the causal relevance of both supervenient properties and relatively relational properties is different in some logical respects from the causal relevance of properties that supervene on no others and non-relational properties, but that this difference is no reason to think that they are not causally relevant at all.³²

³²I have been helped by discussions of various parts of this chapter with Donald Davidson, Sally Haslanger, Noa Latham, Elisabeth Lloyd, Dugald Owen, Karen Pilkington, Greg Ray and Bruce Vermazen. I owe a special debt to Kirk Ludwig, for countless exchanges on externalism, causal relevance and conceptual connections.

Bibliography

- Adams, E. and Rosenkrantz, R. (1980). Applying the Jeffrey decision model to rational betting and information acquisition. *Theory and Decision*, 12:1–20.
- Anscombe, G. (1971). Causality and determination. In Sosa, E. and Tooley, M., editors, *Causation*, pages 88–104. Oxford University Press, Oxford.
- Armstrong, D. M. (1983). *What is a Law of Nature?* Cambridge University Press, Cambridge.
- Bilgrami, A. (1989). Realism without internalism: A critique of searle on intentionality. *The Journal of Philosophy*, 86(2).
- Block, N. (1980). What is functionalism? In Block, N., editor, *Readings in Philosophy of Psychology*, volume 1, pages 171–184. Harvard University Press, Cambridge, MA.
- Burge, T. (1979). Individualism and the mental. In Peter A. French, Theodore E. Uehling, Jr., and Howard K. Wettstein, editors, *Studies in Metaphysics*, volume 4 of *Midwest Studies in Philosophy*, pages 73–121. University of Minnesota Press, Minneapolis.
- Cartwright, N. (1983). Causal laws and effective strategies. In *How the Laws of Physics Lie*, pages 21–43. Oxford University Press, Oxford.
- Churchland, P. (1979). *Scientific Realism and the Plasticity of Mind*. Cambridge University Press, Cambridge.
- Davidson, D. (1963). Actions, reasons, and causes. In Davidson (1980b), pages 3–19.
- Davidson, D. (1970). Mental events. In Davidson (1980b), pages 240–260.

- Davidson, D. (1980a). Emeroses by other names. In Davidson (1980b).
- Davidson, D. (1980b). *Essays on Actions and Events*. Oxford University Press, New York.
- Davidson, D. (1980c). Freedom to act. In Davidson (1980b).
- Davidson, D. (1980d). Hempel on explaining action. In Davidson (1980b).
- Davidson, D. (1980e). Hume's cognitive theory of pride. In Davidson (1980b).
- Davidson, D. (1980f). The material mind. In Davidson (1980b).
- Davidson, D. (1980g). Psychology as philosophy. In Davidson (1980b).
- Davidson, D. (1982). Paradoxes of irrationality. In Wollheim, R. and Hopkins, J., editors, *Philosophical Essays on Freud*. Cambridge University Press, Cambridge.
- Davidson, D. (1984a). First person authority. *Dialectica*, 38(2–3):101–111.
- Davidson, D. (1984b). *Inquiries into Truth and Interpretation*. Clarendon Press, Oxford.
- Davidson, D. (1984c). Thought and talk. In Davidson (1984b), pages 155–170.
- Davidson, D. (1985a). A new basis for decision theory. *Theory and Decision*, 18:87–98.
- Davidson, D. (1985b). Rational animals. In LePore and McLaughlin (1985), pages 473–480.
- Davidson, D. (1985c). Reply to Patrick Suppes. In Vermazen, B. and Hintikka, M., editors, *Essays on Davidson: Actions and Events*. Clarendon Press, Oxford.
- Davidson, D. (1986). A nice derangement of epitaphs. In LePore (1986), pages 433–446.
- Davidson, D. (1987). Knowing one's own mind. *Proceedings and Addresses of the American Philosophical Association*, 60:441–458.
- Davidson, D. (1988). Reply to Burge. *The Journal of Philosophy*, 85(11).

- Davis, W. (1988). Probability and causality. In Fetzer, J., editor, *Probabilistic Theories of Causation*. D.Reidel, Oxford.
- Dennett, D. (1987a). Evolution, error and intentionality. In Dennett (1987b).
- Dennett, D. (1987b). *The Intentional Stance*. The MIT Press, Cambridge, MA.
- Dennett, D. (1987c). Midterm examination. In Dennett (1987b).
- Donnellan, K. (1966). Reference and definite descriptions. *The Philosophical Review*, 75(2):281–304.
- Dretske, F. (1981). *Knowledge and the Flow of Information*. MIT Press, Cambridge, MA.
- Dretske, F. (1983). Précis of knowledge and the flow of information. *The Behavioral and Brain Sciences*, 6:55–63.
- Dretske, F. (1986). Misrepresentation. In Bogdan, R., editor, *Belief: Form, Content and Function*. Clarendon Press, Oxford.
- Dretske, F. (1988). *Explaining Behavior: Reasons in a World of Causes*. MIT Press, Cambridge.
- Fodor, J. (1986). Banish disContent. In Butterfield, J., editor, *Language, Mind and Logic*. Cambridge University Press, Cambridge.
- Fodor, J. (1989). Information and representation. In Hanson, P., editor, *Information, Language, and Cognition*. British Columbia University Press, Vancouver.
- Fodor, J. A. (1987). *Psychosemantics*. The MIT Press, Cambridge, MA.
- Forbes, G. (1989). Biosemantics. In Tomberlin, J., editor, *Philosophy of Mind and Action Theory*, Philosophical Perspectives. Ridgeview Publishing Company, Atascadero.
- Giere, R. (1980). Causal systems and statistical hypotheses. In L.Jonathan Cohen and Mary Hesse, editors, *Applications of Inductive Logic*, pages 251–270. Clarendon Press, Oxford.
- Goodman, N. (1954). *Fact, Fiction and Forecast*. Harvard University Press, Cambridge, MA, 4 edition.

- Gould, S. J. and Lewontin, R. C. (1979). The spandrels of San Marco and the panglossian paradigm: a critique of the adaptationist programme. *Proceedings of the Royal Society of London, Series B*, 205(1161):147–164.
- Grice, P. (1957). Meaning. *The Philosophical Review*, 66(3):377–388.
- Haugeland, J. (1982). Weak supervenience. *American Philosophical Quarterly*, 19(1):93–101.
- Hempel, C. (1974). Reasons and covering laws in historical explanation. In Gardiner, P., editor, *The Philosophy of History*. Oxford University Press, Oxford.
- Hornsby, J. (1982). Review of *Essays on Actions and Events*. *Ratio*, 24(1).
- Jeffrey, R. (1983). *The Logic of Decision*. The University of Chicago Press, Chicago. Second edition.
- Johnston, M. (1985). Why having a mind matters. In LePore and McLaughlin (1985).
- Kahneman, D., Slovic, P., and Tversky, A., editors (1982). *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge.
- Kahneman, D. and Tversky, A. (1982). Belief in the law of small numbers. In Kahneman et al. (1982).
- Kim, J. (1973). Causes and counterfactuals. *Journal of Philosophy*, 70:570–72.
- Kim, J. (1983). Epiphenomenal and supervenient causation. In French, Uehling, and Wettstein, editors, *Midwest Studies in Philosophy*, volume 9, pages 257–270. University of Minnesota Press, Minneapolis.
- Kim, J. (1985). Psychophysical laws. In LePore and McLaughlin (1985), pages 369–386.
- Kripke, S. (1972). Naming and necessity. In Gilbert Harman and Donald Davidson, editors, *Semantics of Natural Language*, pages 253–355. Kluwer.
- Latham, N. (1987). Singular causal statements and strict deterministic laws. *Pacific Philosophical Quarterly*, 68(1):29–43.
- Lehrer, K. (1989). Metamental ascent: Beyond belief and desire. *Proceedings and Addresses of the American Philosophical Association*, 63(3):19–30.

- LePore, E., editor (1986). *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*. Basil Blackwell, Oxford.
- Lepore, E. and Loewer, B. (1987). Mind matters. *The Journal of Philosophy*, 84(11):630–642.
- LePore, E. and McLaughlin, B., editors (1985). *Actions and Events: Perspectives on the Philosophy of Donald Davidson*. Basil Blackwell, Oxford.
- Loar, B. (1981). *Mind and Meaning*. Cambridge University Press, Cambridge.
- Loewer, B. (1987). From information to intentionality. *Synthese*, 70(2).
- Matthen, M. (1988). Biological functions and perceptual content. *The Journal of Philosophy*, 85(1).
- McDowell, J. (1985). Functionalism and anomalous monism. In LePore and McLaughlin (1985).
- McDowell, J. (1986). Singular thought and the extent of inner space. In Pettit, P. and McDowell, J., editors, *Subject, Thought and Context*. Clarendon Press, Oxford.
- McLaughlin, B. (1985). Anomalous monism and the irreducibility of the mental. In LePore and McLaughlin (1985).
- Millikan, R. G. (1984). *Language, Truth and Other Biological Categories*. The MIT Press, Cambridge, MA.
- Millikan, R. G. (1986). Thoughts without laws; cognitive science with content. *The Philosophical Review*, 95(1):47–80.
- Millikan, R. G. (1989a). Biosemantics. *The Journal of Philosophy*, 86(6):281–97.
- Millikan, R. G. (1989b). In defense of proper functions. *Philosophy of Science*, 56(2):288–302.
- Putnam, H. (1975). The meaning of ‘meaning’. In *Mind, Language and Reality: Philosophical Papers*, pages 215–271. Cambridge University Press.
- Putnam, H. (1983). Probability and the mental. In D.P.Chattopadhyaya, editor, *Jadavpur Studies in Philosophy 5: Humans, Meanings and Existences*, pages 161–173. Macmillan India Limited, New Delhi.

- Putnam, H. (1986). Information and the mental. In LePore (1986).
- Putnam, H. (1988). *Representation and Reality*. The MIT Press, Cambridge, MA.
- Salmon, N. (1986). *Frege's Puzzle*. The MIT Press, Cambridge, MA.
- Searle, J. (1983). *Intentionality*. Cambridge University Press, Cambridge.
- Searle, J. (1984). Intentionality and its place in nature. *Dialectica*, 38(2–3).
- Skyrms, B. (1980). *Causal Necessity*. Yale University Press, New Haven.
- Sosa, E. (1984). Mind-body interaction and supervenient causation. *Midwest Studies in Philosophy*, 9:271–82.
- Stampe, D. (1979). Toward a causal theory of linguistic representation. In Peter A. French, Theodore E. Uehling, Jr., and Howard K. Wettstein, editors, *Contemporary Perspectives in the Philosophy of Language*, pages 81–102. University of Minnesota Press, Minneapolis.
- Sternberg, S. (1969). Memory-scanning: Mental processes revealed by reaction-time experiments. *Scientific American*, 57:421–57.
- Stich, S. (1983). *From Folk Psychology to Cognitive Science: The Case Against Belief*. The MIT Press, Cambridge, MA.
- Suppes, P. (1970). *A Probabilistic Theory of Causation*. North Holland Publishing Company, Amsterdam.
- Suppes, Patrick and Zinnes, Joseph (1963). Basic measurement theory. In R. Duncan Luce, Robert R. Bush, and Eugene Galanter, editors, *Handbook of Mathematical Psychology*, volume I, pages 1–76. Wiley and Sons, New York.
- Wilson, M. D. (1978). *Descartes*. Routledge and Kegan Paul, London.
- Wright, L. (1976). *Teleological Explanations*. University of California Press, Berkeley.