Three Views of Language & the Mind

Jeffrey Speaks

Submitted May 16, 2003

A dissertation
presented to the faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for acceptance by the
Department of Philosophy

November 2003

Abstract. The essay which follows is about the relationship between mind and language. Most recent thought about intentionality has it that (i) mental states of individuals are largely, or in the most fundamental cases, independent of social facts about public languages, and (ii) these social facts are derived from, or constituted by, the mental states of individuals. The purpose of this essay is to challenge this individualist orthodoxy (as well as the view of the relationship between mind and action which often accompanies it), and suggest in its place a communitarian picture of intentionality which gives public languages a role to play in the constitution of thought.

Preface

This essay is an attempt to give at least a partial answer to a traditional question about the relative priorities of mind and language: Is language prior to thought, or is thought prior to language?

I defend the claim that the truth lies closer to the first of these alternatives than is usually supposed. But, as will become clear quickly in what follows, this simple-sounding traditional question breaks down into a number of different questions about the relationship between specific linguistic and mental phenomena. Because the issues involved with some of these may often seem far removed from our initial question, it will be useful to begin with a brief overview of the structure of what follows.

As the title indicates, this essay is concerned with three different ways of answering this question; as you might expect, I argue against two, and defend the third. The two views I criticize are both ways of substantiating what I call the thesis of *the priority of the mental*, or, equivalently, *the priority of the individual*. To substantiate this thesis is to defend a claim about language, and a claim about the mental.

The thesis about language is that the meaningfulness of public language expressions is derived from more fundamental facts about the mental states of individuals who use the language. This thesis may take one of two forms:

> [1*a*] Expressions in public languages have only a derivative kind of intentionality. In particular, the meanings of such expressions are constituted by the propositional attitudes — the beliefs and intentions — of speakers of the language.
>
> [1*b*] Expressions in public languages have only a derivative kind of intentionality. In particular, the meanings of such expressions are constituted by the meanings of expressions in private languages — in, that is, the idiolects or languages of thought of speakers of the language.

The thesis about the mental is that the contentfulness of the mental states of individuals is not to be explained by facts about shared, public languages; rather, mental states are constituted in the first instance by facts about the individuals who have those thoughts. Again, this thesis can take two forms:

[2*a*] Facts about the beliefs and thoughts of agents are prior to and independent of facts about the public languages spoken by those agents. In particular, we can say what it is for an agent to have a certain belief in terms of certain properties of internal states of that agent, which do not involve any facts about public languages.

[2*b*] Facts about the beliefs and thoughts of agents are prior to and independent of facts about the public languages spoken by those agents. They are, however, derived from facts about the private languages of such agents. We can explain what it is for an agent to have a certain belief in terms of the contents of representations in that agent's idiolect or language of thought, and can explain what it is for a representation in such a language to have such a content in terms which do not invoke any facts about meaning in shared, public languages.

Each of these theses has, in one form or another, enjoyed wide support in recent analytic philosophy. The two views of language and the mind that I argue against in the first two parts of this essay are the *mentalist* picture of intentionality, which consists centrally of the *a*-theses above, and the *private language* picture of intentionality, which consists of the *b*-theses above.

The structure of the essay is very simple. After Chapter 1, which explains some key methodological notions used in stating foundational claims about intentionality, I turn in Part I to critical examination of the mentalist picture of intentionality. Chapters 2 and 3 are devoted to showing the falsity of [1*a*], and Chapter 4 is targeted against [2*a*].

Many of the problems faced by mentalism can be seen as motivations for the private language picture of intentionality, with is another way of spelling out the thesis of the priority of the individual. Accordingly, in Part II I turn to discussion of this view, arguing in Chapter 5 that thesis [2*b*] is false, and in Chapter 6 that thesis [1*b*] is false.

This opens the way to a defense of a *communitarian* picture of intentionality, which I present in Part III of the essay. Communitarianism is partly defined by the negations of the four theses above. It holds that many of the mental states of agents are constituted by facts about the public languages they speak. In Chapter 7, I show how the failure of individualism motivates such a view, and present a positive account of one such mental state — belief — in Chapter 8. Giving an account of the beliefs of agents partly in terms of the meanings of expressions of their public languages raises the question of what constitutes public language meaning. As it turns out, this question is only one of a number of related questions which threaten the communitarian position with circularity. I turn to discussion of these problems, some possible responses, and some of the philosophical consequences of communitarianism in Chapter 9. Part III leaves a number of important questions unanswered; in particular, I ignore the role of intentions in communitarianism. It is intended not to provide a full foundational account of intentionality, but rather to do no more than to show the possibility of a communitarian alternative to more well-worked out individualist views of thought and language.

These issues have been much discussed in analytic philosophy in the last half-century; accordingly, many of the issues have become quite technical and complex. Sometimes it has

been impossible to avoid delving into technicalities in the main text; but, when possible, I have consigned discussion of complicated but relatively peripheral issues to appendices. Footnotes in the main text indicate when discussion in one of the appendices is relevant.

. . .

This essay has benefited from conversations with and comments from a number of people in Princeton's Department of Philosophy. I cannot single out the many conversations with fellow graduate students over the last few years which have, in a variety of ways, changed the shape of this dissertation; but I can say with certainty that it would have been much worse without their help. Special thanks are due to the members of the Dissertation Seminar of the last three years, and especially to Antony Eagle, who read and commented on large parts of the penultimate version of the essay.

It is is similarly difficult to recall all of the individual instances in which members of the faculty contributed to this dissertation; but special thanks are owed to Paul Benacerraf, Sean Kelly, Mark Greenberg, Jim Pryor, Gideon Rosen, and Scott Soames. In particular, Mark introduced me to many of the topics in the philosophy of mind and language with which this essay is concerned in his seminar on mental content in the fall of 2000. I am sure that my approach to these problems, in more ways than I am aware, bears the stamp of Mark's thinking. In the last several years he has read several drafts of various parts of the dissertation, and invariably has responded with helpful comments on both very general issues in the philosophy of mind and technical details of the relevant piece. Without his help, my understanding of the issues discussed in what follows would have been greatly impoverished.

The person to whom I owe the greatest academic debt is my advisor, Scott Soames. In many places in the dissertation, I note that certain points or formulations are due to the advice of one or another philosopher; I have not followed this practice with respect to Scott's comments, simply because such an acknowledgement would appear on virtually every page. Scott has read every page of what follows, some of them many times, and with each reading uncovered new objections to and implications of my work. In more than one case, whole sections of chapters emerged from Scott's comments. I do not think that it is an overstatement to say that Scott taught me how to do philosophy in a rigorous and systematic way; for this, and much else, I am greatly in his debt.

Of course, anyone who has written a dissertation has personal as well as academic debts to acknowledge. In my case, three stand out: my friends at Princeton, to whom I owe a very happy five years; my parents, to whom I owe ceaseless support and encouragement; and, most of all, Elyse Deeb, to whom I owe more than I can say.

Table of Contents

Chapter 1

Mentalism, Private Languages, & Communitarianism

Contents

1.1 Introduction

In our talk of meaning and content, we ascribe content both to signs — expressions of languages, utterances, gestures — and to the mental states of agents — their beliefs, intentions, desires, and so forth. One fundamental question about language is what it is for a sign to have a certain meaning; one fundamental question about the mind is what it is for an agent to be in a given mental state with a certain content. It is natural to think that these two questions bear some relation to each other; our ability to have contentful thoughts of certain sorts evidently has something to do with our ability to speak a meaningful language. This essay is about these two questions, and about the ways in which an answer to one might figure in an answer to the other.

Refinements aside, something like the following picture of the relationship between mind and language is widely accepted. Mental representation is prior to representation in public languages. What it is for an agent to be in a mental state with a certain content is for that agent to be in a physical state — likely, a brain state — which exhibits certain key second-order properties. By contrast, what it is for an expression in a public language like English to have a certain meaning is for speakers of that language to be in certain mental states which, in a way to be specified, fix the content of the expression. So, on this view, the most fundamental questions about intentionality arise at the level of thought, and are to be answered by finding the properties of brain states constitutive of agents being in the mental states under investigation. Once the contents of mental states are determined in this way, the meanings of expressions of public languages follow in turn.

One key element in this picture of mind and language is a claim which we might call either the priority of the individual or the priority of the mental: the claim that the meaningfulness of expressions of public languages is a kind of derived intentionality, constituted by more fundamental facts about the contents of the thoughts of individual agents. For ease of reference, I shall call views which endorse this claim *individualist* views of mind and language. The central negative thesis of this essay is that individualism is the wrong picture of language and the mind.

In its place, I shall present and defend what I call a *communitarian* view of mind and language. Again putting refinements to the side, the communitarian rejects the thesis of the priority of the mental. According to the kind of communitarianism I defend, the foundations of intentionality are to be found, not in the contents of the thoughts, beliefs, or judgements of agents, but rather in the contents of their perceptions, in the actions they are disposed to perform, and in the meaningfulness of expressions in the linguistic community (if any) of which those agents are members. The contents of the beliefs, desires, and intentions of agents are derived from the contents of their perceptions, along with their dispositions to perform certain kinds of actions; crucially, among the relevant dispositions will be dispositions to perform certain kinds of linguistic acts with meaningful sentences of their language. The communitarian thus takes seriously the claim that public languages can, in a robust sense, serve as a vehicle for the thoughts of agents.

As I have described these two positions, the individualist and the communitarian differ on a number of issues; three among these are particularly important: the relative priorities of linguistic meaning and the content of thought; the relative priorities of individuals and social groups in the explanation of intentionality; and the relative importance of internal states and dispositions to action to understanding the nature of representation. A fourth difference will come to the fore later; for the individualist and the communitarian differ not only in their attitude towards mental and linguistic content, but also in their views of the nature of intentional action.

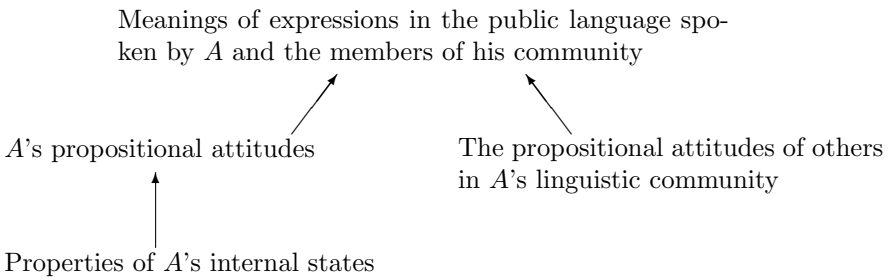
These descriptions paint with a very broad brush; many philosophers would reject both individualism and communitarianism, as stated. The point of giving these descriptions is only to state two very different starting points for thinking about the relationship between mental content and linguistic meaning; as we'll see in what follows, many refinements to each picture are possible. But, even if a fully general characterization of individualism and communitarianism is not easily given, various accounts of intentionality typically fit rather naturally into one or the other camp.[1]

The essay divides into three parts, with the first two devoted to the negative project of arguing against individualism, and the last devoted to the exposition and defense of a version of communitarianism. The discussion of individualism breaks into two parts because there are two different kinds of individualism. According to the first, which I shall call *mentalism*, the fundamental sort of mental representation is to be found in propositional attitudes: for example, the beliefs, thoughts, and intentions of agents. Taking belief as our example, the view is then that what it is for an agent to believe an arbitrary proposition p is for the agent to be in some internal state x which has come property F which qualifies it as the belief p.

[1] A possible exception is a view which does not countenance any sort of relations of explanatory or conceptual priority between facts in this domain; this, and related 'interdependence' views, will come in for more discussion later.

On some views, F will be the conjunction of a property which qualifies x as a belief state, and a separate property which gives it the content p. Some basic propositional attitudes are explained in this way, and others are explained in terms of these. Facts about linguistic meaning are then fixed by some combination of these basic and derived propositional attitudes.[2] Letting 'A' stand for an arbitrary agent, and using arrows to express the relation of one class of facts constituting, or determining, another, this may be represented as follows:

THE MENTALIST PICTURE



As it stands, then, any attempt to fill out the mentalist picture must answer two questions: Which properties of internal states constitute the relevant propositional attitudes of agents?, and Which propositional attitudes of agents in a linguistic community conspire to constitute the meanings of expressions in the language they share?

But mentalism is not the only option open to the individualist; a quite different view of intentionality is provided by the *private language* picture. The proponent of this view will agree with the mentalist that the fundamental kind of representation is mental representation, but, unlike the mentalist, will regard the contents of propositional attitudes as derived from a more fundamental kind of intentionality. On the private language picture, each agent capable of contentful thought possesses a language specific to himself; this private language may either be thought of as a language of thought consisting of a set of mental representations, or as an idiolect or Chomskyan I-language.[3] On this view, the fundamental questions about representation are questions about what it is for an expression or representation in such a private language to have a given content.

Because the private language picture posits an extra level of contentful items not present in the mentalist picture, the private language theorist has more resources than does the mentalist when it comes to the question of what it is for an expression of a public language to have a given meaning. The private language theorist may either take the mentalist course of identifying the source of public language meaning with the propositional attitudes of speakers of the relevant language, or may regard the meanings of words in public languages as constituted directly by the contents of items in the private languages of speakers of the relevant public language.

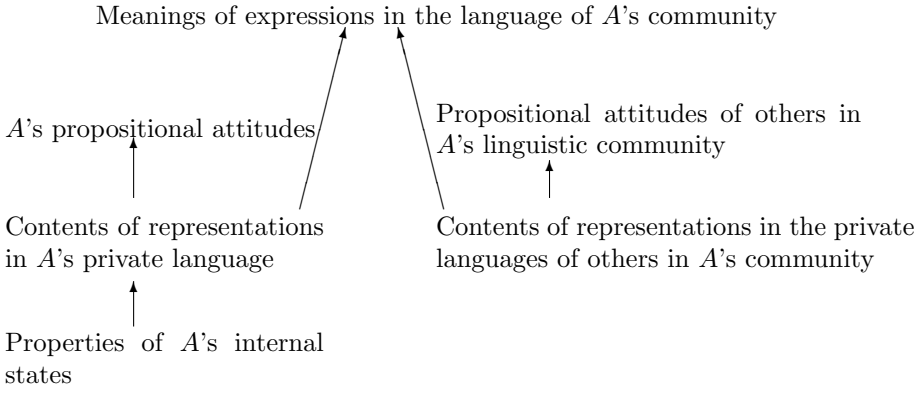
[2]Note that the mentalist need not hold that *all* propositional attitudes may be explained in this way; the key claim is only that linguistic meaning is constituted by some group of those which can.

[3]The difference between these two conceptions of private languages will be discussed in Chapter 5 below. Note that private languages need not be regarded as private in Wittgenstein's sense; that is, they need not be such that they could, in principle, only be spoken by a single agent.

I shall consider a version of the private language picture which takes the latter course, since the attempt to view public languages as constituted by the propositional attitudes of agents will be discussed in the context of the mentalist picture of intentionality.

Given this, the view of intentionality we get from the private language theorist may be represented as follows:

THE PRIVATE LANGUAGE PICTURE



Because she posits an extra class of theoretically relevant facts — contentful representations in the private languages of agents — the proponent of the private language picture has more resources at her disposal than does the mentalist. But this move also means that filling out the private language picture requires answering three rather than two questions: What is it for a representation in a private language to have a given content?, How are the contents of propositional attitudes fixed by the contents of these representations?, and How do the private languages of the members of a linguistic community fix the meanings of expressions the language they share?

Communitarianism is, in the first instance, a rejection of the individualist claim that public language meaning is only a derived kind of intentionality; in its place, the communitarian asserts that the meanings of expressions in public languages play a role in explaining what it is for an agent to be in certain sorts of intentional mental states. But this claim is open to different interpretations on two scores: (i) Is communitarianism the thesis that linguistic meaning is strictly prior to the mental states in question — which seems to require an independent account of linguistic meaning which does not advert to those mental states — or an assertion of the interdependence of public languages and those mental states? (ii) How wide is the class of mental states for which the communitarian claim holds? Intuitively, the communitarian claim sounds plausible for propositional attitudes like assertion, contentious for attitudes like belief, and implausible for mental states associated with perception and bodily sensation. Depending on which direction the communitarian goes on these two issues, her explanatory commitments will correspondingly differ. But any communitarian owes some story about how linguistic meaning plays a role in a constitutive account of some central mental states; in Part III, I will defend such an account of belief.

Each of these three pictures of intentionality raise a number of problems. But before

going on to discuss these, it's worth pausing to discuss some issues common to all three. In particular, each is stated as a view about a certain kind of relation holding between certain classes of facts. One might wonder, first, what this relation is. Above, I've loosely characterized this relation by talking about one class of facts *constituting* another, about what a class of facts *consists in*, and about *what it is* for a certain fact to obtain; but these idioms are far from clear. Further, one might wonder what sorts of things *facts* about the meanings of expressions and the contents of thoughts are; that is, one might wonder what questions about what constitutes different sorts of intentional facts presuppose about the metaphysics of intentionality. The next two sections are devoted to clearing up these two issues.

1.2 Constitutive claims

There are a host of idioms which have been used to express the claim that one class of facts about meaning or content is derivative from, or constituted by, another. Searle writes of what has "content in the first instance" versus what has "derived intentionality;"[4] Fodor of which is "first *in order of explanation*";[5] Schiffer of whether we can give an account of *what it is* for an expression to have meaning in terms of other facts about meaning and content.[6] None of these ways of putting the issue is altogether clear.[7] The aim of this section is to arrive at an interpretation of these notions which gives us a better understanding of what it is to say that one class of facts about meaning and content is constitutive of another. I shall call these *constitutive claims*.

The distinction between saying which facts of a kind obtain, and saying what it is for a fact of that kind to obtain

A first step is made by distinguishing claims about *which* facts about meaning and content obtain from claims about *what it is* for a certain sort of fact about meaning or content to obtain, a distinction often not made in discussions of 'theories of meaning.'[8] It should be clear, however, that there is a distinction between two different sorts of theories of linguistic meaning: semantic theories, which tell us which facts about linguistic meaning obtain, and foundational,

[4]Searle (1983), (1998).

[5]Fodor (2001), 2.

[6]Schiffer (1982).

[7]For example, the metaphor of constitution sometimes used to frame these claims is clearly only metaphorical: there are no facts which stand to intentional facts as the particles which constitute the desk on which I'm writing stand to the desk. The same can be said of the other expressions used to ask foundational questions; for example, when we ask *what it is* for some intentional fact to obtain, this cannot be construed literally as the requirement that the proposition given in answer to the foundational question be *identical* to the intentional fact in question. If this were required, then, since in most cases there will be no sentence in the language which expresses this proposition other than the meaning ascription itself or a trivial rewording, we could answer foundational questions only by simply repeating the intentional fact whose nature we were out to understand.

[8]See, e.g., Lycan (1999) for a possible example of the two sorts of theories being conflated. He treats as competitors, for example, possible worlds semantics — which is a framework for constructing semantic theories — and Gricean theories of meaning, which make claims about what constitutes an expression having a given meaning. But one could endorse both without inconsistency; Brian Loar, for example, does in his (1976).

or constitutive, theories which make claims about what it is for a linguistic expression to have a certain meaning. In the case of facts about mental content, the distinction between the two sorts of accounts is too obvious to need much mention: no one would confuse the project of saying what the beliefs of an agent are with the project of saying what it is for an agent to have a belief with a given content. The two sorts of theories are largely independent of each other; two theorists could, for example, endorse the same semantics for a language, but endorse very different views about the facts in virtue of which expressions have the meanings they do. Pretty clearly, because the priority of the mental is a thesis about the relationship between facts about linguistic meaning and facts about mental content, an evaluation of the thesis of the priority of the mental requires a constitutive, rather than a semantic, theory of linguistic meaning.

Constitutive claims as stating necessary and sufficient conditions

This tells us something about what constitutive claims about meaning and content are not; can we say anything about what they are? A natural and traditional starting point is to require that constitutive claims be substantiated by *analyses*. That is, according to the traditional view, if a class A of facts about meaning or content is constituted by and derived from a class B of such facts, then we must be able to give metaphysically necessary and sufficient conditions for the A-facts obtaining in terms of the B-facts.[9] There is some reason to think that this constraint on constitutive claims is in some ways too strong, and in others too weak; before considering these, though, it is worth rehearsing the intuitive motivation for the traditional claim that theories about what it is for an expression to have a given meaning should provide necessary and sufficient conditions for an expression to have a certain meaning.

This is best done by example. Consider the Gricean claim that facts about what speakers mean by their utterances are constituted by, and wholly derivative from, facts about their intentions. The following simplified version of this claim will suit our purposes: what it is for a speaker to mean p by uttering x is for that speaker to intend in uttering x that her audience believe p on the basis of recognition of this intention. Suppose first that we cannot give *necessary* conditions for facts about speaker-meaning in terms of facts about the intentions of speakers. In this case, there would be at least one fact about speaker-meaning for which there is no corresponding fact about the intentions of speakers; that is, there is at least one possible occasion on which a speaker means p by uttering x and yet does not intend in uttering x that his audience come to believe p on the basis of recognition of this intention. Intuitively, this is enough to show that facts about audience-directed intentions are only part of the story; in this case it seems that, at best, one could make the claim that *some* facts about speaker-meaning are constituted by the intentions of speakers.

Suppose now that we cannot give *sufficient* conditions for facts about speaker-meaning in terms of facts about the intentions of speakers. In this case, there would be facts about speaker intentions of the sort claimed to be constitutive of speaker-meaning in the absence of any facts about speaker-meaning; that is, there would be at least one possible instance in which a speaker intended in uttering x that his audience come to believe p on the basis

[9]This view is shared by many theorists, although often it is an implicit working hypothesis rather than an explicit methodological claim. See, e.g., Schiffer (1972), and the various writings of Grice, in which this requirement is more visible than usual.

of recognition of this intention but in which she did not mean p by uttering x. Again, in this case it seems intuitively right to say that such an example is enough to show that the claim that facts about speaker-meaning are entirely constituted by facts about these audience-directed intentions is false. If a certain kind of fact about the intentions of speakers can occur either with or without the requisite fact about speaker-meaning, then, if facts about speaker's intentions are to play some role in saying what constitutes a speaker meaning something by an utterance, it must be these intentions *along with something else* — which is present in cases of speaker-meaning but absent in the cases where we have the intentions but no speaker-meaning — which constitutes a speaker meaning something by an utterance.

This traditional constraint on constitutive claims can be put a bit more precisely as follows: if a class A of facts about meaning and content is constituted by class B of facts about meaning and content, then there is a true formula of the form

$$\Box\forall x_1\forall x_2 \ldots \forall x_n \ (\mathrm{F}(x_1, x_2 \ldots x_n) \equiv \mathrm{G}(x_1, x_2 \ldots x_n)),$$

instances of which are obtained by assigning n a value and replacing 'F' and 'G' with n-place predicates, which has the following characteristics: (i) all and only the facts in class A satisfy the left-hand side of the biconditional, and (ii) all facts which satisfy the right-hand side of the biconditional are in class B.[10] In order to make the analyses which we'll consider below (slightly) more readable, in what follows I shall often omit quantifiers and the necessity operator when discussing some proposed analysis of a class of facts. Throughout the text, displayed biconditionals should be understood as necessitated biconditionals, in which all variables not explicitly bound by quantifiers are understood to be universally quantified and to have wide scope relative to every operator in the formula other than the necessity operator. For example, the Gricean claim that

> $\Box\forall a\forall x\forall p$ (a means p by uttering $x \equiv a$ intends by uttering x that his audience believe p on the basis of recognition of that intention)

will be written simply as

> a means p by uttering $x \equiv a$ intends by uttering x that his audience believe p on the basis of recognition of that intention.

Claims throughout to the effect that there is an analysis or account or explication of one class of facts in terms of another should be understood accordingly.

By itself, this constraint on constitutive claims is clearly too weak; but there are also some reasons for thinking that it is too strong. One thought along these lines is that we should be looking, not for necessary and sufficient conditions, but only for sufficient conditions: facts on which the class of facts about meaning or content in question supervene. In much contemporary work on 'naturalizing semantics,' the provision of such a supervenience base is the implicit goal.[11]

[10] Because A and B will include possible as well as actual facts about meaning and content, these facts should be thought of as indexed to possible worlds. (The necessity operator in this schema represents metaphysical necessity.)

[11] For an explicit discussion — though no argument for this weakening of the traditional constraint — see Fodor (1990c), 96-7.

This weakening of the traditional constraint on constitutive claims is, I think, a mistake. To see why, it is important to distinguish between the sorts of foundational questions we've been discussing — questions about *what it is* for an expression to have a given meaning, or for an agent to have a given belief — from the question of whether the existence of facts about meaning and content is consistent with physicalism. If the best statement of physicalism is in terms of the supervenience of non-physical facts on physical facts rather than in terms of the reducibility of the non-physical to the physical, then, trivially, all that is required to show the consistency of belief in intentional facts with physicalism is the provision of a physical supervenience base for intentional facts.

But if one is interested in what constitutes having a belief, or what it is for an expression to have a meaning, it is hard to see why sufficient conditions should be more important or revealing than necessary conditions. Since I know of no argument why they should be — and, as mentioned above, there is an intuitive case for the traditional constraint — I propose to stick with the traditional constraint for the time being.

The problem posed by 'family resemblance'

A second, and more powerful reason for thinking that the traditional constraint is too strong derives from Wittgenstein's remarks on "family resemblance" concepts. In general, we neither come to understand words by matching them with necessary and sufficient conditions for their application, nor require for competence that users of a term be able to provide such conditions. ("When I give the description: 'The ground was quite covered with plants' – do you want to say I don't know what I am talking about until I can give a definition of a plant?"[12]) This is certainly the case for the concepts with which we're concerned: meaning, belief, intention, and so on. So why think that there *is* any true analysis of the class of facts about belief, or about linguistic meaning? Isn't the imposition of the traditional constraint just a recipe for failure?

This doubt is reinforced by consideration of the history of attempts to meet the traditional constraint. Stephen Stich has described the situation well:

> The rules of the game have changed very little over the past 2500 years. It goes something like this:
>
> S: (Socrates, as it might be): Tell me please, what is X? (where "X" may be replaced by "justice" or "piety" or "knowledge" or "causation" or "freedom"...)
>
> C: (Cephalus, perhaps, or Chisholm): I will tell you gladly. To be an instance of X, something must be y and z.
>
> S: But that can't be right. For surely you will grant that a is X, but it is neither y nor z.
>
> C: You are quite right. Let me try again. To be an instance of X something must be either y and z or it must be w.
>
> S: I'm afraid that won't work either, since b is w, but clearly it is not X.

[12] *Philosophical Investigations*, §70.

> The game comes to an end when S runs out of counterexamples, or C runs out of definitions. And, though no one has kept careful records in this sport, the smart money usually bets on S.[13]

These doubts present something of a quandary. On the one hand, nothing in our practices of using these concepts seems to indicate that we should be able to give true analyses of them, and this is reinforced by the relative absence of successful analyses from the history of philosophy. On the other hand, if we have to reject the traditional constraint altogether, it is difficult to see how we could even go about evaluating answers to foundational questions.

Though his solution to this quandary is far from precise, I think that Kripke's remarks on this topic in *Naming and Necessity* point the way to an improvement on the traditional constraint. In Lecture II, after giving his objections to the view that the referents of names are fixed by descriptions associated with those names by speakers and then giving his alternative picture of the way in which the reference of a proper name is fixed, Kripke writes

> . . . rather than giving a set of necessary and sufficient conditions which will work for a term like reference, I want to present just a *better picture* than the picture presented by the received views.
>
> Haven't I been very unfair to the description theory? Here I have stated it very precisely — more precisely, perhaps, than it has been stated by any of its advocates. So then it's easy to refute. Maybe if I tried to state mine with sufficient precision in the form of six or seven or eight theses, it would also turn out that when you examine the theses one by one, they will all be false. That might even be so, but the difference is this. What I think the examples I've given show is not simply that there's some technical error here or some mistake there, but that the whole picture given by this theory of how reference is determined seems to be wrong from the fundamentals.[14]

One moral of these remarks is that constitutive claims ought to be viewed as presenting, in Kripke's sense, *pictures* of the nature of the class of facts in question. In order to argue against such claims, then, we should require more than one or two recondite counterexamples; rather, an interesting argument against a view about, for example, what it is to have a belief should contain, in addition to counterexamples to the view, an argument that those counterexamples show the picture of belief embodied in the view to be fundamentally mistaken.

Note that this does not mean that constitutive claims should be presented in a vague way; on the contrary, in what follows I shall present claims about the nature of meaning and various propositional attitudes as attempts to provide strict necessary and sufficient conditions. Rather, it means that we should require more of attempted refutations of these constitutive claims than a few counterexamples.

The notion of priority

Whether or not the traditional constraint is too strong, it is clear that, when taken by itself, it is too weak to give us truth conditions for constitutive claims. For consider the following putative account of what it is for one thing to be to the east of another:

[13] Stich (1994), 351.
[14] Kripke (1972), 93.

x is to the east of $y \equiv y$ is to the west of x

This seems to satisfy the traditional constraint; it seems to give metaphysically necessary and sufficient conditions for one thing to be to the east of another. But, pretty obviously, this is not a very informative story about what constitutes the relational property of one thing's being to the east of another; and the reason why is not far to seek. When we say that a class A of facts is constituted by a class B of facts, then part of what we are saying is that the B-facts must be, in some sense of conceptual or explanatory priority, prior to the A-facts. If *what it is* for a member of class A to obtain is for a certain member of class B to obtain, then there must be some such asymmetry between the two facts. Since necessary equivalence is a symmetric relation, what I've been calling the traditional constraint on constitutive claims needs to be supplemented by another constraint. The difficulty is in getting clear about what this extra constraint of conceptual or explanatory priority amounts to.

A natural thought is that, while not all analyses are sufficient to establish the priority of one class of facts over another, some analyses, in virtue of satisfying certain conditions, do have this property. One way to find out what these conditions are is to look at attempted philosophical analyses which have been thought to show, if true, the intuitive priority of one class of facts about meaning and content over another, and to try to see what characteristics of these analyses have led to this response.

One such analysis is the neo-Gricean analysis of facts about the meanings of sentences in a language in terms of facts about the audience-directed intentions of speakers of the language. This analysis has two stages: (i) an analysis of facts about the meanings of expressions of a language in terms of facts about conventions governing what speakers of the language mean by using sentences of the language, and (ii) the analysis, mentioned above, of facts about speaker-meaning in terms of facts about certain of the intentions of those speakers. One reason why this analysis has been taken to show that facts about intentions constitute, and are prior to, facts about the meanings of expressions in public languages is that, given that non-linguistic animals can have intentions, there is no analysis which runs in the opposite direction: there is no true analysis of facts about the intentions of agents in terms of facts about the meanings of expressions in the public languages spoken by those agents.[15] If this diagnosis of this response to the neo-Gricean analysis is right, then this provides some justification for the claim that true analyses are priority-revealing just in case there is no true analysis running in the opposite direction.[16]

Two points show that this can't be quite right. First, there is a kind of arbitrariness in the way that we demarcate the sets of facts involved. In the case of the Gricean analysis, do we include in the constitutive class of facts all facts about the intentions of agents, or just the sub-class of audience-directed intentions? Surely, it shouldn't matter; if the Gricean analysis is true and if facts about intentions are prior to and constitutive of facts about linguistic meaning, then the particular intentions used in the analysis — the audience-directed intentions discussed above — are prior to and constitutive of linguistic meaning as well. But,

[15]By "non-linguistic animals", I mean animals which use no public language. Strictly, the premise in the text about these animals is stronger than is needed: all that is needed is the claim that it is metaphysically possible that there be a non-linguistic animal which has intentions.

[16]This understanding of 'logically prior' is suggested in Neale (1992), 543. This also seems to have been the working hypothesis in the correspondence between Chisholm and Sellars on the nature of intentionality in the 1950's. See Chisholm & Sellars (1958), especially 215 ff.

according to this proposal, this is not so. For if the Gricean analysis is true, then there *is* an analysis running in the opposite direction: there is an analysis of facts about audience-directed intentions in terms of facts about speaker-meaning.[17]

Second, it may be that the best account of what it is for a B-fact to obtain will involve one or more A-facts. Even if there is no analysis of the B-facts in terms of the A-facts alone, there may be an account of the B-facts in terms of the A-facts along with another class C of facts. In this case we would have a kind of explanatory circularity, and hence no grounds for regarding the B-facts as in any way prior or constitutive of the A-facts.

Both of these problems are solved by the following view of constitutive claims: a class B of facts is constitutive of and prior to a class A of facts just in case there is an analysis of the A-facts in terms of the B-facts, and the B-facts are more basic than, and hence not constituted by, the A-facts. But, while this is likely true, it is not a non-circular account of the truth conditions of constitutive claims, since the specification of these truth conditions makes essential use of the fact that a certain constitutive claim is false. But what are the grounds for saying that such a claim is false?

It seems that one ground could be the provision of an account of the B-facts which make no use of the A-facts: if we can say what it is for a B-fact to obtain without mentioning the A-facts, then it seems plausible that the B-facts are not constituted by the A-facts. This suggests, then, that in order to show that a class A of facts is constituted by and derived from a class B of facts, we should give an analysis of the A-facts in terms of the B-facts, and an analysis of the B-facts in terms of a class C of facts which does not overlap with the A-facts.

Now one might still raise the following doubt: "Take this class C of facts claimed to be constitutive of the B-facts; what guarantees do we have that it is not itself to be explicated in terms of the A-facts? For, if it is, then we again have an explanatory circle, and no grounds for regarding the B-facts as explanatorily or conceptually prior to the A-facts. So perhaps we need also require that this class C of facts be explicated by another class D of facts, to avoid this possibility. But this only delays the problem, since the D-facts might equally well be constituted by the A-facts. So either you have to provide an unending string of analyses, or your original claim that the A-facts are constituted by the B-facts is unjustified." Since this requirement of an unending string of analyses clearly cannot be met, we need some reason to

[17] Similar objections show that two related proposals also fail. One might think, first, that the Gricean analyses show, if true, the intuitive priority of intentions over speaker-meaning because they show that there is an analysis of facts about speaker-meaning in terms of a proper sub-class of the facts about intentions. Suppose that we adopt this criterion, and show that a class A of facts is constituted by class B of facts. Then it seems that for any proper subset A' of A with members $a_1 \ldots a_n$, there will be a proper subset B' of B with members $b_1 \ldots b_n$ such that a_1 is constituted by b_1, a_2 by b_2, and so on. In this case, there will be an analysis of A' in terms of B', and of B' in terms of A'. But then, since there is an analysis of B' in terms of a proper sub-class of A, we get the result that class B' of facts is constituted by the A-facts.

Another possibility is to cash out this notion of priority in terms of the asymmetric supervenience of one class of facts on a class of prior, constitutive facts. But again suppose that the Gricean analysis is true, and that facts about certain audience-directed intentions are constitutive of and prior to facts about speaker-meaning. In this case there would be no asymmetric supervenience of the latter on the former, since there would be no facts about those audience-directed intentions in the absence of facts about speaker meaning. So again we would get the result that while intentions in general are constitutive of speaker-meaning, the class of audience-directed intentions used in the analysis are not.

believe that one class of facts is *not* constituted by another without having to give an account of the former in other terms.

If one thinks of concrete examples, though, this isn't really much of a problem. Suppose (yet again) that the Gricean analysis is a true analysis of facts about linguistic meaning in terms of facts about the intentions of agents. It would, I think, be reasonable to object that this analysis may not be priority-revealing: propositional attitudes in general and intentions in particular *might* be partly constituted by facts about linguistic meaning. So, to allay this doubt, the Gricean proponent of the priority of the mental owes an account of what it is for an agent to intend p. Suppose that the Gricean succeeds in giving such an account; say, in terms of causal relations between brain events and objects in the world. Now the anti-Gricean is supposed to object: "Wait! You still can't conclude that facts about the intentions of agents are constitutive of and prior to linguistic meaning, for the facts used to explicate what it is to have an intention might themselves be partly constituted by facts about linguistic meaning." But this objection is really not very forceful; it is just not plausible that causal relations between objects in the world and brain states, for example, are to be accounted for partly in terms of facts about the meanings of expressions in a public language. What this example shows is that in many cases, we can be quite sure that one class of facts is more basic than, and hence not constituted by or derived from another, even without having an independent account of the former.[18]

This is enough to give us a good idea of what, intuitively, is required to defend the thesis that all facts about the meanings of expressions in public languages are derived from more fundamental facts about mental content. First, the proponent of this thesis should give an account of facts about linguistic meaning in terms of some class of facts about meaning and content. Second, she should explicate this other class of facts in terms of a class of facts which include no facts about the meanings of expressions in public languages. There is no requirement that the analysans make use only of terms from the 'language of physics' or anything of the sort; we need require only that the analysans not include facts about linguistic meaning, and that the facts that it does include be such that we can be sure that they are not themselves partly constituted by facts about linguistic meaning.

These requirements that the proponent of the priority of the mental provide various sorts of analyses should, of course, be understood as subject to the proviso regarding family resemblance concepts and the distinction between theories and pictures in philosophy: one holds the the proponent of this thesis to too high a standard if one requires necessary and sufficient conditions which admit of no intuitive counterexamples at all. Rather, I think, it should be enough if he can provide a compelling *picture* of the nature of linguistic meaning.

[18] It's important to keep in mind, I think, that claims about explanatory priority, like claims about explanation more generally, are often best understood as relative to the aspect of the explanans in which one is interested. Here we're interested in the contents of mental states and of linguistic expressions; hence it will usually be sufficient to ensure explanatory priority that the class of facts used to explain some facts about meaning and content not themselves be facts about the contents of mental states or linguistic expressions.

1.3 Foundational questions & the metaphysics of intentionality

This goes some distance toward clarifying what is meant by talk about one class of facts constituting, or being prior in the order of explanation to, another. But, depending on one's attitude toward intentionality, one might find the idea of asking constitutive questions about either the meanings of expressions in languages or the contents of mental states of agents inappropriate. For many philosophers have seemed to want to deny to claims about the meanings of expressions or the contents of mental states the status of genuine facts.

One might think that the kinds of constitutive claims outlined above presuppose an objectionably inflationary view of facts about meaning and content. For consider a claim like the Gricean account of what it is for a speaker to mean something by an utterance, mentioned above:

> $\Box\forall a\forall x\forall p$ (a means p by uttering $x \equiv a$ intends by uttering x that his audience believe p on the basis of recognition of that intention)

In this formula, 'p' functions as an ordinary objectual variable. This seems to carry a commitment to propositions, and, as Davidson puts it, "meanings as entities."[19] This runs counter to one school of thinking about meaning and content, which rejects this sort of 'inflationary' conception of intentional facts. Wittgenstein, for example, seems to have opposed such a view in §120 of the *Investigations*, when he characterized the view of his interlocutor as follows:

> You say: the point isn't the word, but its meaning, and you think of the meaning as a thing of the same kind as the word, though also different from the word. Here the word, there the meaning. The money, and the cow you can buy with it.

Various reasons have been given for rejecting "inflationary" views of meaning, and adopting one sort or other of deflationism about meaning and content.[20] For this reason, it seems worthwhile to show that the sense in which there are facts about meaning and content which is required for questions about what constitutes a class of such facts to be intelligible includes only minimal assumptions about the metaphysical status of meanings.

It seems to me that all we need presuppose on this score is that ascriptions of meaning and content are truth-apt, and sometimes true. The use of quantification over meanings in biconditionals like the above can, I think, be recast in terms of quantification over ascriptions of meaning or content. For example, we might replace the Gricean account of speaker-meaning with its deflationist translation:

> $\Box\forall s$ ((s is an ascription of speaker-meaning)$\rightarrow$(s, as we actually use it, is true $\equiv$ $f(s)$, as we actually use it, is true))

in which 'f' stands for a function from sentences of the form $\ulcorner a$ means p by uttering $x\urcorner$ to sentences of the form $\ulcorner a$ intends by uttering x that his audience believe p on the basis of recognition of that intention$\urcorner$.

[19] Davidson (1967), 20.

[20] See, e.g., Davidson (1967), Kripke (1982), Schiffer (1987), Johnston (1988), Field (1995).

Another route for the deflationist to go is via a reinterpretation of the quantifiers in these formulae. Any deflationist who rejects objectual quantification into the complement clauses of ascriptions of meaning and content will have to find some way to make sense of ordinary language idioms which seem to be instances of just that; it seems plausible that if there are any ways of doing that by interpreting this quantification as substitutional, then this trick should work for formulas like the above as well. In either case, it doesn't seem that evaluation of claims like that above about the relationship between two classes of facts about meaning or content need incorporate any substantial metaphysical assumptions about meanings.

If this is right, then the sense of 'facts about meaning' required for foundational questions about meaning is one that only an error-theorist about meaning and content, who denies that there are any true ascriptions of meaning or content, or a radical non-cognitivist about meaning, who denies that such ascriptions are truth-apt, should find cause to question. Whether we talk about the conditions under which certain ascriptions of meaning or content are true or about the conditions in which certain intentional facts obtain makes little difference when it comes to these sorts of foundational questions. In this sense, the metaphysical question about the nature of meanings is distinct from foundational questions about the nature of facts about meaning and content. So, while I will continue to use the framework of identifying facts about meaning and content with true Russellian propositions, this way of thinking about facts about the meanings of signs and contents of mental states is not an essential part of the evaluation of the thesis of the priority of the mental. This result should not be surprising, since many philosophers whose work has deflationist elements — notably the later Wittgenstein, Davidson, and Kripke in *Wittgenstein on Rules and Private Language* — focus on questions about what constitute different classes of facts about meaning and content.[21]

To ask whether trying to answer foundational questions about meaning and content is consistent with certain forms of deflationism about intentionality is one thing; to ask whether the project of trying to answer such foundational questions remains an interesting project in the context of deflationism about intentionality is another. While it will not presuppose the falsity of deflationism, the perspective from which this essay is written is, at least in the following sense, resolutely non-deflationary: I take our abilities to have contentful thoughts and speak a meaningful language to be genuine, perfectly objective facts about us. Accord-

[21] We may also remain agnostic about various controversial issues concerning the individuation of facts about meaning and content. We may, for example, ignore questions about whether the fact that Hammurabi believed that Hesperus is Hesperus is identical to the (alleged) fact that Hammurabi believed that Hesperus is Phosphorus.

These disclaimers aside, I do think that the framework of Russellian propositions provides a natural way of thinking about the nature of facts about meaning and content. On this view, facts are identified with true propositions, and so facts about meaning and content are identified with the propositions expressed by true ascriptions of meanings to signs in a language and by true ascriptions of propositional attitudes to agents. These propositions are Russellian propositions: they are thought of as structured objects (rather than as, e.g., sets of possible worlds), and as having as components objects, properties, and functions (rather than Fregean senses, construed as ways of thinking about objects, properties, and functions). On this view, the fact that John believes that Sally is nice is a structured object consisting of an object, John, a relational property, belief, and a proposition, which is itself structured object consisting of an object, Sally, and the property of being nice. Facts about the meanings of linguistic items or symbols are given a similar construal, though the property expressed by "means", unlike the properties expressed by propositional attitude verbs, is a three-place property, holding between a symbol or word, its meaning, and a language or population of speakers.

ingly, the interest I see in trying to answer these questions is that, like the study of any other central human ability, this study forms a part of philosophical anthropology. Whether or not this interest is undercut by deflationism about intentionality is, so far as I can see, an open question.

Part I

THE MENTALIST PICTURE

Chapter 2

Meaning and Intention

Contents

The mentalist picture of mind and language holds that the contents of the mental states of agents — in particular, their propositional attitudes — are the most fundamental level of representation. Above I glossed this as consisting in two claims: (i) we can give an account of other intentional facts, including facts about the meanings of expressions in public languages, in terms of the contents of the propositional attitudes of agents, and (ii) for at least some central or basic propositional attitudes, we can give an account of what it is for an agent to bear that attitude toward a proposition in terms which do not presuppose facts, such as public language meaning, which were explained in terms of propositional attitudes in (i). The next three chapters argue that neither (i) nor (ii) is tenable, and hence that mentalism should be rejected. Here and in the next chapter I focus on claim (i): the view that public language meaning is constituted by certain propositional attitudes of agents.

2.1 Two classes of propositional attitudes

Mentalists, then, share a commitment to the thesis that, insofar as there are facts about the meanings of expressions in public languages, these facts may be explicated in terms of the propositional attitudes of speakers of those languages. Confidence that such an explication is available is, I think, the main reason why issues about the foundations of linguistic meaning

are usually taken to be, at root, issues about the foundations of mental content.[1]

In practice, mentalists defend a stronger claim than that there is an analysis of linguistic meaning in terms of the propositional attitudes of agents. The class of propositional attitudes is very broad, and contains attitudes very closely linked to the use of language, such as saying, asserting, and telling. When mentalists try to give an account of language in terms of the attitudes, they try to account for facts about meaning in terms of facts about what agents intend, believe, or judge, and not in terms of what agents say or assert by uttering sentences of their language. At first sight, this focus on such a restricted class of propositional attitudes seems strange; it would, it seems, be far easier to analyze linguistic meaning in terms of what agents say or assert by their utterances than in terms of what they intend or believe.

This implicit restriction to the intentions, beliefs, and judgements of agents is motivated by the next step in the mentalist program: the explication, in non-semantic terms, of the propositional attitudes employed in the analysis of linguistic meaning. The thought is presumably that facts about the intentions, beliefs, or judgements of agents will be easier to explicate without aid of facts about the meanings of expressions in the public languages spoken by those agents than would facts about what they assert or say. One reason for thinking this is a grammatical distinction between the favored class of propositional attitudes — intending, believing, etc. — and propositional attitudes like saying and asserting.

Some propositional attitudes are more closely linked to the performance of actions than others. In the case of some propositional attitude verbs, which I shall call *action-based* attitude verbs, we can expand an ascription $\ulcorner\alpha\ V$'s that $\sigma\urcorner$ to one of the form $\ulcorner$By ϕing, $\alpha\ V$'s that $\sigma\urcorner$, where 'ϕing' denotes some action of the referent of 'α.' This class of propositional attitude verbs includes, for example, "says," "means," and "asserts"; we can expand an ascription of the form $\ulcorner\alpha$ said that $\sigma\urcorner$ to one of the form $\ulcorner$By ϕing, α said that $\sigma\urcorner$, and an ascription of the form $\ulcorner\alpha$ asserted that $\sigma\urcorner$ to one of the form $\ulcorner$By ϕing, α asserted that $\sigma\urcorner$. For example: "By uttering 'Schnee ist wiess,' John said that snow is white"; "By spreading his arms, the umpire meant that the base runner was safe." By contrast, attitude verbs like "believes" and "desires" are not action-based; a similar expansion of a sentence of the form $\ulcorner\alpha$ believed that $\sigma\urcorner$ to $\ulcorner$By ϕing, α believed that $\sigma\urcorner$ yields a sentence which is at best awkward, and at worst nonsensical.

Why is this distinction between two kinds of propositional attitudes important? The goal of a mentalist treatment of meaning is to show that the meanings of signs are derived from the contents of propositional attitudes which themselves are explicable without reference to public language meanings. The problem with propositional attitude verbs like "asserts" and "says" which the above grammatical distinction illustrates is that they are too closely linked to the meanings of signs. For usually, in a sentence of the form $\ulcorner$By ϕing, $\alpha\ V$'s that $\sigma\urcorner$, the action denoted by 'ϕing' will include mention of a sign which has a meaning in a language or population. In the case of our example sentence — "By spreading his arms, the umpire meant that the base runner was safe" — the action performed has a meaning in its context, and there is a clear link between the meaning of the action or gesture and what the umpire meant by performing the gesture. In cases where the action is the assertion of a sentence, the

[1] For example, in a recent survey of work on the foundations of semantics, Barry Loewer, referring to the claim "that the semantic properties of natural-language expressions are derived from the semantic properties of mental states," "assume[s] that some such view is correct" and "concentrate[s] on the semantic properties of mental states" (Loewer (1997), 108).

connection is even more obvious. In such cases, the obvious way to give an account of what constitutes an agent asserting or saying something will make reference to the meaning of the sentence uttered or gesture performed.[2] But, given the aims of the mentalist program, this is precisely the strategy which the mentalist must avoid.

By contrast, non-action-based attitude verbs are not so closely linked with the meanings of signs; ascriptions of non-action-based attitudes make no reference, either implicit or explicit, to meaningful signs. For this reason, the task of giving an account of facts about such attitudes which makes no reference to the meanings of signs as used in public languages appears more promising. This fact, I think, explains the practice of mentalists of trying to give an account of linguistic meaning in terms of the non-action-based attitudes of belief and intention. For ease of reference, in what follows I shall reserve the term *mental content* for facts about the non-action-based propositional attitudes.

The question facing the mentalist is: how can we give an account of linguistic meaning in terms of mental content? There are, broadly, two going answers to this question. First, one might say that the meanings of sentences are fixed by the typical causes and effects of utterances of the sentence, where these causes and effects are specified in terms of facts about mental content. Or, alternatively, one might think that the meaning of a sentence is determined by the effects which speakers *intend* to bring about about by uttering that sentence. The former route naturally yields an account of meaning in terms of belief; the latter an account of meaning in terms of intention. I discuss connections between meaning and belief in the next chapter, and the connections between meaning and intentions here.

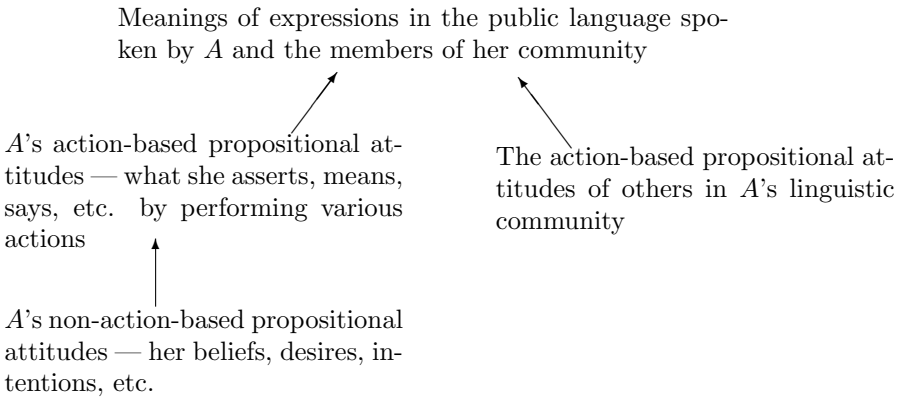
The intention-based route to an account of linguistic meaning will have to use action-based propositional attitudes as intermediaries. To see why, note that the effects which a speaker intends to bring about by an utterance need not be closely related to the meaning of the utterance; metaphor, sarcasm, and jokes are all cases in which there seems to be no very close connection between the meaning of the sentence uttered and the intentions of the speaker. But, one might think, there is, even in these cases, a close connection between the intentions of the speaker and what the *speaker* means, asserts, or communicates by her utterance. This suggests that intentions are better used to account for these action-based propositional attitudes than for the meanings of linguistic expressions.[3] The hope for the

[2]This is not to say that what an agent asserts by uttering a sentence is to be identified with the meaning of the sentence uttered (in the context); agents can assert things by uttering sentences which go beyond the meaning of the sentence uttered. Rather, the point is that it seems that the natural way to go about giving a constitutive account of these sorts of propositional attitudes will make use of the meaning of the utterance or gesture, along with, for example, facts about the intentions of the agent performing the action.

[3]One might reject this diagnosis, and attempt an account of linguistic meaning directly on the basis of intentions, by turning one's attention away from communicative intentions on particular occasions of utterances toward general intentions about word- or sentence- types, such as the intention to use "dog" to mean dog, or the intention to use "Grass is green" to mean that grass is green. It is a point in favor of this strategy that it is plausible to think that such intentions on the part of members of a linguistic community are both necessary and sufficient for the expressions to have the intended meanings in that community. The difficult question for this approach is how these general intentions are to be understood. On the one hand, they may be understood as derivative from the meanings of words in one's community; for example, the contents of one's intentions regarding word-meaning may be derived from (i) a general intention to use words as those around you do, and (ii) the meanings

mentalist would then be to use these action-based attitudes to give an account of meaning, and so to construct, indirectly, an account of meaning in terms of the intentions of speakers. The picture of mind and language envisaged by this indirect mentalist strategy might then be represented as follows:

THE INDIRECT MENTALIST STRATEGY



I shall begin discussion of this indirect strategy by considering the possibility of giving an account of some class of action-based propositional attitudes in terms of intentions; for only after seeing what sorts of analyses of action-based attitudes in terms of mental content are available will we know what action-based attitudes, if any, are available to the mentalist in her attempt to say what it is for an expression of a public language to have a meaning. There is, I shall argue, no way to explicate any of the action-based attitudes in terms of mental content. Rather, any substantive account of the contents of these propositional attitudes will treat them as derivative from facts about linguistic meaning. This result, I shall argue in the closing sections of this chapter, dooms the indirect mentalist strategy.

2.2 Grice on speaker-meaning and intentions

Since Grice's article "Meaning," the most popular choice of an action-based propositional attitude to play the middle role in a mentalist account of linguistic meaning has been the propositional attitude of speaker-meaning. As above, we can identify the class of facts about speaker-meaning with the propositions expressed by a certain class of true propositional attitude ascriptions; here, the relevant class of such ascriptions will be those of the form ⌜By

of those words as used by those around you. On this deflationary reading of these intentions, I think that it is plausible to think that all or most competent users of a term must have them; but, on this interpretation, these intentions are of no use to the mentalist, since their contents are derived from the facts about public language meaning to be explained. On the other hand — and this is the route forced upon the mentalist advocate of this strategy — the intentions may be understood as constituted independently of social facts about public languages. But then the mentalist owes a story about what does constitute these facts about the language-directed intentions of speakers; and I am skeptical that any such account could be given. The reasons for this skepticism will be clear from the discussion in Chapter 4 of the problems in giving a mentalist account of the contents of beliefs.

uttering x, α meant that σ⌝.[4] Grice's idea was that facts about the meanings of linguistic expressions can be explicated in terms of what speakers mean by uttering sentences in which they occur, and that what speakers mean by uttering such sentences can be explicated in terms of facts about a particular class of intentions of those speakers.

In the last fifty years, many revisions of Grice's analysis of speaker-meaning, some of considerable complexity, have been proposed. My aim is to argue that Gricean accounts are mistaken in principle rather than flawed in their details; unfortunately, however, there is really no way to substantiate the former claim without an examination of the details. There is a persistent thought that, although *this* version of the Gricean account fails, there is a true version just around the corner. To show that this is not the case, I shall argue that the basic version of the analysis faces fundamental objections and that subsequent refinements, far from making the account more plausible, merely exacerbate these problems. This will of necessity involve some counterexample-mongering; but throughout I hope to keep one theme in view: the Gricean account of speaker-meaning fails, time and again, by failing to recognize the role of the meanings of sentences in determining what speakers mean by uttering those sentences. This, I shall argue, strongly suggests that the order of explanation is precisely the opposite of what the Gricean usually takes it to be.

The starting point for consideration of the Gricean analysis of speaker-meaning in terms of the intentions of speakers is the following:

[G] a means p by uttering $x \equiv a$ intends in uttering x that
(1) his audience come to believe p,
(2) his audience recognize this intention, &
(3) (1) occur on the basis of (2)[5]

One important point to note about this formulation is that it requires that speakers have self-referential intentions; they must intend that their audience come to have a certain belief on the basis of their recognition of that very intention. That is, the expression "this intention" in clause (2) refers to the intention whose content is given by the conjunction of (1), (2), and (3). It is therefore essential to this version of the Gricean account that an agent who means something by an utterance have a single, conjunctive, self-referential intention rather than three separate intentions, the content of which are given respectively by (1), (2), and (3). In employing this sort of formulation, I differ from most Griceans, including — at least after 1957 — Grice himself. To say why this sort of formulation is to be preferred would take us too far astray at present. For now, suffice it to say that, in my view, there is nothing particularly problematic about these sorts of self-referential intentions, and that without them there is no plausible way to avoid a large class of counterexamples first raised by Strawson in his

[4] In what follows, I focus on cases of speaker-meaning in which an agent means something by uttering a sentence, rather than by performing a non-linguistic act. This is an artificial restriction intended to make discussion of the Gricean account simpler. For similar reasons, I restrict attention to utterances of indicative sentences.

[5] See Grice (1957), 219. As noted in §1.2 above, displayed biconditionals should be understood as elliptical for formulae such that (i) there is a necessity operator with wide scope over the formula, and (ii) all free variables are bound by universal quantifiers with wide scope over everything in the formula other than the necessity operator. I omit the modal operator and quantifiers only in the interests of making the constitutive claims under consideration more readable.

"Intention and Convention in Speech Acts."[6]

As I said, in my view this account does not give a very good picture of the nature of speaker-meaning. By way of showing this, I shall argue that none of the three conditions on speaker-meaning is necessary; that a general argument against intention-based accounts can show us that this defect will be shared by any descendant of the Gricean approach; and that these problems are the result of the thesis, which is at the heart of the Gricean program, that the meanings of linguistic expressions are to be explained by what speakers mean by using them, rather than the other way around.[7]

2.2.1 Cases of reminding and examination

Grice's account requires that, in order to mean p by an utterance, a speaker must intend that his audience come to believe p.[8] But, as has been noticed, a speaker need intend no such thing. One way to bring this out is by considering what happens when one reminds one's audience of something.

Suppose that you forget your friend's name, and I say to you, "Your friend's name is 'John'," meaning by uttering this that your friend's name is "John." I do not intend that by hearing my utterance you should come to believe that your friend's name is "John"; you already believe this, and I know that you do. Now, while this is indeed a problem for Grice's account as formulated, it is, as Grice noted, easily remedied. For, though I do not intend that you come to believe that your friend's name is "John," I do intend that you come to be in a closely related mental state with the same content: that of, as Grice puts it, *actively* believing that your friend's name is "John." Roughly, we can take actively believing p to be equivalent to believing p and entertaining or thinking p. So we may revise the Gricean account accordingly: it is a necessary condition on speaker-meaning not that a speaker intend that his audience come to believe something, but that he intend that his audience come to actively believe something.

But this solution is only a temporary fix: it is not a necessary condition on meaning p by an utterance that a speaker intend that his audience actively believe p. Suppose, for example, that I am in an oral examination in the philosophy of language, and that Saul Kripke is one of my examiners. In response to a question, I say, "The meanings of proper names are equivalent to the meanings of descriptions." By uttering this I mean that the meanings of proper names are equivalent to the meanings of descriptions, but I do not intend that Kripke come to believe (or actively believe) this; I know perfectly well that he will not. So our amended version of the

[6]For an extended discussion of Strawson's argument, different proposed solutions, and a defense of the coherence of this kind of self-referential intention, see Appendix A, pp. 199 ff. below.

[7]One objection I will not discuss in the text is the oft-cited objection that the Gricean analysis gets epistemic priorities backwards, since we typically come to know the intentions of speakers (in part) by knowing the meanings of the sentences they utter, rather than the other way around. (See, e.g., Biro (1979), Platts (1979).) This objection rests on a misunderstanding of the Gricean program. The goal is not to give a theory of the epistemology of language use, but rather to say what it is for a speaker to mean something by an utterance.

[8]As noted above, the version of Grice's account under consideration requires that the speaker have a single conjunctive intention; in assuming that speakers must also have three intentions with the contents given by the appropriate versions of the three clauses, I am assuming that intention, like belief, distributes over conjunction.

account still has the unwanted implication that I did not mean that the meanings of names are equivalent to those of descriptions.[9]

Now, the defender of the Gricean account may object that these are a superficial sort of counterexample; for although in either case I do not intend that my audience come to have a certain belief, I certainly do intend that my audience come to be in a certain mental state whose content is the same as what I meant by my utterance. In this case, it seems that, although I did not intend that Kripke come to actively believe that the meanings of proper names are equivalent to the meanings of descriptions, I did intend that he actively believe *that I believe* that the meanings of names are equivalent to the meanings of descriptions. One might think, then, that we should divide utterances into two classes; those in which the speaker intends to bring about a certain belief in his audience, and those in which the speaker intends to reveal something about his own beliefs to an audience. This was Grice's idea; he labelled the former class *protreptic*, and the latter *exhibitive* utterances.[10]

How can this distinction be incorporated into our account of speaker-meaning? The obvious move is to appeal to a disjunctive account; that is, to claim that an agent a means p by uttering x iff *either* a intends in uttering x that (1) his audience come to actively believe p, (2) his audience come to recognize this intention, and so on, *or* a intends in uttering x that $(1')$ his audience come to actively believe that a believes p, $(2')$ ....[11] And this sort of disjunctive analysis, though a bit unwieldy, seems suited to handle each of the proposed counterexamples discussed so far.

Unfortunately, the move to a disjunctive account weakens the conditions to such an extent that they are no longer sufficient for speaker-meaning p. Suppose that I say, "I believe that the meanings of names are equivalent to the meanings of descriptions." I intend by uttering this that you believe that I believe that the meanings of names are equivalent to the meanings of descriptions; but by the second disjunct of this account it follows that I meant by my utterance that the meanings of names are equivalent to the meanings of descriptions. But of course I didn't mean this, but only meant that *I believe* the proposition in question.[12]

[9] Examples with the same moral include cases of examination in which a student is asked a question, and gives an answer which, he believes, the teacher already believes; moreover, since the teacher is presently asking him the question, it would be reasonable for him to think that she actively believes it. Hence he does not intend that she come to actively believe it.

[10] Grice (1969), 111.

[11] Fully spelled out, the account would look like this:

a means p by uttering $x \equiv$

(a) a intends in uttering x that

(1) his audience come to actively believe p,

(2) his audience recognize this intention, &

(3) (1) occur on the basis of (2); *or*

(b) a intends in uttering x that

$(1')$ his audience come to actively believe that a believes p,

$(2')$ his audience recognize this intention, &

$(3')$ $(1')$ occur on the basis of $(2')$

Neale (1992) also suggests that a disjunctive account, presumably along these lines, may be required.

[12] A further problem is that, in many cases where I mean p by uttering x, I both intend that my audience believe p and intend that my audience believe that I believe p. In these cases, the present version of the account yields the (often incorrect) result that by my utterance I meant both p and

It is no solution to this problem to regard all utterances as exhibitive; the unwanted implication would follow from an account of speaker-meaning in terms of the second disjunct of this account alone. Moreover, this revision would not be a substantial advance over the purely protreptic account we had in hand before; for while this move would correctly classify the cases of oral examinations discussed above as cases of speaker-meaning, it would give rise to a different class of counterexamples which do not arise under purely 'protreptic' analyses like that with which we began: namely, cases in which speakers mean p by an utterance, intend that their audience believe p, intend that they come to this belief on the basis of the recognition of their intention, but do not intend that their audience believe that they believe p.[13]

So, if it is to be of any use to us, we need a different way of implementing Grice's distinction between exhibitive and protreptic utterances. And, one might think, a candidate is ready to hand: we should simply make Grice's distinction between exhibitive and protreptic utterances explicitly part of the account by supplementing our disjunctive account. This might run as follows: an agent a means p by uttering x iff *either* if x is a protreptic utterance, then a intends in uttering x that (1) his audience come to actively believe p, ... *or* if x is an exhibitive utterance, a intends in uttering x that (1′) his audience come to actively believe that a believes p, ...[14]

---

that I believe p. It seems that some notion of a 'primary intention' of the speaker is required to rule these cases out; below, I give some reasons for thinking that there may be no easy way to integrate this notion into a version of the Gricean analysis.

A tempting idea is to make the first clause, rather than the right-hand side as a whole, disjunctive; that is, to require that a speaker who means p by an utterance intend by that utterance that his audience either come to believe p or come to believe that he believes p. But this runs into precisely the same problem. Moreover, it is not much in the spirit of Grice's intuitive distinction; the distinction between exhibitive and protreptic utterances is intended to allow that speakers may have either of two sorts of intentions, and not to require that they have one disjunctive sort of intention.

[13]The following is an example of this kind: Bob is an air-traffic controller at Newark Airport, who relays information from the head controller to arriving pilots. Bob does not believe this information; he thinks that the head controller is involved in a conspiracy to cause air accidents. Nevertheless, Bob has an overly strong sense of duty, and feels it his duty to convey the information passed to him from the head controller to pilots. So when the head controller tells him that Runway 2 is open for landing, Bob says over his radio to the appropriate pilot, "Runway 2 is open for landing." Bob intends by uttering this that the pilot should believe that Runway 2 is open for landing, and intends that he should come to that belief on the basis of recognition of his intention. Nevertheless, Bob certainly does not intend that the pilot come to believe that he, Bob, believes that Runway 2 is open for landing; Bob doesn't believe this, and would say so were an accident to occur. See McDowell (1980) for further objections to treating all utterances as exhibitive.

[14]Once again, the full version of the account:

a means p by uttering $x \equiv$

(a) if x is a protreptic utterance, then a intends in uttering x that

(1) his audience come to actively believe p,

(2) his audience recognize this intention, &

(3) (1) occur on the basis of (2); *or*

(b) if x is an exhibitive utterance, a intends in uttering x that

(1′) his audience come to actively believe that a believes p,

(2′) his audience recognize this intention, &

(3′) (1′) occur on the basis of (2′)

But this account, like its predecessors, fails; the reason why it does so is instructive. The goal of Gricean accounts of speaker-meaning — insofar as they are to be of any use to the mentalist — is to give a non-circular account of facts about what speakers mean by their utterances. The problem with our latest version is that it does not avoid circularity. Facts about which utterances are exhibitive and which protreptic figure in this account; and it is not clear that there is any way to distinguish between these two classes of utterances which does not employ facts about what speakers mean by their utterances. The loose definitions of these notions I gave above — that an exhibitive utterance is one by which the speaker means p and intends that the audience actively believe that he believes p, while a protreptic utterance is one by which the speaker means p and intends that the audience actively believe p — will clearly not do. Because this version of the account is only acceptable to the mentalist if the distinction between exhibitive and protreptic utterances can be made out in such a way as to avoid this sort of circularity, and because this seems most unlikely, it seems that no version of the distinction between exhibitive and protreptic utterances is of use to the Gricean who claims that what it is for a speaker to mean something by an utterance is for that speaker to have certain communicative intentions.

What have these examples shown us so far? The Gricean account relies on there being a correlation between what a speaker means by an utterance and the contents of mental states that he intends that his audience come to be in. In each case discussed so far, there is a mental state ready to do the work the Gricean needs done; the problem is that *which* mental state this is varies from case to case. Sometimes it is believing what the speaker meant; at other times it is believing that the speaker believes what he meant. The problem we are encountering is that there seems to be no good way to combine these two sorts of utterances in an account of speaker-meaning.

Is this really a serious problem for the Gricean? Note that no one, whether a Gricean or not, should deny that there are bound to be *some* systematic connections between what speakers mean by their utterances and what beliefs they intend to bring about; everyone should agree that often meaning something by an utterance involves intentionally conveying information, and that often conveying information involves intentionally bringing about new beliefs in others. What is at issue is how these systematic correlations should be explained. The Gricean claims that they are to be explained by the fact that *what it is* for a speaker to mean p by an utterance is for him to intend to bring about a certain belief in an audience via a certain kind of mechanism. What we have found so far is that the correlations between speaker-meaning and intentions to bring about beliefs are not of a sort to support the Gricean's claim.

The Gricean is not, however, quite out of options. For why, he might ask, should we focus on intentions to bring about *beliefs*? Perhaps the right intentions are of a different sort; perhaps, as Stephen Neale has suggested, they are intentions to cause agents to *entertain* certain propositions.[15] Taking up this suggestion, the Gricean might advance the following thesis: what it is for an agent to mean p by uttering x is for that agent to intend that his

Note, though, that this formulation can't be quite right; it follows from this that any utterance of a sentence x by a speaker a which is neither exhibitive nor protreptic is an instance of the speaker meaning p by uttering x, *for any proposition* p. I ignore this in what follows, assuming that the problem could be solved by complicating the left-hand side of the biconditional.

[15] Neale (1992).

audience entertain the proposition p on the basis of recognition of that intention.[16]

This amendment, however, surrenders sufficiency for breadth. This can be seen by considering utterances of complex sentences. Suppose I say to you, "Either 'John' is your friend's name or I am mistaken", meaning thereby that either "John" is your friend's name or I am mistaken. I do intend by uttering this, in keeping with Neale's suggestion, that you entertain the disjunctive proposition that either "John" is your friend's name or I am mistaken. But presumably I also intend that you entertain the proposition that I am mistaken, given that there is no way to entertain a disjunctive proposition other than by entertaining its disjuncts. But if this is so, then Neale's suggestion yields the incorrect result that I meant by my utterance that I am mistaken. Similar problems arise in the case of conditionals.

The natural response on the part of the Gricean here is to appeal to the notion of a primary intention. That is, a proponent of the foregoing version of the analysis might reply to this criticism that the right-hand side of the Gricean analysis should be modified to read, "*a primarily* intends by uttering x that ...", with the ellipsis filled in by the appropriate version of the analysans. As used by Griceans, the notion of a primary intention seems to be that originally suggested by Schiffer: "To specify one's primary intention in doing X ... is to give one's reason for doing X," as opposed to specifying an intention which one has because one has a given primary intention.[17] But it should be clear that modifying the Gricean analysis along these lines will yield an account which is far too restrictive. Suppose that your car runs out of gas, and I say to you, "There's a gas station around the corner." Clearly, though I do intend that you come to believe that there is a gas station around the corner, I only have this intention because I intend you to come to believe that you can fill your car up around the corner; so the former is a secondary intention and not, in the relevant sense, the reason for my utterance. Nevertheless, it does seem as though I meant by my utterance that there is a gas station around the corner; and this is a counterexample to any version of the analysis restricted to the primary intentions of speakers.

A more fundamental problem with the switch from believing to entertaining propositions is that the conditions given make it almost impossible, in many contexts, for speakers to mean things by their utterances. Suppose I know that, usually, when you hear a certain sentence which in the context means p, you immediately, almost reflexively, entertain p. This is, after all, not a farfetched scenario; it is the condition that normal speakers are in with respect to sentences they understand. But if I know this, then I can hardly intend that you come to entertain p *on the basis of your recognition of my intention* when I utter this sentence; I know that you'll entertain p upon hearing this sentence, quite independently of your recognition of my intentions. So it seems that, for several reasons, the Gricean was right to focus on belief rather than other propositional attitudes. But, as we saw above, there is no true account in terms of intentions to bring about beliefs to be had.

The point of plodding through this series of counterexamples, revisions, and yet more counterexamples is to dispel the illusion that the Gricean account of speaker-meaning is just a correction or two away from being adequate. A pattern here is worth noting: as soon as one

[16] This is not Neale's version of the thesis; Neale suggests that we drop clause (3) from the account altogether. That is, he suggests that we should drop the requirement that the speaker intend that the audience entertain the proposition *on the basis of* recognition of the speaker's intention. This proposal is, I think, pretty clearly on the wrong track; I discuss it at some length in the next section.

[17] Schiffer (1972), 62.

hole in the Gricean account is patched up, another (or, more often, several) appear. Here is one diagnosis, which I hinted at above: the Gricean has latched onto a genuine phenomenon in the links between what speakers mean by utterances and the effects those speakers intend to bring about by those utterances. There are also, as should be clear from the examples used above, systematic links between the meanings of sentences and what speakers mean by uttering those sentences. The question is how these two classes of links are to be explained. The Gricean's idea is that the first sort of link should be used to explain the second: that intentions to bring about effects explain what speakers mean by their utterances, and that this in turn explains why the sentence has the meaning that it does. The failure of the first stage of this explanatory strategy thus far might lead one to think that that the relationship between these two classes of systematic links should be viewed differently; perhaps, for example, facts about speaker-meaning should be explained jointly by facts about what sentences mean and facts about what speakers intend. Were this the case, one would expect what we have in fact found so far: that every attempt to give an account of speaker-meaning in terms of intentions alone seems to say something substantial and true about many cases, but goes awry in cases where, in some sense as yet to be explicated, what the speaker means depends largely on what the sentence she uttered means.

This is, as I say, one diagnosis of what is going on in these cases; but the Gricean is likely to offer another. The Gricean should, I think, remind us that, as noted in §1.2 above in connection with Wittgenstein's comments on family resemblance concepts, there is no reason to think that the concepts we employ in ordinary language can be given strict analyses. Rather, the best that we can hope for is to provide conditions which capture the central cases of the phenomenon in question. And it is not unreasonable to think that our original formulation [G], for all that has been said so far, succeeds in doing just that. The cases which have posed problems for that account — cases of reminding, answering examination questions, and the like — should, on this diagnosis, be regarded as side cases, and hence not as a constituting a decisive argument against the Gricean picture of speaker-meaning.

As we'll see in the next two sections, though, these cases are not the only ones which pose a problem for the Gricean. There are, in addition, two very broad classes of language use which do not fit the Gricean paradigm.

2.2.2 Persuasive discourse

The main examples of the first class of cases come from persuasive discourse. The Gricean account requires for a speaker to mean p not only that she intend that her audience come to believe p and that her audience recognize this intention, but also that her audience come to this belief *on the basis of that recognition*. The problem is that, when giving an argument of some kind, a speaker will typically intend that her audience come to believe the conclusion on the basis of belief in the premises, and *not* on the basis of a recognition of the speaker's intention. This class of counterexamples extends across a very wide range of cases; it includes not only cases of formal argumentation in classrooms and courtrooms, but any use of language in which a speaker intends to bring about a belief in her audience on the basis of previously adduced information rather than recognition of speaker's intention.[18]

[18] Another class of counterexamples to clause (3) of the account, though a less significant one, comes from a sub-class of the class of cases of reminding, discussed above. Suppose that you cannot

Although Grice was the first to notice this problem for his account of speaker-meaning, he never offered a solution.[19] But clearly, if the account is to cover all cases of speaker-meaning, some solution for these cases must be found; denying that these are genuine cases of speaker-meaning does not seem to be a viable option.[20] It is tempting to try to solve these cases by treating them as exhibitive utterances; one might think that, even if a speaker does not intend that the audience believe p on the basis of recognition of this intention, the speaker might intend that the audience believe that the speaker believes p on the basis of this intention.[21] But in neither case will the amendment solve the problem; we can imagine cases in which the speaker knows that the audience already knows that the speaker believes the proposition of which he is being reminded or persuaded, or cases in which the claim that a speaker has a certain belief is in fact the conclusion of the argument.

The only response on offer to these problems which counts them as cases of speaker-meaning is a radical one, which has been proposed by Stephen Neale. Neale proposes that condition (3) simply be dropped from the account, to yield the following:

> a means p by uttering $x \equiv a$ intends in uttering x that
> (1) his audience come to believe p, &
> (2) his audience recognize this intention.[22]

This version of the Gricean account has two main advantages: it correctly characterizes the cases of persuasion and reminding discussed above as genuine instances of speaker-meaning, and thus gets closer to providing a necessary condition for speaker-meaning; and it simplifies the analysis and so limits the intentions we are forced to attribute to speakers.

But even so, it fails to provide a sufficient condition for speaker-meaning. Consider a case in which we have a surgeon capable of manipulating a patient's brain so as to make him

---

remember your friend's name, though you feel as though it is on the tip of your tongue. I say to you, "Your friend's name is 'John'," by which I mean that your friend's name is "John." But, although this seems clearly to be a case of my meaning something by an utterance, it does not satisfy the conditions laid down by the account. Again, clause (3) requires that I intend that you come to believe p — in this case, come to believe that your friend's name is "John" — on the basis of your recognition of my intention that you come to believe this proposition. But in this case, I don't intend any such thing; I know that the name is already on the tip of your tongue, and that mere mention of the name will be sufficient for you to remember your friend's name. In this situation, I can hardly intend that you remember this on the basis of recognition of my intention that you do so. This class of cases is thus different than the class of cases of reminding discussed above; no distinction between activated and dispositional belief will solve the class currently under consideration.

[19]See Grice (1969), 107. None of the points Grice goes on to make in this article — e.g., his distinctions between exhibitive and protreptic utterances and between activated and dispositional belief — helpful as they are with other cases, do anything to solve the problems the present class of cases create for clause (3). Puzzlingly, though, by the end of "Utterer's Meaning and Intentions" he takes himself to have given necessary and sufficient conditions for speaker-meaning (115-6).

[20]Schiffer (1972) seems inclined in some places to deny that these are genuine instances of speakers meaning something by an utterance.

[21]This is suggested in Rumfitt (1995).

[22]See Neale (1992), 547-9. Neale's account is somewhat different than this one, since his solution to the cases of deception does not appeal, as this does, to self-referential intentions. In addition, as discussed above, Neale requires that the speaker intend that the audience merely entertain the proposition meant, rather than believe it.

disposed to assent to certain sentences. The surgeon, in this case, intentionally manipulates the patient in such as way as to make the patient disposed to accept both "S" and "My surgeon intends me to believe that S". Plausibly, the patient's being so disposed is sufficient to guarantee that the patient believes that S, and believes that the surgeon intends that he believe that S.[23] By Neale's account, then, it follows that by manipulating the brain of his patient, the surgeon meant that S by manipulating the brain of the patient, which seems clearly to be incorrect.

This case may seem too fantastic to be convincing; but the basic idea behind it is not fantastic at all. The idea is that there are many ways to cause people to believe a proposition p without uttering something by which you mean p. Cases of brain surgery, drugging, and brainwashing are only the most dramatic; there are also various indirect ways of bringing about beliefs in other people. But, on Neale's proposal, all that is required for a speaker to mean p by an action is that he intend to bring about two beliefs in someone else: p and that the speaker intends that person to believe p. The point of the above example was only to illustrate the fact that if one intends to bring about each of these beliefs, but does not intend that there be any significant connection between the audience coming to have these two beliefs, it will seem strange to describe what you've done as 'meaning p by your action.' In effect, Neale's proposal is to drop the most fundamental part of Grice's thought about speaker-meaning: the thesis that speaker-meaning is to be analyzed in terms of an intention which is to be fulfilled on the basis of its own recognition. If this part of the program is dropped, then it seems better to take another track altogether rather than to defend a thesis like this one.[24]

2.2.3 Speaker-meaning without intended effects

The second broad class of cases which are incorrectly classified by the Gricean account are cases in which the speaker means something by an utterance but does not intend to bring about any beliefs in his audience at all. Indeed, this is a special case of the more general fact that it is not a necessary condition on meaning something by an utterance that the speaker intend to bring about any effects in an audience.[25]

There are many sorts of cases which illustrate this point. Perhaps the clearest are cases in which a speaker means something by an utterance but has no audience, and so *a fortiori* does not intend to bring about beliefs in his audience. Some such cases can be handled by expansions of the Gricean analysis. For example, I might mean that trespassing is not allowed by erecting a sign to that effect on my property even though there is no audience present when I put the sign up; in this sort of case, the Gricean can say that there is an

[23] Note that we need not imagine the patient asserting these sentences in robotic fashion immediately after waking from surgery; the patient may do so, reflectively, some time after the surgery. I assume to ensure that clause (2) is satisfied that (for unambiguous non-context-sensitive sentences) $\ulcorner\alpha$ comes to believe that $\sigma\urcorner$ entails $\ulcorner\alpha$ recognizes that $\sigma\urcorner$.

[24] Neale proposes to drop clause (3) because he finds the examples Grice originally used to motivate this clause uncompelling. I agree with Neale about this; but examples like the case of the surgeon discussed above show that clause (3) is required to ensure the sufficiency of the Gricean conditions.

[25] This limitation on Gricean accounts of speaker-meaning was first emphasized by John Searle in his *Speech Acts* (1969), 42 ff. This point is also an effective objection to the neo-Gricean theory advanced in Armstrong (1971).

intended audience, which should be specified by a description: I intend that anyone who reads this sign come to believe that trespassing is not allowed on the basis of recognition of this intention. But, as Chomsky and Harman have emphasized, at least one substantial class of audienceless utterances resists incorporation into a Gricean account of speaker-meaning: uses of language in calculation or in thought.[26]

There are a host of such cases in which a speaker seems to mean something by an utterance or action but seems to have no intentions at all regarding an audience: cases of writing in a diary, writing down figures in a check book, working through a philosophical problem aloud when taking a walk. There are similar cases to do with the expression of emotions; cursing someone angrily after they've left the room, for example. How is the Gricean to account for these cases?

One might identify the intended audience in these cases with oneself.[27] It is not implausible that these cases represent, in some sense, a dialogue with oneself; the problem with the suggestion comes with integrating it with the rest of the Gricean analysis. Are we to say that by writing something in my diary I intend to bring about a certain belief in myself by recognition of this (my own) intention? I already know that I have the relevant belief; this is presumably part of the explanation of my writing as I did in my diary, and not the intended effect of my writing.

Another approach, suggested by Grice, is to say that, for some description D, in these cases the speaker intends that, were someone who satisfies D present, they would come to have the relevant belief on the basis of recognition of this intention.[28] The appropriate description will differ from case to case; in many cases, it might simply be "a competent speaker of my language." The immediate problem is that among the things I say in the absence of an audience may be things that I would never want anyone to know. Suppose, for example, that I've committed some terrible crime, of which I'm deeply ashamed. I might state the crime in the absence of an audience; of course I do not intend that, where someone present, they would come to believe that I was the author of the crime. I intend that, were someone to be present, they would misunderstand me, take me to be joking, or be otherwise deceived.

These audienceless cases point to a fundamental problem with the Gricean account: intending to bring about effects by one's utterance is not an essential part of meaning something by an utterance. Audienceless cases are one way of making this manifest, but there are others which do the same. Consider an innocent person arrested on charges of espionage by his own government; under torture, he continues to claim that he is not a spy, simply because he feels that it is his duty to do so. He knows that his torturers will not believe him; at this point he doesn't even care if they do. Nonetheless, the torturers are his audience, and he does mean by his utterance that he is not a spy.[29] In general, there is simply nothing to prevent my meaning something by an utterance directed at a particular audience while being perfectly

[26] See especially Chomsky (1975), and Harman op. cit. One might be inclined to argue that such cases can't legitimately be used against the Gricean mentalist, since this position is committed only to there being an explication of public language meaning along Gricean lines and not to there being an adequate Gricean treatment of individual languages, like languages of thought. The problem with this objection is just that the cases at issue are clearly cases of agents meaning something by utterances of sentences in a public language.

[27] As do, for example, Schiffer (1972), at 79-80, and Avramides (1989), at 65-6.

[28] Grice (1969). See also Hyslop (1977).

[29] The case is a variant of one presented in Harman (1974).

indifferent as to whether that audience comes to acquire any beliefs at all as a result of my utterance.[30]

2.2.4 Meaning, speaker-meaning, & Moore's paradox

I mentioned at the outset of this chapter that there is a persistent thought to the effect that, even if we have not yet arrived at the right form of the Gricean account, some true version is out there to be discovered. The discussion of the problems raised by revising Grice's original account has been designed to discourage this thought. I suggest that one source of these problems is a connection between the meaning of the sentence uttered by the speaker and what the speaker meant by uttering the sentence; this connection may be expressed by the following principle:

[M/S-M] If an agent utters a sentence which means p sincerely and seriously (without sarcasm, irony, etc.), then the agent means p by her utterance.

This principle connecting meaning and speaker-meaning has been in the background of many of the arguments of the preceding sections, for many of the cases for which the Gricean has been unable to account have been cases in which among the things meant by the speaker was the meaning of the sentence in the context of utterance. Inasmuch as he seeks to account for speaker-meaning wholly in terms of the intentions of speakers, it is unsurprising that this principle should pose a problem for the Gricean.

As it turns out, though, this principle can also be employed in a general argument involving Moore's Paradox which shows that no intention-based account of speaker-meaning can be correct. G. E. Moore drew attention to the oddness of uttering sentences like

It is raining, but I do not believe that it is raining.

However odd it is to utter such a sentence — and it is difficult to imagine a scenario in which a speaker who understands this sentence could utter it sincerely and seriously — there are contexts of utterance in which a sentence of this form is clearly true. To construct such a context, after all, we need only select a speaker who does not believe some truth, find some sentence 'S' which expresses this truth, and assign that speaker as the speaker of $\ulcorner S$, but I do not believe that $S\urcorner$ in the context.[31]

Other quasi-paradoxical sentences may be generated by focusing, not on belief, but on intention; the one relevant to evaluation of the Gricean account of speaker-meaning is

It is raining, but I do not intend you to believe that it is raining.

[30] The case of prayer is another counterexample of the sort under discussion, in which the speaker clearly means something by her utterance but does not seek to bring about any effects in her audience. There is clearly an intended audience; but, since the person praying believes her audience to be omniscient, she does not intend to bring about any beliefs in her audience by her utterance. Yet it is clear that she means something by her utterance.

[31] Here I am thinking of contexts abstractly, either as 'centered worlds' or as a set of parameters, such as a speaker, audience, time, world, place, etc.

As with Moore's original paradoxical sentence, it is clear that there are some contexts of utterance in which the sentence is true. The key question here is: are there contexts in which a speaker could utter this sentence seriously and sincerely?

It seems so. For consider the following three examples:

> An unfaithful husband is found out by his wife; he might deeply regret his act and love his wife, but know that there is no way to convince her of this. Still, he might well say to her, "I love you, though I don't expect (intend) you to believe that." It seems clear that what he says might well be true.[32]
>
> A harbors ill feelings for B. Moreover, he knows that B knows this, and that B will disbelieve anything that he says. A calculates to himself the best way to cause B the most discomfort, and hits on the following strategy: he will tell B that it is raining outside, knowing that B will upon hearing this immediately come to believe that it is not raining outside. Then, he hopes, B will fail to bring an umbrella on his next trip outside, get wet, and feel all the worse for knowing that his enemy A had warned him truthfully in advance. For added confusion, A also decides to tell B, truthfully, that he does not intend him to believe that it is raining. So A says to B: "It is raining, but I don't intend you to believe that it is raining."
>
> Suppose that A has just taken a truth serum, knows that B hates being wet, and, as above, harbors ill feelings for B. While gazing at B malevolently, he says, "It is raining, but I don't intend you to believe that it is raining."

In each of these cases, an agent utters a sentence which has a meaning; and in each of these cases, it seems that the agent utters the sentence seriously (they are not being sarcastic, ironic, joking, etc.). But if this is right, then it follows from the above principle connecting linguistic meaning with speaker-meaning that each of the agents in these examples means by their utterance what the sentence means in the context of utterance.

With these cases in mind, we can then argue as follows:

[1] The speaker (A) utters a sentence to his audience (B) which means (p & A does not intend B to believe p)

[2] The sentence uttered by the speaker is true.

Since the sentence uttered was in each case true, each of its conjuncts must be true; so we get as a further premise the second conjunct:

[3] A does not intend B to believe p ([1],[2])

But since the utterances are serious (nonsarcastic, etc.), we can derive a further claim about what the *speaker* means by her utterance:

[32] Thanks to Jonathan Beere for this example.

[4] By her utterance, A means (p & A does not intend B to believe p) ([1],[M/S-M])

So we have reached the conclusion that some conjunctive proposition is meant by the speaker in these contexts. To get the desired conclusion, we must employ a further premise:

[5] The propositional attitude relation expressed by "means" distributes over conjunction; that is, indexicality aside, a sentence of the form $\ulcorner\alpha$ means that σ and $\sigma'\urcorner$ entails $\ulcorner\alpha$ means that $\sigma\urcorner$ and $\ulcorner\alpha$ means that $\sigma'\urcorner$

Premise [5] seems to me intuitively quite plausible; how could a speaker mean that p and q without also meaning that p? The case seems even clearer when speaker-meaning is laid alongside other action-based propositional attitudes, such as assertion. It seems clear that assertion distributes over conjunction; if John asserts p & q, then John asserts p.[33] Indeed, virtually all of the action-based propositional attitudes which come to mind — telling, informing, saying, communicating, commanding — clearly do distribute over conjunction. It would be surprising if, lone among the members of this class, the propositional attitude of speaker-meaning did not distribute over conjunction.[34]

The problem for the Gricean is that [4] and [5] jointly entail

[6] By her utterance, A means p

[33] If claims about assertion entail claims about speaker-meaning — which is not altogether implausible — then this would be enough to guarantee the truth of [5].

[34] One might object that the relation expressed by "means" when used between symbols, rather than agents, and propositions clearly does not distribute over conjunction. For example, even though "'It is raining and arithmetic is incomplete' means that it is raining and arithmetic is incomplete" is true, "'It is raining and arithmetic is incomplete' means that arithmetic is incomplete" is false. But the propositional attitude of speaker-meaning, inasmuch as it is a relation between agents and propositions rather than between symbols and their meanings, seems much more closely related to other action-based propositional attitudes than to this use of "means."

It is worth noting that, even without the supposition that speaker-meaning distributes over conjunction, we could derive the result that, in the scenario described above, it is true both that A does not intend that B believe that it is raining, and that A means by uttering "It is raining, but I do not intend you to believe that it is raining" to B that it is raining and A does not intend B to believe that it is raining. That is,

(By uttering x to B, A means p & q) & $\neg$(A intends that B believe p)

While this does not immediately entail the falsity of the Gricean account, it seems to me that it does have counter-intuitive consequences. If the first conjunct of this formula is true, then, by the Gricean's lights, A must intend that B believe p & q; thus, according to the Gricean, it is a necessary truth that in any context C which meets the constraints described above, the speaker A's intentions as regards B must be as follows:

(A intends that B believe p & q) & $\neg$(A intends that B believe p)

But this seems implausible. Suppose that A is a sophisticated philosopher of language, who knows that belief distributes over conjunction, and thus that it is impossible that B believe p & q without believing p; he would never intend that B believe p & q without also intending that B believe p.

This is problematic because the conjunction of [6] and [3] is inconsistent with the Gricean claim that what it is for a speaker to mean p by an utterance is for that speaker to intend to bring about the belief p in her audience.

Of course, we have by now discussed a number of other cases in which having this sort of audience-directed intention fails to be a necessary condition for meaning something by an utterance. The interest of the present argument is that it does not depend on the details of specific formulations of the Gricean account. The examples discussed above began with an utterance of the form $\ulcorner S$, but I do not intend that you believe that $S\urcorner$; as such it counts against the version of Griceanism which takes intentions to bring about believing p to be constitutive of meaning p; but we just as easily could have begun with utterances of the form $\ulcorner S$, but I do not intend that you believe that I believe that $S\urcorner$ or $\ulcorner S$, but I do not intend that you actively believe that $S\urcorner$ or $\ulcorner S$, but I do not intend that you entertain the proposition $S\urcorner$. So this is a class of counterexamples with substantial generality; it seems likely that, whichever audience-directed intention is seized upon by a particular version of the Gricean analysis, we could construct quasi-paradoxical sentences of a form appropriate to refute that analysis.

2.2.5 Assessment of Grice's account

So where does this leave the Gricean analysis? We have identified several broad classes of cases of speaker-meaning which the analysis does not capture, and presented one argument for the conclusion that no modification of the analysis will be able to capture all cases of speaker-meaning. Even without this general argument, it seems very unlikely that the Gricean account could be modified so as to solve each of the problem cases; cases of examination answers, persuasive discourse, and cases in which there is no intended effect in an audience all point to *different* problems with the account. It is implausible in the extreme, I think, to claim that all of these exceptions are, strictly speaking, not cases of speakers meaning anything by utterances. The right conclusion is that if the Gricean analysis is to play any role in a mentalist account of language, it cannot be that of an analysis of speaker-meaning.

Further, the cases discussed above seem to show not only that the Gricean account of speaker-meaning fails to provide exceptionless necessary and sufficient conditions for a speaker meaning p by an utterance, but also that the Gricean account is a fundamentally mistaken picture of speaker-meaning. The two basic tenets of the Gricean account are that speaker meaning is essentially a matter of intending to bring about certain effects, and of intending that these effects be brought about via recognition of the intentions of the speaker; and, as we have seen, neither is at all essential to a speaker's meaning something by an utterance. Furthermore, even in cases in which a speaker *does* intend to bring about beliefs in an audience via the Gricean mechanism, the contents of these intended beliefs are often not a reliable guide to what the speaker meant by her utterance.

This claim is reinforced by the fact that there is a pattern to the cases which pose problems for the Gricean account of speaker-meaning. In each of these cases, the speaker meant something which was also the literal meaning of his utterance, and the Gricean conditions counted the speaker as not having meant this by his utterance. One moral to draw from this is that, if one's utterance means p, then it is very easy for one to mean p by one's utterance. This result also suggests that of the two notions, linguistic meaning and speaker-meaning, the former is, contra the Gricean order of explanation, the more fundamental one. If there is any

true non-trivial theory about what it is to mean something by an utterance, that account will have to be given partly in terms of the meanings of the expressions uttered by speakers.[35]

But might the Gricean account fare better as an account of some other propositional attitude relation? Nothing in the indirect mentalist strategy, after all, requires that the action-based propositional attitude which bridges the gap between linguistic meaning and non-action-based propositional attitudes be speaker-meaning.

Some evidence for this comes from a grammatical distinction between two classes of action-based propositional attitude verbs. For some such verbs, an ascription of the form ⌜By ϕing, α V's that σ⌝ can be expanded to an ascription of the form ⌜By ϕing, α V's that σ to β⌝, which includes mention of an audience as the indirect object of the verb. The Gricean analysis makes essential reference to an audience; it thus seems plausible that, if it is a successful analysis of any propositional attitude relation employed in ordinary speech, it will be one expressed by a verb in this class. Given this, it is striking that the propositional attitude of speaker-meaning is not in this class; there is no natural sentence ⌜By uttering x, α meant that σ to β⌝. Verbs which naturally take a singular term referring to an audience as indirect object are, unsurprisingly, verbs very closely related to communication, such as "communicated," "told," "informed."[36]

Might it be the case that the Gricean analysis is an adequate analysis of communication? If so, then perhaps the following analysis would be true:

a communicates p to his audience by uttering $x \equiv$
  a intends in uttering x that
    (1) his audience come to actively believe p,
    (2) his audience recognize this intention, &
    (3) (1) occur on the basis of (2)

But meeting these conditions is neither necessary nor sufficient for communication. The conditions are not sufficient because "communicates" is a success verb; if I say that I communicated something to an audience, that entails that my audience understood what I was trying to get across; otherwise, it would be natural to describe the situation as one in which I tried to communicate something, but failed.[37] The conditions are not necessary because, in many cases, we would take the statement of the conclusion of an argument as a case of communication; but, as we saw before, persuasive discourse does not meet condition (3) of the Gricean analysis.

The first of these problems seems also to be a problem for the construal of the Gricean analysis as an analysis of the propositional attitude expressed by "informs." There does seem

[35] For a sketch of such an account, see Appendix C, "A communitarian account of speaker-meaning."

[36] Sentences of the form ⌜S, but I do not intend you to believe that S⌝, discussed in the last section, provide further indication that this is the best way to view the Gricean analysis. I claimed that there are contexts of utterance such that an utterance of such a sentence in those contexts would be true; but it is no doubt also true that it is difficult to imagine a situation in which an utterance of such a sentence would be conversationally appropriate. This indicates that these would not be useful sentences in communication between speakers; and because the truth of such sentences can be used in an argument against the Gricean analysis, it seems plausible that this shows that the Gricean analysis is best regarded as an analysis of a propositional attitude more closely linked to communication than is speaker-meaning.

[37] This point is made in Davis (1999).

to be a sense of the verb "tell" which suits the Gricean analyses more closely.[38] "Tell" is not a success verb, and there is something strange about describing the statement of the conclusion of an argument as a case of telling an audience something. Nevertheless, it does seem that, if I give an argument for a proposition p and you reported me as having told my audience p, you would have said something which is, though admittedly a bit incongruous, true. Further problems come from the special class of cases of reminding in which the information of which the audience is to be reminded is on the tip of the audience's tongue; in the case above, it seems clear that I told my audience that his friend's name is John, even though I did not satisfy the conditions set by any version of the Gricean analysis.

Moreover, whether or not the propositional attitude of speaker meaning distributes over conjunction, it is clear that communicating, telling, and informing all do. But then it is clear that there is a modified version of the general argument given in §2.2.4 above which will refute the Gricean analysis when considered as an analysis of communicating, telling, or informing. Indeed, the argument is much stronger in these cases than it was in the case of speaker-meaning.

2.2.6 Two interpretations of Gricean accounts

Suppose that this conclusion is correct: there is no account of any action-based propositional attitude verb in terms of facts about mental content. This need not be the end of the indirect mentalist strategy. There are two ways to regard Grice's analysis of speaker-meaning: either as an account of the class of facts corresponding to true sentences of the form ⌜By uttering x, α meant that σ⌝, or as a stipulative definition of a technical notion of S-meaning, which includes some but not all facts about speaker-meaning. The fact that Grice and his followers were concerned to revise their account in response to counterexamples shows that, for the most part, the Gricean analyses of speaker-meaning have been thought of as an analysis of the pre-theoretic notion of speaker-meaning. On this interpretation, the Gricean account fails; but the Gricean can still retreat to viewing her analyses as stipulative definitions of S-meaning, and try to use this stipulatively defined propositional attitude to give an account of linguistic meaning. This would still yield an account of linguistic meaning which is Gricean, if not in letter, then at least in spirit.

To recap, this stipulative definition of the propositional attitude of S-meaning runs as follows:

> a S-means p by uttering $x \equiv a$ intends in uttering x that
> (1) his audience come to actively believe p,
> (2) his audience recognize this intention, &
> (3) (1) occur on the basis of (2)

The main consequence of this retreat is that the Gricean account of speaker-meaning does not succeed in making the class of facts about what speakers mean by their utterances serviceable for use in a mentalist account of language. For this reason, it is not sufficient for the mentalist to give an account of linguistic meaning in terms of, for example, conventions governing what speakers mean or would mean by certain utterances; such an account will have to be in terms

[38] Schiffer (1982) suggests that a version of the Gricean analysis might be an appropriate analysis of telling.

of conventions governing what speakers S-mean, or would S-mean, by certain utterances. And the preceding examples show that these are importantly different: the mentalist will have to give an account of linguistic meaning in terms of a sub-class of the possible uses of language, which does not include cases of persuasive discourse, the use of language in calculation, and the other instances of speaker-meaning not counted as such by the Gricean account. In this sense, the Gricean slogan that "the meaning (in general) of a sign needs to be explained in terms of what users of the sign do (or should) should mean by it on particular occasions; and so the latter notion is in fact the fundamental one"[39] is, even if true, irrelevant to the mentalist project of reducing linguistic meaning to mental content. And this is because, as we have seen, action-based propositional attitudes like speaker-meaning, assertion, and communication are partially derived from the facts about linguistic meaning we are now trying to explain.[40]

What philosophical motivations might the switch from analysis of speaker-meaning to stipulative definition of S-meaning have? Some neo-Griceans have admitted that the class of facts about S-meaning is narrower than the class of facts about speaker-meaning, but have claimed that instances of speaker-meaning which are not cases of S-meaning are, in some sense, derivative uses of language.[41] The problem for these revisions of the Gricean program has always been in spelling out exactly what this conceptual dependence, or derivativeness, amounts to, and in substantiating these claims.

The present understanding of the mentalist program provides one way of making these claims explicit. It may be the case that (i) we can give an account of linguistic meaning in terms of facts about S-meaning, and (ii) the cases of speaker-meaning which are not cases of S-meaning may then be accounted for in terms which appeal to facts about linguistic meaning. Were this the case, the dissociation of S-meaning from speaker-meaning would be only a small setback for the cause of mentalism, and we would have good grounds for agreeing with the neo-Gricean claim that the recalcitrant cases of speaker-meaning are, in some sense, derivative from cases of S-meaning.[42]

At this point, it is worth noting an ambiguity in our formulation of mentalism. So far I have been glossing mentalism as the view that we can give an account of the meanings of expressions of public languages in terms of facts about some class C of propositional attitudes of speakers of the language, and that we can give an account of the propositional attitudes in C without appeal to facts about the contents or meanings of representations in any language. This formulation of mentalism can indeed survive through the move from analysis of speaker-meaning to stipulative definition of S-meaning. A stronger mentalist claim, however, must be given up. One might want to claim that this foundational class C of propositional attitudes can be accounted for in non-linguistic terms, but that *all* propositional attitudes may be given a

[39] Grice (1957), 217.

[40] I try to say more about the sense in which these attitudes are derived from facts about linguistic meaning in Appendix C, pp. 213 ff. below.

[41] Different versions of this thesis are presented in Clark (1975), Bennett (1976), Suppes (1986), and Bar-On (1995). For some leanings in this direction by Grice himself, see Grice (1987), 357.

[42] There is another way of spelling out this claim. The neo-Gricean mentalist might take the following course: (i) stipulatively define a notion of S-meaning, (ii) give an account of speaker-meaning in terms of S-meaning, and (iii) go on, as usual, to give an account of linguistic meaning in terms of speaker-meaning. An advantage of this is that, as we'll see in §2.3, it avoids some of the problems which face the neo-Gricean move sketched above. This may be a promising route, though I don't see how (ii) could be carried out. Thanks to Jonathan Beere for suggesting this possibility.

constitutive account without bringing in facts about the contents of linguistic representations. We have now seen that the mentalist must give up this strong claim, and in so doing, endorse a bit of communitarianism: she must admit that, in the case of some action-based propositional attitudes (speaker-meaning, assertion, telling, etc.) an agent's bearing one of these attitudes toward a proposition is sometimes explained by reference to a sentence with a meaning which is used by the agent in meaning, or asserting, or communicating something by an utterance.[43]

One might have the following worry about the path I have suggested for the mentalist: most mentalist accounts of linguistic meaning on offer employ the notion of speaker-meaning in the analysans; even if these accounts are true as they stand, it is not at all clear that changing each reference to speaker-meaning to a reference to S-meaning will preserve truth, especially since S-meaning will be the narrower notion. As we'll see in the next section, this doubt is a serious one; there is no true account of linguistic meaning in terms of what speakers do or would S-mean by their utterances.

2.3 Convention and linguistic meaning

Since the publication of David Lewis's *Convention* in 1969, it has become standard practice for mentalist proponents of the indirect strategy to try to account for facts about the meanings of linguistic expressions in terms of conventions governing the use of those expressions by speakers to mean things.[44] Although, as in the case of the Gricean analysis of speaker-meaning, there have been many variations on this approach, I hope to stick with fairly simple versions of convention-based accounts of linguistic meaning, and to present objections which, insofar as they are effective, apply to other variants as well.[45]

Before going on, it should be noted that there are many different uses to which a notion of convention similar to the ones discussed below may be put; the present role of a link in a mentalist account of meaning in terms of mental content is only one of them. Lewis, for example, is out largely to rehabilitate the thesis that language is conventional, and thereby to provide an answer to one of Quine's doubts about analyticity; one might also be interested simply in an analysis of our everyday notion of convention. This section is an argument for the thesis that the notion of convention will not help the mentalist reduce linguistic meaning to mental content; I leave it open whether the notion of convention, defined either as by Lewis or by those who came after him, might be a useful tool for accomplishing some of these other ends.

[43]Now, the mentalist might object that I have so far only argued that Grice's picture of speaker-meaning fails; it might yet be that another succeeds. This is indeed an open possibility so far; in my view, it is not a very promising one. For a discussion of the reasons why, see the discussion of the views of John Searle and Christopher Peacocke in Appendix B below.

[44]See, e.g., Schiffer (1972), (1982); Bennett (1973), (1976); Loar (1976), (1981). Indeed, Grice himself is virtually the only Gricean not to endorse this approach as the right way to understand linguistic meaning. For a brief discussion of Grice's proposals, see footnote 14 in §3.2 below. Note also that Lewis himself does not endorse the variant of the convention-based approach discussed here; his views are part of the subject of the next chapter.

[45]More precisely, the versions of the account I will consider will each have an analysans weaker than that of any of the fully developed analyses of convention which have been put forward. My arguments will all be that these analyses fail to provide necessary conditions for linguistic meaning; if these arguments go through, they will also apply to accounts which have a stronger analysans.

The simplest way to present the thesis of the convention-theorist is to define convention in terms of mutual knowledge, where a and b mutually know p just in case a knows p, b knows that a knows p, a knows that b knows that a knows p, and so on, and vice versa.[46] Using this notion, we can state the view of the convention-theorist as follows:

[C] x means p in a population $G \equiv$
(1) almost all members of G utter x only when they mean p by uttering x,
(2) almost all members of G mutually know (1), &
(3) (1) obtains because of (2)[47]

But, as argued in §2.2, there is no true explication of facts about speaker-meaning in terms of mental content; indeed, we saw good reason to think that what speakers mean by their utterances is partly constituted by, rather than constitutive of, the meanings of the sentences they utter.[48] For this reason, these facts can't be employed by the mentalist in his account of linguistic meaning, if the goal is ultimately to reduce linguistic meaning to mental content. So we need to replace mention of speaker-meaning with the stipulatively defined notion of S-meaning, which of course *is* (stipulatively) defined in terms of mental content.

That is, we should change the first clause of the above account to read as follows:

(1*) almost all members of G utter x only when they S-mean p by uttering x

But this raises an immediate problem: changing clause (1) of the account also changes the contents of clauses (2) and (3); on our new account, clause (2) requires that almost all members of G mutually know that almost all members of G utter x only when they S-mean p, and so attributes to speakers a belief part of whose content is the relation of S-meaning. But the relation of S-meaning is a technical notion which was just defined a few pages back; what grounds are there for saying that speakers must have such beliefs? There may well be no expression of their language which they understand which expresses this relation; and it is hard to see how they could have acquired the belief through non-linguistic channels. This seems to me a decisive objection to our revised theory.[49]

[46] This is the route of Schiffer (1972), ch. 5.

[47] I ignore complications introduced by indexicality and ambiguity. Showing that the notion of convention cannot be used to give an account of the meanings of expressions in simplified languages will suffice to show that it cannot serve the mentalist's purposes for languages which do include indexicals either; inclusion of indexicality and ambiguity makes the mentalist's job harder rather than easier. None of the objections I'll raise against convention-based analyses of linguistic meaning will turn on either issue.

I state the account as giving the meaning of an expression in a population rather than in a language because it has become standard, following David Lewis (1975), to treat languages as abstract objects: functions from expressions and contexts to contents. Given this conception of languages, foundational questions about meaning should not be thought of as asking what it is for an expression to have a given meaning in a language, but rather what it is for an expression to have a given meaning in a population.

[48] For a further defense of this claim, see Appendix C below.

[49] Note that this argument against the revised version of [C] can't be defused by appeal to Lewis's distinction between knowledge *in sensu composito* and knowledge *in sensu diviso*. (See Lewis (1969),

To this problem, it might seem that the convention-theorist has a ready answer. The failure of Gricean accounts of speaker-meaning bars him from appealing to facts about speaker-meaning in clause (1) of his account; but there is no reason why he cannot appeal to facts about what speakers *know* about what other speakers mean by their utterances. This is, effectively, to replace clause (1) with (1*), while leaving (2) with the content it had in [C].[50] Here the only reference to speaker-meaning is within the scope of "mutually knows" in clause (2). Since no mention of this relation occurs outside the scope of a propositional attitude verb, this formula should be acceptable to the mentalist, if true; it is an analysis of facts about linguistic meaning in terms of facts about what speakers S-mean by utterances and about what speakers know. Both of these are, or are analyzed in terms of, facts about mental content.

In fact, though, the mentalist cannot appeal to what speakers know about what other speakers mean by utterances any more than he can appeal to speaker-meaning simpliciter; for to make use of the former is also to make use of the latter. Any account of what it is for an agent to know p will involve p itself, outside the scope of any propositional attitude verb. For example, the traditional analysis of knowledge as justified true belief may be stated as follows:

> a knows $p \equiv a$ believes p, a is justified in this belief, and p

But, if some account of knowledge which has this feature turns out to be true — and it is hard to imagine an account which would not have this feature — then any account of linguistic meaning which appeals to facts about knowledge of speaker-meaning will, de facto, appeal also to facts about speaker-meaning themselves. And this is just what the mentalist, for familiar reasons, cannot do.[51]

64-8, and Loar (1976), 157-9.) The difference between these two interpretations of clause (2) is as follows:

> Knowledge *in sensu composito*: almost all members of population G mutually know that: $\forall x \forall p$ (x means p in $G \rightarrow$ almost all members of G utter x only when they S-mean p)

> Knowledge *in sensu diviso*: $\forall x \forall p$ (x means p in population $G \rightarrow$ almost all members of G mutually know that: almost all members of G utter x only when they S-mean p)

In the text, I describe the knowledge attributed to speakers as knowledge *in sensu composito*; and one might object that only the weaker attribution of knowledge *in sensu diviso* is required. But, even if we restate (2) so as to attribute only knowledge *in sensu diviso*, the objection above stands. The problem was not that it is implausible to attribute to speakers mutual knowledge of a general claim about all the sentences of their language, but rather that speakers need have no knowledge, whether particular or general, involving the technical concept of S-meaning.

[50] More formally, the account would be as follows:

x means p in a population $G \equiv$

(1) almost all members of G utter x only when they S-mean p by uttering x,

(2) almost all members of G mutually know that almost all members of G utter x only when they mean p, &

(3) (1) obtains because of (2)

[51] One natural move would be to replace the notion of mutual knowledge with that of mutual belief, defined in the obvious way. I'm not sure whether this works. One worry is that it may be too

So it does seem that the move from speaker-meaning to the narrower notion of S-meaning poses serious problems for the mentalist following the indirect strategy laid out in this chapter. But this is not, it seems to me, the fundamental problem with convention-theoretic approaches to linguistic meaning. In my view, the fundamental problem with this view of meaning is that it substantially over-intellectualizes language use.

To see this, recall the argument given above against the theory obtained by replacing every mention of speaker-meaning in [C] with a mention of S-meaning. The argument there was that, given that S-meaning is a technical notion known only by a subclass of those people who know something about the philosophy of language, it is bizarre to think that beliefs about what others S-mean by their utterances should be a necessary condition on membership in a linguistic community. But, once we see this, is it any more plausible to claim that it should depend on mutual knowledge of facts about what speakers do or would *mean* by their utterances? It only seems so to us because we happen to speak a language in which a commonly used predicate expresses the propositional attitude of speaker-meaning; but surely a meaningful language could develop which lacked this characteristic. As applied to speakers of such a language, the requirement that speakers have beliefs about what other speakers would *mean* by their utterances is surely as unmotivated as the requirement that we have beliefs about what other speakers would *S-mean* by their utterances.[52] In short: the convention-theorist requires more knowledge of speakers than is plausibly necessary.[53]

weak; members of a population might be deceived about who else is in the population, and have no justification for their beliefs.

A further problem is that, as argued above and in Appendix C, facts about speaker-meaning are constituted partly by, and hence derivative from, facts about linguistic meaning. But the convention-theorist we are considering takes facts about what speakers *know* about facts about speaker-meaning to be constitutive of, and hence prior to, facts about linguistic meaning. The problem is that the relations of constitution and priority seem clearly to be transitive; and, if this is right, it follows that the mentalist under consideration is committed to the claim that speakers' knowledge of facts about speaker-meaning is constitutive of and prior to facts about speaker-meaning themselves. But this claim seems incoherent; how could it be the case that, for some fact F, what it is for F to obtain is for certain individuals to *know* that F obtains? On the contrary, it seems clear that F must obtain in order for anyone to know that it obtains.

[52]Antony Eagle has suggested that the stable signalling strategies discussed in Skyrms (1996) may be examples of quasi-languages which develop without this sort of mutual knowledge.

[53]It is also worth noting a reply to the objection from over-intellectualizing language use which, while not entirely implausible, is not open to the mentalist. The reply is to take a kind of deflationary view of certain mental states by taking them to be constituted by certain sorts of linguistic behavior which are related via a public language to those mental states. An example of this sort of view is the link that is standardly supposed to obtain between understanding an expression (of a language), and knowing the meaning of the expression (in the language). One might think that to have the meta-linguistic knowledge that expressed by "'S' means p", it is sufficient that (i) one satisfy certain communal standards with respect to "S", and (ii) for "S" to mean p. Just so, one might think that, for an agent to know that a speaker meant p by her utterance, it is sufficient that (i) she be able to respond in generally appropriate ways to the utterance, and (ii) that the speaker in the context meant p by her utterance. I call these views 'deflationary' because they are views according to which a certain kind of knowledge comes cheap. The reason why this kind of deflationary view is not open to the mentalist trying to give a convention-theoretic account of meaning should be clear: the account in question explains what it is for the agents in question to have the knowledge partly in terms of the facts they have knowledge of. But these facts are not independent of linguistic meaning, and hence

This brings to a close the discussion of the indirect mentalist strategy of giving an account of some action-based propositional attitude verb in terms of mental content, and then using facts about that action-based propositional attitude to give an account of linguistic meaning. The first half of the strategy can't be carried out; facts about assertion, speaker-meaning, and communication are derived from linguistic meaning, rather than the other way around. And, partly for this reason, the attempt to give a mentalist account of linguistic meaning using the notion of convention faces problems of principle in even getting off the ground.[54]

Given this result, it is worth asking whether the mentalist ought to set aside action-based propositional attitudes and the indirect strategy, and seek instead an account of linguistic meaning directly in terms of the beliefs of agents. In the next chapter, I consider two attempts to do just that.

are not available to the mentalist.

[54] As mentioned briefly above, there is one other way of carrying out the indirect mentalist strategy which I haven't discussed: Grice's attempt to give an account of linguistic meaning using the notion of a 'resultant procedure' of speakers for meaning things by their utterances in his (1968). I discuss this briefly below in connection with David Lewis's theory of meaning (p. 49, note 14).

Chapter 3

Meaning and Belief

Contents

3.1 Ramsey on meaning and belief

Although most of the work done on mentalist accounts of linguistic meaning has employed the indirect strategy typified by the Gricean account — that is, explicating linguistic meaning in terms of an action-based propositional attitude, and that action-based attitude in terms of mental content — there is another, more direct route, which has received little attention.[1] This is the approach to linguistic meaning suggested by Frank Ramsey in his 1927 paper, "Facts and Propositions."

Ramsey characterized his view as a form of pragmatism, and summarized it as follows:

> The essence of pragmatism I take to be this, that the meaning of a sentence is to be specified by reference to the actions to which asserting it would lead, or, more vaguely still, by its possible causes and effects.[2]

Discussions earlier in the paper indicate that when he linked the meaning of a sentence with the causes and effects of asserting a sentence, Ramsey had in mind especially the propositional attitudes — in particular, beliefs — which a competent speaker would have by virtue of asserting a sentence. Though we should keep in mind that Ramsey is not explicitly pursuing a reduction of meaning to the mental, and that he nowhere offers an explicit theory about what it is for a sentence to have meaning, Ramsey's remarks suggest giving an account of the meaning of a sentence in terms of the content of the belief a speaker would acquire by accepting that sentence in that context. On the face of it, this seems to deliver the result the mentalist sought from the Gricean account of meaning, and far more directly.

[1] An exception is a mention in Loar (1981).
[2] Ramsey (1927), 51.

It is important to see the philosophical differences between this sort of account and the convention-theoretic account of the last chapter. Though superficially similar, they present very different pictures of the nature of linguistic meaning. Recall that one of the basic worries about the convention-theorist's account was its tendency to over-intellectualize membership in a linguistic community: its tendency, that is, to make such membership depend upon mutual knowledge of various more or less recondite facts about what speakers would mean (or S-mean) by uttering various sentences. Ramsey's account stands in stark contrast to this; in Ramsey's view, meaning depends not on the beliefs of speakers about other speakers' use of language, but simply on correlations between the linguistic usage and beliefs of speakers of a language. This turn away from the over-intellectualizing tendencies of the indirect strategy of the last chapter is part of the shift from Grice's account of meaning in terms of what effects speakers intend to bring about by utterances, to Ramsey's account of what effects utterances actually have.

Problems formulating the account

How are we to formulate a Ramseyan account of speaker-meaning? Again abstracting away from indexicality and ambiguity, the following may seem to be the right formulation:

> (x means p in a population G) $\equiv \forall a$(a is a member of $G \rightarrow$ (a accepts $x \rightarrow a$ believes p))

But if the "$\rightarrow$" connective in the consequent of the right-hand side of this biconditional is understood as the material conditional, then the formula is clearly false: if x is a sentence which no member of G accepts, then for *any proposition* p, the formula entails that x means p in G. But of course a sentence may be a meaningful sentence of a language, even if every speaker of the language thinks the sentence false.

The natural move is to solve this problem by making the consequent of the right-hand side into a counterfactual conditional. Using the connective "$\Box\!\!\rightarrow$" to represent the counterfactual conditional, this yields the following:

> [R] (x means p in a population G) $\equiv \forall a$(a is a member of $G \rightarrow$ (a accepts x $\Box\!\!\rightarrow a$ believes p))

But this faces a similar problem. Suppose that every member of our linguistic community believes that the sentence "0=1" is false. Is it true that, were we to accept "0=1", we would thereby believe that 0=1?

This seems not to be true on a possible worlds interpretation of counterfactuals. The world most similar to the actual world in which I would accept this sentence is one in which it has a different meaning. Surely, after all, there is a world in which this sentence means something different — something which might be believed to be true by someone who understands the sentence — which is more similar to the actual world than any in which I believe that 0=1.[3]

One might reply that this is a misinterpretation of the counterfactual: the worlds relevant to the evaluation of the counterfactual are restricted to those in which the agents in question

[3]Note that it is a necessary condition on accepting a sentence that I understand the sentence; the attitude of acceptance must be interpreted this way if [R] is to be true. More on this below.

are members of the population G, and a world in which "0=1" has a meaning different than its actual meaning in my language is one in which, in the relevant sense, I fail to be a member of G. But this is just to require that the meaning of "0=1" be held constant between the actual world and the possible world(s) relevant for evaluation of the counterfactual; and this makes the account circular, since it builds facts about linguistic meaning into an account of what it is for a linguistic expression to have a meaning.[4]

The problem to which this mathematical example points is that an agent can't just believe whatever she likes, and so can't arbitrarily accept a sentence which she understands. Because our language is not constrained in such a way that only believable sentences are meaningful, some sentences express propositions which are so clearly false that anyone who understands the sentence — and thus is eligible to accept it — would reject it. Once we try to imagine what the agent *would* believe were he to accept it, we must, to get the right result, stipulate — whether implicitly or explicitly — that the meaning of the sentence remain fixed. And this makes the proposed analysis of linguistic meaning circular.

Meaning, belief, & fineness of grain

A second problem with this Ramseyan account arises from the fact that facts about the meanings of sentences are, in a certain sense, more fine-grained than facts about the beliefs of agents.

As noted above, belief seems to distribute over conjunction; that is, it is a necessary consequence of the truth of a sentence of the form $\ulcorner\alpha$ believes that σ and $\sigma'\urcorner$ that, in that context, the following sentences are also true: $\ulcorner\alpha$ believes that $\sigma\urcorner$ and $\ulcorner\alpha$ believes that $\sigma'\urcorner$.[5] Now consider a conjunctive sentence x whose meaning may be represented as the conjunctive proposition $< AND < p, q >>$.[6] Then each of the following claims are true:

$\Box\forall a$ (a accepts x in $C \rightarrow a$ believes p & q)

$\Box\forall a$ (a accepts x in $C \rightarrow a$ believes p)

[4]It is important to be clear about what kind of circularity is involved here. Strictly, the account needn't build in facts about which propositions are the contents of which sentences; rather, it must build in facts about which sentences *mean the same thing as* or *mean something different than* other sentences. The argument assumes that this kind of circularity is as vicious as the more straightforward circularity of including facts about meaning in the analysans.

Nor, I think, will more complex counterfactuals solve the problem. Consider the following replacement for the simple counterfactual used in the consequent of the right-hand side of [R]: were a to believe p and accept x, then, were it not the case that a accepted x, it would not have been the case that a believed p. Intuitively, the suggestion is that we consider the nearest possible world w in which a believes p and accepts x; we then consider the world $w*$ most similar to w in which a does not accept x, and compare the beliefs of a and w and $w*$ to arrive at the meaning of x. The meaning of x will be the content of that belief which a has at w but lacks at $w*$. Here essentially the same problem arises. On pain of circularity, we can't stipulate that that x maintain its actual meaning at w; but if x might have had a different meaning at w, there is no reason to believe that the difference in a's beliefs between w and $w*$, if such there be, should be revealing of the actual meaning of x. Thanks to Caspar Hare for suggesting this possibility.

[5]As above, I simplify by ignoring the need to relativize claims about meaning to contexts of utterances.

[6]Here "p" and "q" are used as names of propositions.

$\Box\forall a$ (a accepts x in $C \rightarrow a$ believes q)

The problem is that it follows from [R], along with the truth of these three claims, that x is three ways ambiguous, as used in C: x means p & q in C, x means p in C, and x means q in C. But this is incorrect: conjunctive sentences are not, in general, ambiguous in this way. To use the example mentioned above, "It is raining and arithmetic is incomplete" does not mean that arithmetic is incomplete.

While this does show that this account is false, it seems that there is an easy way to revise it in answer to this objection. We can limit the account of meaning to simple sentences — or, better, to non-conjunctive sentences — and give a separate account of the meanings of conjunctive sentences in terms of the meanings of their parts.[7] A disadvantage of this move is that it leaves the mentalist with the task of giving an account of what constitutes the fact that "and" means what it does; but perhaps there will be some way of doing this.[8]

But this idea does not go far enough; there are counterexamples similar in form to those based on the distribution of belief over conjunction even for simple sentences. To come up with such a counterexample, one needs only two non-conjunctive propositions p, q such that it is a necessary consequence of an agent a's believing p that a also believes q. If there are such propositions, then our account will be falsified by any sentence x and context C such that x means p in C and x does not mean q in C.

The most obvious candidates are cases in which one of the propositions is a trivial consequence of the other. Consider, for example, the proposition that Lassie was a brown dog and the proposition that Lassie was a dog. Is it possible for an agent who accepts, and therefore understands, the sentence "Lassie was a brown dog," and who believes that Lassie was a brown dog, not to believe that Lassie was a dog? I have not been able to think of any situations in which this would plausibly be the case. If there are no such cases, then our account entails that "Lassie was a brown dog" is ambiguous between meaning that Lassie was a brown dog and meaning that Lassie was a dog; this is clearly incorrect. Because "Lassie was a brown dog" seems to be a fairly ordinary simple sentence in every respect, it is unlikely that any restriction of the class of sentences for which our account should hold will be effective in ruling this sort of case out.

Another such class of cases exploits the trivial equivalence — paradoxical cases aside — of a proposition p and the proposition which attributes truth to p. Suppose that an agent accepts the sentence "It is true that Lassie was a dog", and so by weak disquotation believes that it is true that Lassie was a dog. Is it possible that such an agent could yet fail to believe that Lassie was a dog? Again, it seems unlikely.

It is tempting to try to solve this problem by appealing to a distinction between beliefs

[7] I.e., we should give the following sort of account:

$\Box\forall x\forall p\forall G$ {x is not a conjunction $\rightarrow$ [(x means p in a population G) $\equiv$
$\forall a$(a is a member of $G \rightarrow$ (a accepts x $\Box\!\!\rightarrow a$ believes p))]}

[8] A natural thought is that we can explain the meaning of "and" in its role as a sentence connective in terms of its role as a word which joins two predicates to form a single predicate, and so follow what Evans (1977) calls the Tarskian approach to giving a semantics for "and." But this is no help in the present context, for a variant of the problem with compound sentences arises with compound predicates. E.g., it seems plausible that, necessarily, if a believes that Socrates was a Greek and a philosopher, then a believes that Socrates was a Greek.

gained 'directly' as a result of accepting a sentence, and beliefs gained 'indirectly.'[9] In one sense, this is clearly correct; if x means p in C, then there is a more direct link between the acceptance of x in C by a and a's believing p than between a's acceptance of x and a's coming to believe any other propositions. But it seems to me that the only way to make sense of this distinction between beliefs directly and indirectly arrived at on the basis of acceptance of sentences will be in terms of the meaning of the sentence accepted; and, if this is right, then the proposed restriction is useless from the point of view of the mentalist. For how else could the distinction be made out? It seems unlikely that, on any account of causation, a's believing p is the cause of a's believing q rather than the reverse; it is not as though a comes to believe p and then, a moment later, comes to believe q. Rather, it is a fact about belief that the former is metaphysically sufficient for the latter. And if the distinction is not to be made in causal terms, then I don't see any way to draw it which does not rely on facts about the meanings of sentences.

This is a serious problem for the mentalist who seeks to give a Ramseyan reduction of linguistic meaning to mental content. The idea was to give an account of linguistic meaning by exploiting correlations between sentences accepted, their meanings, and the beliefs of agents. But facts about the distributions of beliefs among agents in a population may be such as to make these correlations ill-suited to serve as the basis for an account of linguistic meaning. More specifically, we've seen that there are some necessary connections between the beliefs of agents; for some distinct propositions p, q it is a necessary consequence of the fact that an agent believes p that that agent believes q as well. The problem is that an analogous fact does not hold for linguistic meaning; it is never the case that for distinct propositions p, q it is a necessary consequence of the fact that some sentence S means p that S also means q. In this sense, assignments of meanings to sentences are more fine-grained than assignments of beliefs to agents; and this makes it difficult to see how we could give an account of the former in terms of the latter.

The analysis of understanding

A third basic objection to the Ramseyan approach to meaning is that it rules out a certain kind of view of understanding, which is a very natural one for the mentalist to adopt. First, note that if some variant of the Ramseyan account is to be true, then "accepts" must be interpreted in such a way that it is a necessary condition on an agent accepting a sentence that the agent understand the sentence. Were it not so interpreted, and thought of instead as something like expressing approval of a sequence of marks or sounds or believing of a sentence that it is true, then the account would be falsified by cases in which an agent accepts a sentence without understanding it, because, in such cases, the agent would not come to acquire the relevant belief.

What is it to understand a sentence? The mentalist takes the meanings of public language expressions to be fixed by the propositional attitudes of users of the language; hence it will be very natural for her also to claim that agents' understanding of those expressions is, similarly, to be explained by the propositional attitudes of agents. In particular, the mentalist is likely to

[9] This might be stated as follows:
$\Box\forall x\forall p\forall G$ [(x means p in a population G) $\equiv \forall a(a$ is a member of $G \rightarrow$
(a accepts x $\Box\!\!\rightarrow a$ directly comes to believe p))]

think that an agent's understanding of a sentence is to be explained by that agent's *knowledge* of the meaning of that sentence. So, given the discussion of acceptance above, this means that a proponent of the Ramseyan account who endorses this sort of 'cognitivist' view of linguistic understanding will appeal to knowledge of linguistic meaning in her account of linguistic meaning.

But now recall the discussion of knowledge in §2.3 above: the problem is that, eventually, the mentalist who takes this route will have to give an account of what it is for an agent to know p, and any account of what it is for an agent to know p will involve p itself, outside the scope of any propositional attitude verb. This is a problem in the present case because p — the object of the agent's knowledge — is a fact about linguistic meaning. This poses a problem for the mentalist since our story about what it is to understand a sentence — and hence also our story about what it is to accept a sentence — will make essential reference to the fact that the expression has the meaning that it does. And this is precisely what the mentalist must avoid.[10]

There is one other worry about formulations like [R], which is a bit hard to put clearly. The worry is that it seems as though a mentalist use of this sort of account of meaning — one which takes facts about belief to be constitutive of facts about the meanings of sentences, rather than the other way around — gets the intuitive order of explanation wrong. It seems natural to say that John believes p *because* or *in virtue of the fact that* he accepted x in C, and x means p in C; it seems most unnatural to say that x means p in C in virtue of the fact that any agent accepting x in C would thereby come to believe p. Ramsey seems to have shared this view about explanatory priority; he writes in one place that the sense in which a predicate 'R', as used in a sentence 'aRb', "unites 'a' and 'b' then *determines* whether it is a belief that aRb or that bRa."[11] But, as discussed at length above, intuitions about explanatory or conceptual priority are pretty vague and difficult to pin down; given this, the point made in this paragraph should be regarded less as an objection to the Ramseyan mentalist than as expressing the worry that an account of meaning and content which employed some version of [R] might not quite answer to the mentalist claim that mental representation is prior to, and constitutive of, linguistic meaning.[12]

[10] Further, we will again have a case in which the mentalist is committed to thinking that some class of facts — facts about linguistic meaning, in this case — is constituted by knowledge of this very class of facts. And, as argued above, this seems incoherent.

[11] Ramsey (1927), 41. The italics on "determines" are mine.

[12] It is worth noting, furthermore, that no accounts like [R] which account for linguistic meaning on the basis of the results of accepting sentences can be the whole story about what it is for an expression to have a meaning. [R], after all, is an account only of the meanings of whole sentences in contexts, and whole sentences are not the only signs which have meanings: sub-sentential expressions have meanings as well. There is no way to expand [R] to yield an account of the meanings of such expressions, since we do not accept such expressions as we accept sentences, and do not have beliefs whose content is anything other than a whole proposition. So, even if some version of [R] can be defended, the mentalist still owes an account of the meanings of expressions in terms of the meanings of whole sentences. I haven't raised any objections against the possibility of giving an account like this; the point here is just to note that it is a further project. See Appendix D for more discussion.

3.2 Lewis on conventions of truthfulness and trust

One natural response to these problems is to think that, while Ramsey's view cannot serve as a constitutive theory of meaning, the basic idea behind it — giving an account of meaning on the basis of correlations between linguistic behavior and belief — might yet be correct. One way to develop this thought is to embed Ramsey's central claim in a more sophisticated theoretical structure. A natural choice for such a structure is one we have already discussed: that of *convention*.

A good way to understand the view of meaning developed in David Lewis's seminal book *Convention* is as giving just this kind of development of Ramsey's thought. Indeed, his view is, in its essentials, closer to Ramsey's view than to that of the convention-theorist discussed in the previous chapter, who focuses on conventions governing what speakers mean (or S-mean) by their utterances. But, as I shall argue, Lewis's account is open to substantially the same objections as was Ramsey's.

Lewis takes the relevant conventions to be conventions of truthfulness and trust in a language.[13] His account may be adapted to state an analysis of a sentence having a given meaning in a population as follows:

[L] x means p in a population $G \equiv$
- (1a) ordinarily, if a member of G utters x, the speaker believes p,
- (1b) ordinarily, if a member of G hears an utterance of x, he comes to believe p, unless he already believed this,
- (2) members of G believe that (1a) and (1b) are true,
- (3) the expectation that (1a) and (1b) will continue to be true gives members of G a good reason to continue to utter x only if they believe p, and to expect the same of other members of G,
- (4) there is among the members of G a general preference for people to continue to conform to regularities (1a) and (1b)
- (5) all of these facts are mutually known by members of G[14]

As regards the mentalist task of reducing linguistic meaning to mental content, Lewis's use of the notion of convention is an improvement over the more common version discussed in §2.3: Lewis's account, if true, gives an account of linguistic meaning directly in terms of the beliefs and knowledge of speakers of a language, without employing the sorts of action-based propositional attitudes which posed so many problems for the indirect Gricean strategy.

However, the first objection to Ramsey's account of meaning arises again here in connection

[13] See especially Lewis (1975), 166-169.

[14] I should note that I have revised Lewis's account in several ways, one of which is substantive. Lewis's original account contained the following six clauses, from which I have, in version [L] above, dropped the fifth:

[L*] x means p in a population $G \equiv$

with both of clauses (1a) and (1b); I'll focus on (1b) in what follows. Clause (1b) requires that, if a sentence S means p in a population G, then, ordinarily, if a speaker a of G hears an utterance of S, a will come to believe p, if he didn't already. For the reasons discussed above, this must be interpreted as a counterfactual: if a speaker a were to hear S, then, ordinarily,

(1a) ordinarily, if a member of G utters x, the speaker believes p,
(1b) ordinarily, if a member of G hears an utterance of x, he comes to believe p, unless he already believed this,
(2) members of G believe that (1a) and (1b) are true,
(3) the expectation that (1a) and (1b) will continue to be true gives members of G a good reason to continue to utter x only if they believe p, and to expect the same of other members of G,
(4) there is among the members of G a general preference for people to continue to conform to regularities (1a) and (1b)
(5) there is an alternative regularity to (1a) and (1b) which is such that its being generally conformed to by some members of G would give other speakers reason to conform to it
(6) all of these facts are mutually known by members of G[15]

I omit clause (5) from the above account, because I think that, as Tyler Burge has argued in his "Knowledge and Convention" (1975), this clause makes Lewis's conditions on linguistic meaning too strong. Burge pointed out that (5) need not be mutually known by speakers for them to speak a meaningful language. Consider, for example, the case in which speakers believe that there is no possible language other than their own, and hence that there is no alternative regularity to (1a) and (1b). They might, for example, have encountered no other groups of people speaking different languages, and may believe that their language was handed down to them directly by God. But surely the fact that they have false beliefs about the origin and status of their language does not stop the expressions of their language from having meaning.

Burge uses a similar example to show that (5) not only need not be mutually known, but also need not even be true. Suppose that other speakers in the linguistic community described above came to obey other regularities than (1a) and (1b). Speakers of that language might come to think that these others were committing a mortal sin; they would take themselves to have no reason to conform to the alternative regularity. Admittedly, these are not mainstream cases. But the fact that, on Lewis's original account, false beliefs about the nature of language and resistance to linguistic change could deprive expressions of an apparent language of their meaning seems sufficient reason to drop clause (5) from the above account, leaving only the clauses listed in the main text. The criticisms of Lewis's account in the text are independent of this problem.

I should also note, because I don't discuss this sort of account of linguistic meaning explicitly in the main text, that the sorts of cases Burge discusses also seem to show that Grice's account of expression meaning is false. Roughly, Grice suggests that a sentence S means p in the language of a group G if and only if both of the following conditions are met: (i) some or many of the members of members of G have the procedure of uttering S when they mean p, and (ii) their retention of this procedure is conditional one other members of G doing the same (see especially Grice (1968), 127, definition D3). The second sort of Burge counterexample to Lewis shows that (ii) is not a necessary condition on expression meaning. Of course, many of the problems discussed above in connection with theories which try to give an account of meaning in terms of conventions governing what speakers mean by utterances arise here as well.

a would come to believe p. But, as above, we run into problems when S is a sentence, like "0=1", which is such that almost everyone who understands it knows it to be false. It is simply not true that, were most competent speakers to hear an utterance of this sentence, they would come to believe that 0=1. Substantially the same problem arises with (1a).

In response to this objection, a proponent of [L] might revise (1b) to require only that, if a speaker of the language were to *accept* "0=1", then ordinarily that speaker would believe that 0=1. But, as we saw above in the discussion of Ramsey's views, this won't solve the problem, for it seems that we face the following dilemma. On the one hand, we might restrict the possible worlds relevant to evaluation of the counterfactual to those in which "0=1" retains its actual meaning; but, in this case, we've built facts about linguistic meaning into what was supposed to be a non-circular account of what it is for an expression to have a meaning. On the other hand, we might leave the possible worlds relevant to evaluation of the counterfactuals unrestricted; but, in this case, it seems clear that the nearest possible world in which most competent users of sentences like "0=1" would accept them is one in which those sentences have a different meaning than they actually do — and, in this case, the agent will not, in general, have the belief that 0=1, as Lewis's account requires.[16]

This problem about the formulation of (1b) is connected to the third objection raised above against Ramsey. I argued that incorporating the attitude of acceptance toward sentences into an account of meaning runs the risk of ruling out an natural mentalist explanation of linguistic competence. Lewis's account, as it stands, is not open to this objection: it is phrased entirely in terms of utterances of sentences and hearings of utterances of sentences. But of course one can often utter or hear a sentence without believing what the sentence says; in effect, the proponent of [L] will have to find a way to rule such cases out. And it seems that, as above, this will have to be done by appealing to the attitude of acceptance toward sentences, where this is an attitude which presupposes understanding of the sentence in question.

Most importantly, the second objection raised above against Ramsey applies also to [L]: the distribution of beliefs in a population may be such as to make an account of meaning in terms of correlations between beliefs held and sentences uttered impossible. In fact, as above, general facts about belief seem to make such an account impossible.[17] This can be seen by focusing on (1a) and (1b). Let S be a sentence which means p in a population G, and let q be a proposition which is, though distinct from p, a trivial consequence of p. Suppose that all speakers of a language believe p only if they also believe q, and that this is mutually known. This is clearly a possible situation; here are three plausible actual examples:

S ="Lassie was a brown dog"
p =the proposition that Lassie was a brown dog
q =the proposition that Lassie was a dog.

S ="It is true that Lassie was a dog"
p =the proposition that it is true that Lassie was a dog
q =the proposition that Lassie was a dog.

[16] Again, similar problems arise with (1a).

[17] Provided, that is, that the language in question is sufficiently complex to include sentences like those discussed below.

S =“σ and σ'”
p =the proposition that σ and σ'
q =the proposition that σ.

The objection is that, in such cases, [L] will wrongly count S as ambiguous in G between p and q.

Letting the analysandum be the false claim that S means q in G, it is clear that clauses (1a), (1b), and (2) are satisfied. Clause (3) is also satisfied; the expectation that (1a) and (1b) will continue to be true gives speakers a good reason to continue to utter S only if they believe q, and to expect the same of other speakers of the language. As clause (4) requires, there is a general preference for this regularity in behavior to continue. And again, as regards clause (5) there seems no reason why a difference in the analysandum between the fact that S means p in G and the fact that S means q in G should make any difference.[18] Hence, according to [L], we get the result that, in each of the above cases, S means q in G, and hence that S is ambiguous in G between meaning p and meaning q: “Lassie was a brown dog” is ambiguous in English between meaning that Lassie was a brown dog and meaning that Lassie was a dog, and any conjunctive sentence “σ and σ'” is three ways ambiguous between meaning that σ, meaning that σ', and meaning that σ and σ'. These results seem clearly incorrect.

It is difficult to see how a proponent of [L] might respond to this problem. Perhaps she should just bite the bullet: accept the claims about meaning and ambiguity to which the argument leads as a result of her theory. But this move is really not very plausible. It is simply not true that every conjunctive sentence of English is, just by virtue of its being a conjunction, semantically ambiguous. So accepting the conclusion will involve identifying, for each p/q pair above, the propositions p and q. But on any view of meanings, the meaning of a sentence either determines or consists in its truth conditions or assertability conditions; and it is clear that, in many cases of the sorts described above, both the truth conditions and assertability conditions of p and q will diverge.[19]

[18] It is perhaps worth noting that clause (5) of the rejected version [L*], discussed in note 14 above, is satisfied by these cases as well: if clause (5) of [L*] is satisfied when the analysandum is the fact that S means p in G, then it will be satisfied as well for the fact that S means q in G; if other speakers began to utter S only when they believed r, and r was a proposition about some subject matter totally unconnected to the subject matter of p and q, this would give a member of G a good reason to conform to this new regularity.

[19] Another route would be to use a technique employed by Scott Soames in *Beyond Rigidity* in his account of semantic content. One could defend [L] by saying that x means p in a population G just in case the following two conditions are satisfied: (i) the claim that x means p in G satisfies (1a)-(5), and (ii) for any q such that $p \neq q$, if the claim that x means q in G satisfies (1a)-(5), then the fact that the claim that x means p in G satisfies (1a)-(5) explains why the claim that x means q in G satisfies (1a)-(5), and not the converse. The worry about this is that it makes the analysis circular. Suppose that the fact that, ordinarily, members of G come to believe that it is true that Lassie was a dog upon hearing “It is true that Lassie was a dog” explains the fact that, ordinarily, members of G come to believe that Lassie was a dog upon hearing “It is true that Lassie was a dog”, and not the converse. It is hard to see what could substantiate this claim of explanatory priority other than the fact that the sentence heard has the same content as one but not the other of the two beliefs; but this is precisely what the mentalist proponent of G *can't* appeal to. This is related to the distinction between beliefs ‘directly’ and ‘indirectly’ gained discussed in connection with Ramsey's account above, on p. 3.1.

In one sense, the claim that language is conventional is, as Lewis says, "a platitude — something only a philosopher would dream of denying."[20] It is a platitude that the words of our language could have had different meanings, and that they owe their meanings largely to the use to which we put them. But the claim that language is conventional, when this is taken as the claim that facts about the meanings of expressions in our language are derived from the mutual knowledge of speakers concerning what members of a population mean or would mean by their utterances (or what they believe or would believe when making or hearing utterances), is far from platitudinous. Indeed, the arguments of this section, along with those of §2.3, indicate that it is false.

For discussion of further quasi-technical problems with the project, see Appendix D, "Convention, serious circumstances, and word-meaning."

[20]Lewis (1975), 166.

Conclusion

In the last two chapters, I've discussed the two ways that a mentalist might go about giving an account of linguistic meaning in terms of facts about mental content: the indirect strategy of giving both an account of linguistic meaning in terms of an action-based propositional attitude and an account of the contents of the chosen action-based propositional attitude in terms of mental content, and the direct strategy of giving an account of the meanings of expressions in a language in terms of the contents of some of the mental states of speakers of that language.

The prospects of the indirect strategy appear very dim indeed. As argued above, the Gricean account of meaning faces problems from so many different sides that there appears little hope of arriving at a satisfactory account along those lines; and the kinds of cases which pose problems, as I argued above, suggest that speaker-meaning is best understood as derivative from linguistic meaning.[21] This conclusion, moreover, seems to extend to the other action-based propositional attitudes: asserting, informing, communicating, and the rest. With these results in hand, it seems all but impossible to give an account of linguistic meaning acceptable to the mentalist which will account for linguistic meaning in terms of conventions governing any of the action-based propositional attitudes.

In my view, the two direct strategies discussed — Lewis's account of linguistic meaning in terms of conventions of truthfulness and trust in a language, and Ramsey's pragmatist view — are more hopeful courses for the mentalist to take than the more popular Gricean indirect strategy. Nevertheless, both run up against the same two fundamental problems: (i) Meaningfulness is not constrained by believability; and it is difficult to see how the approaches of Lewis or Ramsey could be formulated so as to apply to sentences which virtually any competent speaker of the language would reject. (ii) Facts about the distributions of beliefs among agents may, in a given population, be such as to make an assignment of meanings to sentences based on a correlation of sentences uttered and accepted with beliefs held impossible; general facts about belief such as those discussed above seem to show that, in a language of sufficient complexity, this will always be the case. Indeed this latter point is, I think, another way of putting the motivation for pursuing the indirect strategy in the first place: facts about what speakers assert and mean by utterances are obviously more reliably correlated with the sentences they utter and the meanings of those sentences than are facts about what those speakers believe. These seems to indicate that there is something fundamentally wrong with the approach, and that the beliefs that agents typically hold when uttering and accepting a given sentence should be regarded as being explained by, rather than as explaining, the meaning of that sentence. So in the end it seems as though the direct strategy, at least in the forms I've considered, is also on the wrong track.

Of course, this is an argument by cases: I've argued that no extant account of linguistic meaning will suit the mentalist's purposes, and this is compatible with the claim that there is some such account which no one has yet constructed. In the absence of specific proposals, this claim is difficult to evaluate. But some indication that the claim is false is provided by the sorts of problems the mentalist runs into. These are not a matter of simple counterexamples

[21] See Appendix B, "Conditions of satisfaction and the expression of belief", for a discussion of non-Gricean but mentalist accounts of speaker-meaning, and Appendix C for a sketch of a communitarian account of speaker-meaning.

which seem susceptible of solution by fiddling with the formulations of any of the accounts considered. Rather, there are in each case explanations of why these cases pose problems for the version of mentalism in question, and these explanations seem to point to the conclusion that the project is fundamentally misguided.

If all of this is right, then the argument of the last two chapters is enough to give us the conclusion that mentalism is false: linguistic meaning is not completely constituted by more fundamental facts about the contents of the mental states of agents. I hope that it is also enough to show, not only that there are likely no necessary and sufficient conditions, given exclusively in terms of facts about mental content, for a linguistic expression having a given meaning, but that the mentalist view of language is, to use the language of §1.2, the wrong *picture* of what it is for an expression to have a given meaning.

In a way, though, these criticisms do not really strike at the heart of the mentalist picture. The basis of that picture was that a class of basic propositional attitudes are the most fundamental kind of representational state, and that, accordingly, they are not to be explained in terms of the representational properties of linguistic items. This means that the mentalist is committed to there being some story about what it is for an agent to believe p which makes use neither of relations between that agent and expressions of her private language, nor of relations between that agent and a public language of which she is a speaker. Even if the foregoing shows that the mentalist program cannot be carried out, it might still be the case that the view of mental states with which the program began is correct. In the next chapter, I consider whether this portion of the mentalist picture of representation — the view that facts about the mental states of agents can be explicated without reference to facts about the contents of expressions in any language, whether public or private — can be vindicated.

Chapter 4

Belief and Belief States

Contents

4.1 From mentalism to functionalism

A mentalist account of a propositional attitude is one which treats it as prior to, and hence independent of, linguistic meaning. This is so whether the meanings are meanings of expressions in public languages, as the communitarian would have it, or of expressions in internalized idiolects or languages of thought, as private language theorists would have it.

As emphasized in the last two chapters, the class of propositional attitudes is very broad, and includes many attitudes, such as speaker-meaning, which now seem clearly not to be open to a mentalist account. But, of course, the mentalist view of the attitudes is strongest, not when applied to attitudes like speaker-meaning or assertion, but to attitudes like belief, intention, and desire.

In this chapter, I will focus attention on belief. This is the attitude to which mentalists pay most attention, and among those usually taken to be most amenable to mentalist treatment.

I shall argue that there is no true mentalist account of what it is for an agent to believe p; if this conclusion is right, then I think that there will be good reason to extend it to the other attitudes, and to conclude that both sides of the mentalist picture of meaning and content are mistaken: there is no account of what it is for an expression to have a meaning in terms of mental content, and no account of facts about mental content which does not advert to facts about linguistic meaning.

As it happens, all going mentalist accounts of belief fit a certain mold: they are all variants of functionalism, broadly construed. The purpose of this section is to say why this is more than current philosophical fashion: given a few plausible assumptions, the mentalist picture of mind and language entails a functionalist theory of belief. We'll then be well-positioned to see what options the functionalist has for developing her views.

Inasmuch as beliefs are typically attributed to agents, an account of what it is to believe p must be given at least partly in terms of properties of agents. Since different agents have different beliefs, a constitutive account of belief must explain this divergence; and it is hard to see how one could do so except by making properties of believers — whether relational or monadic — part of one's account of belief.

The natural next question is: which properties of agents? Here two possibilities suggest themselves. The beliefs of agents may either be constituted by the internal states of agents (or properties thereof) or by their dispositions to action.[1]

At this stage of the argument, some might object that this list of options is not exhaustive: it might be, after all, that facts about the beliefs of agents are not constituted by facts about either their internal states or dispositions to action, but rather by *normative* facts about them. This is, I think, a genuine possibility; but it is not, I think, a real alternative to explaining beliefs in terms of internal states or dispositions of agents. For few who would be tempted by an explanation of belief in terms of normative facts would be content to rest with these normative facts as the full constitutive story about belief. Rather, it is natural to think, some story would be owed about the facts in virtue of which those normative facts obtain with respect to the relevant agents. Such a story might be compatible with mentalism, or it might not be; that is, it might or might not make use of facts about private or public languages. But if it is to be compatible with mentalism, it seems clear that the constitutive account of the relevant normative facts will be given either in terms of the internal states of agents (or properties thereof) or their dispositions to action.

Suppose first that the mentalist takes the second option, and claims that what it is for an agent to believe p is for that agent to have certain dispositions to action. This is to endorse a kind of behaviorism about belief, which might be stated as follows:

a believes $p \equiv \exists\phi$ (a is disposed to ϕ & $B(\phi, p)$)[2]

Here 'B' stands for a relation between actions and propositions; the hope is that such a relation could yield an account of the propositions believed by an agent in terms of the actions that

[1] This is a false dichotomy on views of the metaphysics of dispositions on which dispositions are identified with their categorical bases; but this view shouldn't be presupposed in our initial outline of the mentalist's options.

[2] Recall that biconditionals such as this one are to be understood as prefixed by the necessity operator, with all free variables bound by universal quantifiers taking wide scope over everything in the formula other than the necessity operator. As above, I omit the quantifiers and necessity operator for simplicity.

agent is disposed to perform. It seems to me that the most decisive argument against this sort of account is a very simple one: in the case of many beliefs, it is difficult to see what the relevant behavioral dispositions might be. This is easiest to see in the case of relatively sophisticated beliefs, such as, for example, the belief that arithmetic is incomplete. The only behavioral dispositions which are reliably correlated with the holding of this belief are, it seems, dispositions to verbal behavior, among which might be the disposition to accept some sentence which means that arithmetic is incomplete. But this is not the sort of thing of which the mentalist can avail himself; the starting point of the mentalist program is, after all, the thought that mental states like belief are more fundamental than, and hence not constituted by, facts about the meanings of expressions in public languages. So the mentalist who wants to be a behaviorist about belief cannot appeal to dispositions to accept sentences with certain meanings in public languages. She cannot, moreover, let relations to tokens of sentence-types do duty for relations to sentences with their meanings; the sentences in question could have meant something else, in which case relations to them would not be reliably correlated with the belief in question, and one might believe that arithmetic is incomplete by accepting a sentence other than "Arithmetic is incomplete" which is synonymous with it.[3]

So it seems that a mentalist should not endorse behaviorism, but rather should take the first option mentioned above: say that what it is for an agent to believe p is for that agent to be in a certain internal state. But there are two ways to interpret this claim. According to the first, for any proposition p, the property of believing p is constituted by a certain first-order non-intentional property of agents (e.g., the property of being in a certain brain state). According to the second, the property of believing p is constituted by a second-order property of first-order states of agents (e.g., the property of being in some state with such and such relational properties). I shall, following standard usage, call the first of these an *identity theory* of belief, and the second a *functionalist theory* of belief.

Identity theories are usually taken to founder on the problem of multiple realizability: the problem that, for some mental property F, two agents may both instantiate F while having no physical properties in common. Take, for example, the property of believing that grass is green. An identity theory about this property will say that what it is for an agent to believe that grass is green is for that agent to instantiate some physical property F; that is, she will claim that

a believes that grass is green $\equiv Fa$

[3]It is worth noting that many of the versions of behaviorism advanced in its heyday were explicitly not mentalist in character. See, for example, the discussion of belief in Ryle (1949), 134-135.

At least two other arguments are often cited as decisive refutations of behaviorism, both of which can be found as early as 1957 in Chapter 11 of Roderick Chisholm's *Perceiving: A Philosophical Study*: (i) Behaviorist accounts of mental states will always make use of an agent's dispositions to perform certain kinds of intentional actions; but, since intentional actions are partially constituted by the beliefs and desires of agents, this sort of account is circular. (ii) Connections between dispositions to action and mental states are can always be overridden by other beliefs and desires which will make it possible to have the relevant mental state *without* the disposition, and so provide a counterexample to the proposed behaviorist analysis. I am inclined to think that (ii) is a technical problem which can be overcome, and that (i) presupposes a questionable thesis in the philosophy of action; so I do not regard either of these to be as definitive as the simpler objection given in the text. See Part III below for a discussion of these and several other traditional arguments against behaviorism.

But, no matter what we take F to be — the property of being in a certain neural state, or whatever — it seems that the above formula will be obviously false: there could be neuron-free agents capable of beliefs. This problem is, I think, enough to show that identity theories are incapable of giving an account, for any mental property, of what it is for an agent to instantiate that property.[4]

We began by noting that an account of what it is for an agent to believe p must be given in terms of properties of agents, and by noting two candidate classes of properties: properties of being disposed to do certain things, and properties of being in certain internal states. We further divided the second class of properties into two categories: properties of being in certain first-order states, and properties of being in first-order states with certain second order properties. I have argued that the mentalist should rule out two of these three options: behaviorism, which focuses on dispositions to action, and identity theories, which focus on properties of being in certain first-order states. This leaves functionalism as the only mentalist theory of belief still standing; as we'll see, however, functionalism is less a theory than a class of widely disparate theories which fit a certain mold.

Above I characterized functionalism about belief as the claim that what it is for an agent to believe p is for that agent to be in some state with a second-order property which qualifies it as the belief that p. So the obvious question to be answered by a functionalist theory is: what are the relevant second-order properties? Really, though, this question is at least in principle divisible into two parts. The functionalist owes an account of the properties of first-order states which makes those states beliefs with a certain content; but one might want to give separate accounts of the second-order properties which make certain states *beliefs*, and the second-order properties which make those states have a certain *content*.

One reason for so dividing functionalist accounts of belief can be given by outlining an implausible, but coherent, functionalist story. One might want to give a functionalist account of a number of propositional attitudes — not only belief, but also, say, intention and desire. And one might think that the contents of states of agents are, in general, fixed by their causes; so what it is for an internal state to have content p is for that internal state to have the property of being caused by p. But, the story might continue, not all internal states with content are beliefs; some are intentions, and some desires. So our account of belief (and of intention and desire) is not complete when we identify the second-order property which fixes the content of an internal state; we also need to identify the second-order properties in virtue of which some of those states are belief states, some intention states, and so on.

Of course, this theory is absurd for a number of reasons.[5] But it does illustrate one motivation for giving separate accounts of content and of what it is for a state to be a belief state. Given this possibility, we can say that a functionalist account of belief will be an instance of the following schema:

[4]The identity theorist might respond to this problem by saying that what it is to have a certain belief is to be in one of a (perhaps very long) disjunction of internal physical states. Aside from the implausibility of this sort of view, it falls prey to the converse of multiple realizability: it is not only the case, presumably, that a belief might be realized by several different types of physical states, but also that a given physical state might realize many different beliefs. Then being in one of this disjunction of physical states would not be a *necessary* condition for having the belief in question.

[5]E.g., it entails that one can only desire things which are already the case.

[F] $\Box\ \forall a \forall p$ (a believes $p \equiv \exists x$ (x is a belief state of a & x has the content p for a))

instances of which are obtained by (i) defining the '___is a belief state of ___' relation, and (ii) defining the '___has the content ___for ___' relation.[6] Each of these relations will be second-order (relational) properties of states of agents.

What, one might ask, makes instances of this schema versions of *functionalism*? So far, after all, we have said a bit about the role played by second-order properties in such a theory, but have said nothing about functional properties or functional roles. The historical answer is that many theorists who have offered theories of the form of [F] have defined these two relations in terms of the functional roles of internal states; in terms, that is, of their causal relations to some combination of other internal states, input factors like perception, and output factors like actions. But I will, in a (perhaps not so slight) abuse of terminology, call instances of [F] which define neither the '___is a belief state of ___' nor the the '___has the content ___for ___' relation in terms of functional roles "functionalist" as well. This does pick up one use of "functionalism" in the literature; but it is importantly different from a more restrictive use of the term to pick out views which are not only instances of [F], but also identify one or both of the above relations with certain kinds of functional roles.[7] Which way one uses "functionalism" is a purely terminological question; but it will be important to be clear in what follows that functionalist accounts of belief are just instances of schema [F].

On this construal, functionalism about belief is less a single theory than a broad class of theories of belief. As it turns out, virtually all going accounts of belief can be viewed as versions of functionalism, in this broad sense. Because the purpose of this chapter is to examine the possibilities for giving a constitutive account of belief within a mentalist framework, and because, as argued above, the only hope for such an account lies with functionalism, this chapter will be devoted to examining some of these accounts. Rather than just going through these theories in a piecemeal fashion, though, it will be useful to have an overview of what the possibilities for developing such an account are.

As noted above, any version of functionalism must say what properties of internal states make it the case that some are belief states, and that some have a particular content. In doing so, the first question the functionalist faces is

> Given that the relevant properties of internal states are relational properties, are these properties restricted to 'internal' relations between those states and other internal states of the agent, or do they include 'external' relations between those states and features of the environment of the agent?

I shall call theories which take the former course *solipsistic* versions of functionalism.[8] In the next section I shall argue briefly that solipsistic functionalism cannot provide a plausible account of belief; I shall then turn to outlining four kinds of nonsolipsistic, or externalist, functionalist theories.

[6] This should not be taken to rule out the possibility that a functionalist might want to define these two relations together.

[7] In the terminology of Fodor (1985), these more restrictive uses of the term might be called, respectively, functionalism about belief states and a functionalist semantics for belief states.

[8] The term is due to Harman (1987).

4.2 Solipsistic theories of content

A wide variety of theories of content have been advanced in recent years; in order to see the motivation behind these different theories, it will be useful to have a grasp of some of the challenges a theory of content must meet. The constraints on such theories which follow are only, it should be noted, *prima facie* constraints; as part of her theory of content, a mentalist may show that one or more of these constraints is not a genuine constraint on a theory of content. The following are five apparent facts about belief:

> *Determinacy of content.* At least sometimes, it is determinately true that an agent believes a proposition p (rather than determinately true that the agent believes one of a list of propositions $p, p_1, p_2, \ldots$, and indeterminate which of them she believes).
>
> *Fallibility.* Agents capable of having beliefs sometimes make mistakes, either by having false beliefs or by incorrectly inferring one proposition from another.
>
> *Externalism.* The beliefs of an agent do not, in general, supervene on the intrinsic properties of that agent.
>
> *Shared beliefs.* Often, two people can have the same belief.
>
> *Belief retention.* Usually, acquisition of new beliefs need not interfere with maintaining past beliefs.

These are, I think, intuitively plausible claims about belief which a theory of content should either explain or explain away.

How might a theory of content meet these constraints? One might think the following: the possible belief states of an agent form a system. Which beliefs an agent has depends upon which of these possible belief states are actual, and what the contents of those belief states are. The content of a belief state depends upon its place in the agent's (potential) system of belief states.[9] The following simplified theory [T] provides one way of making this thought explicit; its importance is not that it is widely held (it isn't), but rather that seeing the ways in which it fails will bring to light some of the motivations behind recent attempts to construct theories of content.

> [T] What it is for an agent a to believe p is for there to be some belief state x of a which has counterfactual relations to other belief states (possible and actual) of a which mirror the relations of entailment (i.e., necessary consequence) between p and other propositions.

The idea is that there are certain relations between the belief states of a of the following form: if a were in belief state x, then a would be in belief state y; if a were in belief states x and

[9] Recall that the functionalist about belief owes an account of what it is for an internal state to be a belief state, and of what it is for an internal state to have a certain content. I am focusing on the second of these questions here, supposing that the solipsistic functionalist has already given a satisfactory account of belief states.

x', then a would be in belief state z; and so on, for all of a's possible belief states. There are also relations of necessary consequence between propositions; perhaps, for propositions p, p', q, and r, these relations are as follows: p entails q, and p and p' jointly entail r. Imagine, *per impossible*, that there are only four possible belief states of a, and that there are only four propositions which exhibit this structure of entailments. Then [T] would yield the following result: a believes p iff a is in x, a believes p' iff a is in x', a believes q iff a is in y, and a believes r iff a is in z.

It is easy to see that [T] fails to meet any of the constraints on a theory of content listed above; what is notable is that it fails almost all of them for the same reason. Any functionalist theory of content will have to assign contents to belief states on the basis of the relations they bear to something; the important point about [T] is that it says that only the counterfactual relations between internal states are relevant to determining the contents of those states. This restriction of the determinants of belief contents to intra-psychological relations between belief states is the source of [T]'s failure to account for fallibility, externalism, the determinacy of content, and the phenomena of belief retention and shared belief.

First, because [T] takes the counterfactual relations between belief states to determine the contents of those states, the inferences agents are disposed to draw — which involve coming to be in one belief state as a result of having come to be in another — must infallibly track relations of entailment between propositions. This can be illustrated by example. Suppose that an agent a is not infallible in her inferences; in particular, suppose that a sometimes comes to believe q as a result of coming to believe p, though the truth of p has no bearing on the truth of q. According to the functionalist picture, this acquisition of beliefs occurs by virtue of a's coming to be in two new belief states x, y, the second as the result of the first. So among the relations between belief states in a's psychology will be the following: if a were in x, then a would be in y.[10] But then [T] entails that the proposition assigned as content to y will be a necessary consequence of the proposition assigned as content to x; and this will result in the wrong characterization of a's beliefs, since it was stipulated that the truth of p has no bearing on the truth of q.[11] Hence the proponent of [T] must hold that the stipulated situation is an impossible one. But the situation is clearly not impossible; agents can and often do make faulty inferences of the kind described.[12]

Second, it is clear that [T]'s restriction of the determinants of belief to facts about the relations between beliefs runs counter to the apparent dependence of the contents of an agent's beliefs on facts external to the agent. In the thought-experiments of Putnam and Burge which illustrate this dependence, the internal relations between the belief states of the relevant agents are identical, and yet their beliefs different.[13] [T] cannot account for this difference.

Third, [T] fails to yield a determinate assignment of beliefs to agents. The problem is

[10] One can imagine the relation being more complicated — there may, e.g., be other belief states required to bring about a's being in y. But the simplest case is enough to illustrate the point.

[11] Of course, [T] could be modified so that the relation between the propositions assigned as contents of belief states so related is something weaker than necessary consequence; but, since that wouldn't solve the problem — because p bears *no* relevant relation to q — I stick with the simpler version here.

[12] Note that the kind of infallibility required by [T] is not the requirement that agents have only true beliefs; [T] makes room for false beliefs. The infallibility in question is infallibility in drawing inferences, whether from true beliefs or false ones.

[13] See, e.g., Burge (1979), 27.

that there may be more than one structure of propositions isomorphic to an agent's system of beliefs, and [T] does nothing to choose between these different structures. It seems that this problem will arise no matter how extensive an agent's system of belief states, as is shown by the following example from Stalnaker:

> Mary is angry at Fred, her neighbor. She wants him to suffer, and believes that he will suffer if she plays her cello badly in at three o'clock in the morning. That, at least, is one hypothesis ... Here is another: Mary wants *Albert* to suffer, and believes that *Albert* will suffer if she plays her cello at three in the morning. ... The *only* difference between the two proposals, let us suppose, is that in the perverse hypothesis, Albert is everywhere substituted for Fred. ... Not only do belief and desire interact to produce the same actions, according to the two hypotheses, but there is also an exact correspondence between the beliefs hypothesized ... by the two competing accounts ... they are equivalent with respect to the mechanisms they postulate.
>
> The same point will hold for any such substitution, not only of individuals for individuals, but of properties for properties, or whole propositions for whole propositions. All that is required is that certain internal structure be preserved.[14]

This degree of indeterminacy is unrealistic; there is surely a fact of the matter as to whether my beliefs (which seem to me to be) about my friend are really beliefs about him rather than about some individual with whom I am not acquainted.

Fourth, [T] has trouble accounting for the phenomenon of belief retention. As stated, [T] seems to imply that the beliefs of an agent are extraordinarily sensitive to changes in the agent's belief system as a whole. Under [T], the beliefs of an agent depend on an isomorphism between the agent's system of belief states and relations between propositions; the problem is that it seems that even a small change in this system of counterfactual relations between belief states will result in a change of assignments of content to many of the belief states in the system. For, as noted in the previous paragraph, the problem for a theory like [T] is not that, given a structure of belief states of the sort outlined above, there aren't any structures of propositions isomorphic to it; the problem is that there are too many. So, given a change from system of beliefs S to system S', there will be many structures of propositions isomorphic to the new system which are not isomorphic to the old one; and there is nothing in [T] to rule out these new structures from assigning different propositions to belief states not directly involved in the change from S to S'. This poses a problem in accounting for the retention of beliefs because, usually, no special effort is required to maintain one's believing p while changing one's mind about the relationship between two propositions q, r apparently unrelated to p; but, if [T] were right, this sort of change in one's system of belief might involve one's giving up — involuntarily, as it were — the belief p.

Fifth, the same sensitivity of the beliefs of an agent to that agent's overall system of beliefs makes sharing beliefs more difficult than we ordinarily think. We do not hesitate to ascribe a belief to several members of a group of people; but, given the extreme unlikeliness of their all sharing the same system of beliefs, there is no reason to think that there will be any belief which [T] will assign to each.

[14] Stalnaker (1984), 17-18.

Each of these five problems arises as a result of [T]'s restriction of the factors which determine the content of a belief state to internal relations between belief states.[15] Now there are responses open to the proponent of a theory like [T]: she may try to rescue the possibility of fallible inferences by restricting the relevant counterfactual relations to those which hold under certain ideal conditions; she may try to account for externalist intuitions by treating this dependence of content upon the environment as a kind of indexicality; she may try to account for the phenomena of belief retention and belief sharing by treating reference to sameness of beliefs over time or across individuals as loose talk for similarity of beliefs over time or across individuals. Nevertheless, problems like the five I've listed have been enough to convince many that [T] is on the wrong track altogether: that the contents of belief states are fixed, not by relations to other belief states, but, at least in part, by relations between those states and objects, properties, or states of affairs in the world.

In what follows, I shall presume that if there is a viable functionalist account of belief, it will be a nonsolipsistic one.

4.3 Four kinds of externalism

We began by asking what it is for an agent to believe p; in keeping with individualism, we have taken it to be a constraint on answers to this question that it not invoke facts about the meanings of expressions in public languages. Given this constraint, we concluded that of the three possible answers to the question of what constitutes having a belief — behaviorism, identity theory, and functionalism — only functionalism has any real plausibility. In the last section we concluded that the most promising forms of functionalism will be ones which take the second-order property constitutive of a state of an agent's being a belief state with a certain content to be a relational property involving not only other internal states of the agent, but also aspects of the agent's environment.

This leaves us with four options, which may be brought out by considering two questions for the nonsolipsistic, or externalist, functionalist:

Are the second-order properties constitutive of belief *exclusively* comprised of rela-

[15] It is clear how the first three problems stem from this restriction; the issue may be less clear in the case of the latter two. The idea is this: once one takes counterfactual relations between belief states to be the sole determinants of the contents of those states, then it seems that one has no choice but to take *every* such relation to be constitutive of the content of those states. (This is what leads to the sensitivity of content assignments to changes in the system of belief states as a whole.) For how could the counterfactual relations between belief states relevant to content assignments be delimited? [T] is supposed to take us from a system of belief states, described only in terms of counterfactual dependencies between those states, to an assignment of content to each state in the system. The problem is that it seems that only *given* an assignment of contents to these states could some counterfactual relations be deemed more important than others. The problem is thus one of circularity: the structure of the system of belief states is supposed to determine the contents of those states, but any principled restrictions on what relations between states are to count as part of the system can only be drawn after the content of the states is determined. I hope it is clear that this claim is different from the claim that there is no principled distinction between analytic and non-analytic sentences or entailments. The difference is the difference between the claim that some sentences are true by virtue of meaning alone and the claim that expressions come to have the meaning that they have by virtue of their place in sentences true by virtue of meaning alone.

> tions between internal states and the external world, or are some internal relations between states of the agent relevant?

Note that in the previous section we only argued that solipsistic versions of functionalism — versions which only make use of such internal relations — cannot succeed. This still leaves it an open question whether some of these internal relations might be relevant.

The second question is:

> Are the contents of belief states derived from the contents of constituents of those states?

To answer this question in the affirmative is to endorse the idea that a theory of the contents of internal states is in the first instance an account of the contents of *mental representations*, where (roughly) these stand to belief states in the same relation as sub-sentential expressions stand to sentences of a language. This is, more or less, to take beliefs to be underwritten by a language of thought.

Since these two questions are independent, we can define four kinds of functionalism by one's answers to them:

Four kinds of externalist functionalism	*Contents of belief states are basic*	*Contents of belief states derived from contents of mental representations*
Only external relations relevant	Indication theories	Informational theories
External & internal relations relevant	Functional role theories	Conceptual role theories

The terms used for these four kinds of nonsolipsistic functionalism have all been used in the literature to signify a variety of things; the point of introducing the terms via this chart is only to fix terminology, and to keep straight a number of distinctions which some uses of these terms elide. Of these four kinds of nonsolipsistic functionalism, only the two in the left column are available to a mentalist. Recall that the mentalist endorses the claims (i) that the meanings of expressions in public languages are derived from the contents of the mental states of speakers of those languages, and (ii) that these propositional attitudes — the beliefs, desires, intentions and so on of individual agents — are not constituted by any more basic sort of intentional state. Claim (i) is consistent with all four types of nonsolipsistic functionalism; but claim (ii) is inconsistent with informational and conceptual role theories of belief.

This is because informational and conceptual role theories of belief take the contentfulness of beliefs to be derived from a more fundamental source of intentionality: the contents of mental representations in an agent's language of thought. As such, the natural home of these accounts of belief is the private language picture, which takes the meanings of expressions in individual languages of agents to be the most fundamental kind of intentionality. That picture of intentionality is the topic of Part II of this essay.

As purpose of this chapter is to argue that there can be no successful mentalist account of belief, it will suffice to show that there can be no successful indication or functional role theory of belief.

4.4 Content and indication

We can begin by exploring the plausibility of theories located in the top left corner of our diagram of possible individualist positions: indication theories, which take only relations between belief states and the external world to be relevant to fixing content (and do not explain these relations in terms of more fundamental relations between constituents of belief states and objects and properties in the world).

The fundamental question to be answered by theories of this sort is: given that the contents of belief states are fixed by relations to an external world, what exactly are the character of these relations? A natural answer to this question has it that *causal* relations between belief states and facts about the world are the determinants of content.

Broadly causal theories of mental content have occupied center stage in the philosophy of mind for the last twenty-odd years. The reason for this popularity is not far to seek. Causal theories, after all, promise to resolve each of the five problems with our solipsistic theory of the previous section: (i) agents are not required to be infallible in their inferences, since the contents of belief states are not fixed by their counterfactual relations to other belief states; (ii) room is made for externalist intuitions, since facts external to believers are given a role in determining the contents of their belief states; (iii) no widespread indeterminacy of content arises, since causal relations to states of affairs can, presumably, decide between the sorts of competing hypotheses about Mary's beliefs entertained above; and because the content of a belief state is not fixed by its place in an agent's system of belief states, its content is not sensitive to changes in that system in a way that poses problems for the phenomena of (iv) belief retention or (v) belief sharing.[16]

4.4.1 From a simple causal theory to the causal-pragmatic theory

One sophisticated and plausible account of belief along causal lines was defended by Robert Stalnaker in *Inquiry*. To see the motivations behind adopting a theory like Stalnaker's, consider first the following very simple causal account of belief:

> Necessarily, an agent believes p iff there is some state of the agent that the agent is in because p is the case

We can see how Stalnaker's theory emerges from this simple causal theory by considering two problems that show that this theory is false as it stands.

[16]It is worth noting that endorsing an indication theory of content does not commit one to taking the relevant relations between belief states and the world to be causal; one could, like Jerry Fodor, take the content-constituting relations to be laws connecting events in the world and facts about internal states of an agent (see Fodor (1990*a*). Indeed, Fodor's move here has some advantages over causal accounts, and seems to preserve features (i)-(v) listed above. But Fodor's move fits better with informational theories of meaning, which are discussed as part of the private language picture of intentionality in the next chapter.

The first problem is that this theory cannot account for the possibility of false beliefs. One way of expressing this is that this simple causal theory faces what Jerry Fodor has called "the disjunction problem,"[17] so called because simple causal theories misrepresent false beliefs as true disjunctive beliefs. When an agent mistakenly comes to believe p, the agent forms the belief because some other fact q is the case. Suppose for the sake of example that this is a very simple case of error; whenever the agent comes to believe p, this is either because the agent is correct, and p is the case, or because the agent has made a certain mistake, and formed the belief because q is the case. Because this simple causal theory identifies the content of a belief state at a world with its causes in that world, it entails that, contra our original supposition, the agent does not falsely believe p after all. Rather, since the agent is always in this state because either p or q is the case, the simple causal theory says, wrongly, that our agent is not making a mistake, but rather has the true disjunctive belief (p or q).

Stalnaker's response to this problem, following the lead of other like-minded theorists, is to say that the content of an internal state of an agent is not fixed by what actually causes the agent to be in that state, but rather by what *would* cause the agent to be in that state, were the agent in optimal conditions.[18] Optimal conditions are conditions in which an agent's cognitive system is functioning perfectly; the intuition is that the content of a state is not determined by actual causes of that state, but rather by its causes in conditions where various factors which block the ideal functioning of an agent's belief forming mechanisms, such as illusions and cognitive shortcomings, are absent. The key point as regards the disjunction problem is that these optimal conditions must be such that, were the agent in optimal conditions, she would have no false beliefs.[19] This solves the disjunction problem, since it makes room for the possibility that an agent may be in a certain state which has the content p despite the fact that the agent was not actually caused to be in that state by p being the case. Adding this reference to optimal conditions to our simple causal theory yields the following modified causal theory of belief:

> Necessarily, an agent believes p iff there is some state of the agent such that, were the agent in optimal conditions and in that state, the agent would be in that state because p is the case.

Following Stalnaker, this may be expressed by saying that the contents of states of agents are determined by what they *indicate*.

It is worth pausing for a moment in the development of Stalnaker's theory to note that the optimal conditions response to the disjunction problem, or the problem of making room for

[17] See Fodor (1987).

[18] See especially Stampe (1979). Relevantly similar views may be found in Dretske (1981); Fodor (1980).

[19] But note that, on pain of circularity, the optimal conditions cannot be specified in terms of an agent's beliefs being true; the truth of an agent's beliefs when in optimal conditions is supposed to be a consequence of being optimal conditions, which are specified independently. This seems particularly worrisome when one notes that optimal conditions will have to rule out illusions, for it is natural to think that an illusion is just a perceptual experience which represents the world as being some way that it is not. But this view of illusions defines them partly in terms of the contents of states of agents; and optimal conditions cannot, on pain of circularity for the indication theorist, be defined partly in terms of the contents of states of agents. For some skepticism about the possibility of giving a non-circular specification of optimal conditions which will meet this constraint, see Schiffer (1986).

false beliefs, is only one of two ways in which mentalists who give causal theories of meaning can go. One might not appeal to what *would cause* one to be in a certain state in certain counterfactual conditions, but rather to a subset of those things which *actually have caused* agents to be in that state, either in an agent's own developmental history or in the history of the species of which the agent in question is a member. This sort of view — often called the *teleological* theory of content — seems better applied to mental representations than to belief states. Its solution to the problem of error is essentially historical; and, because many beliefs are had only once, and new beliefs had all the time, it seems likely that many belief states will lack the history required for this sort of solution to the problem of error. For this reason, I defer discussion of teleological theories until Part II of this essay, which is devoted to the private language picture of intentionality.

In this section, then, we can restrict ourselves to solutions to the disjunction problem which involve appealing to optimal conditions in one form or another. The addition of optimal conditions to the simple causal theory, however, is not enough to solve another problem: many states of agents indicate things but, intuitively, are not beliefs. As Stalnaker points out,

> ... if a bald head is shiny enough to reflect some features of its environment, then the states of that head might be described in terms of a kind of indication — in terms of a relation between the person owning the head and a proposition. But no one would be tempted to call such states belief states.

Even clearer examples are not difficult to come by; the temperature of pavement indicates the temperature of the air above the pavement, but it would be very odd to describe the pavement as believing anything about the temperature of the surrounding air. The moral is that, because only some of the states that indicate something are belief states, we need to add an account of belief states to our causal theory.

Stalnaker's idea is that while causal relations of indication determine the contents of belief states, their status as belief states (rather than some other sort of state) is determined by their connections to action:

> Beliefs have determinate content because of their presumed causal connections with the world. Beliefs are *beliefs* rather than some other representational state, because of their connection, through desire, with action.[20]

But what is the needed connection, through desire, to action? Earlier Stalnaker tells us that

> To desire that P is to be disposed to act in ways that would tend to bring it about that P in a world in which one's beliefs, whatever they are, were true. To believe that P is to be disposed to act in ways that would tend to satisfy one's desires, whatever they are, in a world in which P (together with one's other beliefs) were true.[21]

For Stalnaker, then, what it is for an agent to have a certain belief is both for that agent to be in an internal state which indicates something, and to be disposed to act in certain ways. Neither the states of the bald man's head nor the temperature of pavement are beliefs because

[20] Stalnaker (1984), 19.
[21] Stalnaker (1984), 15.

neither the bald man nor the pavement is disposed to act appropriately on the basis of what the states indicate. We may express this "causal-pragmatic" theory of belief as follows:

> Necessarily, an agent believes p iff
>
> (i) there is some state of the agent such that, were the agent in optimal conditions and in that state, the agent would be in that state because p or something which entails p is the case, &
>
> (ii) the agent is disposed to act in ways that would tend to satisfy his desires in a world in which p together with his other beliefs is true[22]

Were this theory of belief true, it would be, from the mentalist perspective, quite an achievement. For the following diagram illustrates how naturally it fits together with neo-Gricean accounts of linguistic meaning to provide a fairly comprehensive picture of the relationship between mind and language:

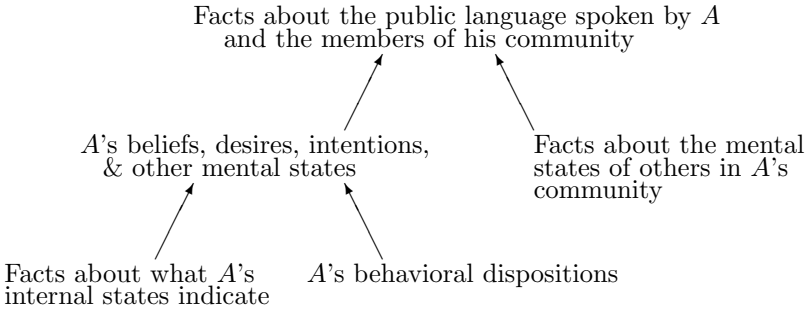
[22]This causal-pragmatic theory may be expressed more formally as follows:

> $\Box\ \forall a \forall p$ (a believes $p \equiv (\exists x$ (a is in state x & (a is in optimal conditions $\Box\!\!\rightarrow a$ is in x only because p or something which entails p is the case)) & (a is disposed to act in ways which tend to satisfy his actual desires in a world in which p and a's other beliefs are true)))

There is some question as to how each of these clauses should be interpreted. As stated, clause (i) claims that a state indicates p if, in the nearest world in which an agent is in that state and is in optimal conditions, that agent is in that state because p. But suppose that there is one uniquely nearest possible world meeting this description. Then the agent's being in this state — like almost any other state of affairs — will be caused by, and hence explained by, an enormous number of other states of affairs. It will be implausible to count the agent as believing very many of these to be the case. Perhaps the implausible ones will be ruled out by clause (ii) of the account; or perhaps the interpretation of clause (i) should be revised so as to count as part of the content of a state only causes of that state common to a class of possible worlds in which the agent is in optimal conditions. I leave this as an open question in what follows.

As regards clause (ii), it is surely too strong to require that an agent be disposed to act in such a way as would satisfy her desires in *every* world in which her beliefs are true; a more plausible interpretation seems to be that the agent must be disposed to act in a way which would satisfy her desires in the *nearest* world in which all of her beliefs are true. I assume further that the intended reading of (ii) requires that the agent be disposed to act in ways which satisfy his actual desires in a world in which all of his beliefs are true, rather than that he be disposed to act in ways which, in a world in which all his beliefs were true, would satisfy his desires *in that world*. This seems clearly to be what Stalnaker had in mind.

THE CAUSAL-PRAGMATIC THEORY WITHIN THE MENTALIST PICTURE



But Stalnaker's version of the indication theory of belief is not correct as it stands. In the next few sections, I shall present four arguments in support of this conclusion; we will then be in a position to return to the question of whether these objections require a modification of Stalnaker's version of the mentalist picture, or whether they instead require replacing the mentalist picture of the relationship between thought and language with another.

4.4.2 The conjunction problem

It is widely agreed that Stalnaker's appeal to optimal conditions solves the disjunction problem.[23] What has not been noticed is that this modification of the simple causal theory only trades in the disjunction problem for what I shall call the "conjunction problem," which is equally damaging to this sort of causal theory.

Suppose that we have an agent A who believes a proposition Q. On Stalnaker's view, there must be some belief state b of A which indicates Q, so that, if we let 'O' abbreviate the predicate 'is in optimal conditions,' the following claim is true:

$$(\mathrm{O}(A)\ \&\ A \text{ is in } b)\ \Box\!\!\rightarrow (A \text{ is in } b \text{ because } Q)$$

The problem is that, if this claim is true, then so is the following:

$$(\mathrm{O}(A)\ \&\ A \text{ is in } b)\ \Box\!\!\rightarrow (A \text{ is in } b \text{ because } (Q\ \&\ \mathrm{O}(A)))$$

The first formula above says that, in the nearest possible world(s) in which A is in optimal conditions and A is in b, A is in b because Q is the case.[24] But, of course, Q is not the whole explanation for A's being in state b. It could have been the case that Q was true, and that A was not in b; A could have been tricked, or confused, or under the influence of heavy drugs. The reason why we can be sure that none of these is the case in the possible worlds under

[23] It is not that everyone agrees that the causal-pragmatic theory of belief is true; rather, everyone agrees that, whatever might be wrong with it, it is not that it makes no room for the possibility of false beliefs.

[24] As is standard, I use 'nearest possible world to w in which p is the case' as shorthand for 'the possible world most similar to w in which p is the case.'

consideration is that the antecedents of the above counterfactuals specify that A is in optimal conditions. Hence the fact that A is in optimal conditions in the worlds under consideration is a significant part of the explanation of the fact that, in these worlds, A is in b, and it is true to say that A is in b because Q is true *and* A is in optimal conditions.[25] Indeed, this is the more complete explanation.

Since this argument generalizes to all agents and belief states, this gives us the conclusion that, necessarily, for any agent a, internal state x, and proposition p,

$((\mathrm{O}a \ \&\ a \text{ is in } x) \ \Box\!\!\rightarrow (a \text{ is in } x \text{ because } p))$ iff
$((\mathrm{O}a \ \&\ a \text{ is in } x) \ \Box\!\!\rightarrow (a \text{ is in } x \text{ because } (p \ \&\ \mathrm{O}a)))$

In other words, an internal state of an agent indicates a proposition p just in case it indicates the conjunction of p with the proposition that the agent is in optimal conditions. But, given the above statement of Stalnaker's causal-pragmatic account of belief, this entails that an agent believes a proposition p just in case the agent believes the conjunction of p with the proposition that she is in optimal conditions.[26] But this is clearly false.

We can draw out a further consequence using the fact that belief distributes over conjunction. This is an independently plausible claim about belief; but in the present context it is worth noting that it need not be taken on as an extra assumption, but rather is entailed by Stalnaker's account of belief. According to Stalnaker, one can believe p either by being in a belief state x which is such that, were optimal conditions to obtain, the agent in question would be in x only because p is the case, or by being in a belief state x such that, were optimal conditions to obtain, the agent would be in x only because of something which entails p being the case. Since conjunctions entail their conjuncts, Stalnaker is committed to the claim that anyone who believes $p \ \&\ q$ also believes p and believes q; this claim, along with the conclusion of the above paragraph, entails that, necessarily, for any proposition p, if an agent believes p then that agent also believes that she is in optimal conditions. Again, this conclusion is clearly incorrect; it is not the case that, for an agent to believe a proposition, that agent must believe that she is in optimal conditions. An agent can have beliefs while believing that she has some false beliefs.[27]

How should the optimal conditions theorist reply? The natural move is to say that the fact that an agent is in optimal conditions should not be allowed to count as part of the

[25] I think it would also be true simply to say that A is in b because A is in optimal conditions; this would be only a partial explanation, but no less correct for that. I use the more conservative version in the text because that is all that is required by the argument.

[26] Strictly speaking, there is a missing step here. The causal-pragmatic theory requires for an agent to believe p not only that the agent be in some state which indicates p, but also that the agent be disposed to act in certain ways; one might think that this second clause can come to the aid of the first by ruling out states which indicate that the agent is in optimal conditions from counting as beliefs. But this is not so. The second clause of the account requires that the agent be disposed to act so as to satisfy her desires in a world in which all of her beliefs are true. But, because a world in which the agent is in optimal conditions is a world in which all of her other beliefs $p_1 \ldots p_n$ are true, if she is disposed to act so as to satisfy her desires in a world in which $p_1 \ldots p_n$ are true, she is thereby also disposed to act so as to satisfy her desires in a world in which $p_1 \ldots p_n$ and the proposition that she is in optimal conditions is true. So the simplification in the text is harmless.

[27] Note that, because the proposition that an agent is in optimal conditions is contingent, this is not a version of the well-known 'problem of deduction' for possible worlds semantics. More on this problem in §4.4.4 below.

explanation for the agent's being in one belief state rather than another; rather, we should treat these optimal conditions as 'background conditions' for the explanation.[28] To build this into the account, the causal-pragmatic theorist might then modify her account of the indication relation to say that a state x indicates a proposition p just in case, were the agent in optimal conditions, the agent would be in x only because p and the agent is in optimal conditions, where the proposition that the agent is in optimal conditions is barred from being a value of 'p.'

While this does block the above argument, it doesn't really address the underlying problem. Note that someone's being in optimal conditions is a matter of many different facts obtaining: that the agent's sensory systems are working appropriately, that there are no convincing illusions in the vicinity of the agent, that the agent is not under the influence of mind-altering drugs. The problem is that just as the general fact that an agent is in optimal conditions is part of the explanation for his being in x in the nearest possible world in which he is in optimal conditions, so each of these aspects of his being in optimal conditions is part of the explanation for his being in this belief state. But then it follows, using a line of argument exactly parallel to that used above in developing the conjunction problem, that the indication theory of content entails that each of these aspects of the agent's being in optimal conditions is also part of the content of x. And this is a mistake, for the reasons given above.

So to make her response to the initial formulation of the conjunction problem stick, the optimal conditions theorist must rule out not only the proposition that a is in optimal conditions as a possible value of 'p,' but also every aspect of a's being in optimal conditions. But there are very many such aspects; and it is not unusual for agents to believe that some of these aspects obtain. For example, it is part of my being in optimal conditions that my visual system be functioning properly. In fact, I believe that my visual system is functioning properly at the moment. But how can the modified indication theory give an account of this belief? One wants to say that I believe that my visual system is functioning properly because I am in a belief state which is such that, were I in optimal conditions, I would be in that

[28] A different line of response is to modify the definition of indication to say that what a state indicates is not fixed by its causes in optimal conditions, but rather by the facts with which the state *covaries* under optimal conditions. There are a number of technical problems with this proposal, which stem from the fact that the class of possible worlds which determine the facts with which the state covaries must be delimited in some way to exclude worlds in which the state has a different content than in the actual world. But a more pressing problem is that the proposal requires that optimal conditions must be such that, when an agent is in optimal conditions, she is not only infallible, but also *omniscient*. Were this not the case, there would be worlds in which, for a state x with content p, p is the case and yet the agent in question is not in x; but this would be enough to stop x from covarying with p in the possible worlds under consideration. Because it is hard to imagine what optimal conditions could be such that they would satisfy this requirement, it seems that the use of covariation is only plausible in the context of a theory which takes mental representations, rather than belief states, as fundamental. More on the prospects of a covariational account of the contents of mental representations below, in §5.3.4.

Neither, it should be noted, does relativizing optimal conditions to beliefs solve the problem; on this sort of revision, an exactly parallel argument may be run to show that any agent who believes p believes that he is in optimal conditions with respect to p; but this is no more plausible than the original result. And this relativization of optimal conditions is in any case less attractive when one notes that, to avoid circularity, optimal conditions must be relativized to belief *states*, rather than to the contents of those states.

belief state only because my visual system is functioning properly. But, to give this sort of explanation, the optimal conditions theorist must allow that the proposition that my visual system is functioning properly can be a value of 'p' in the above formulation of the indication theory; and this is precisely what she must deny if she is to block the conjunction problem. The class of propositions for which this problem arises will be very widespread, since there are many aspects of an agent being in optimal conditions. So it looks as though a more serious revision of the causal-pragmatic account is required to solve the conjunction problem.[29]

4.4.3 Problems with counterfactuals

Above we saw that the existence of false beliefs shows that the contents of belief states cannot be fixed by actual causal relations and that, for this reason, it is natural for the causal theorist to turn instead to causal relations in certain other possible worlds. The conjunction problem

[29] Another line of response is to question a premise on which the above argument is based: namely that, in worlds in which agents are in optimal conditions, the fact that the agent is in optimal conditions is part of the explanation for his being in a certain state. I supported this claim with the intuition that, had the agent not been in optimal conditions, he might not have been in that state; he might have been in some un-optimal condition which made him less apt to form true beliefs. This is, I think, a strong intuition. One might think, however, that it conflicts with plausible views about explanation. Consider, for example, a theory of explanation which identifies explanations with causal explanations, and identifies causation with counterfactual dependence. On such a theory, the fact that the agent is in optimal conditions in a world w is part of the explanation for his being in a certain state x only if, in the most similar world to w in which the agent is not in optimal conditions, the agent is not in x. It is certainly not clear that the latter condition is met; quite possibly, the most similar such world is one in which the agent is in nearly optimal conditions, and still forms the belief (and so comes to be in state x) as before. If so, one might reply, the original intuition should be rejected, and the conjunction problem blocked.

But to this we can make the same rejoinder as to the objection in the text above. Being in optimal conditions is a matter of many different factors obtaining; all that is required to generate the conjunction problem is that one of these factors be part of the explanation for the agent coming to be in the state in question. And it is plausible that, for any belief, there is some such factor which will meet the criterion for explanations discussed in the preceding paragraph. Consider, for example, any belief formed on the basis of vision. It is part of an agent's being in optimal conditions with regard to visual beliefs that his retinal nerve be attached; were his retinal nerve not attached, he would not have come to be in the state which he in fact came to be in on the basis of seeing something. So this is part of the explanation for his being in that state; but surely an agent can come to believe that there is a fire truck in front of him on the basis of eyesight without also believing that his retinal nerve is attached.

It's also worth noting that these strictures on explanations pose a challenge to Stalnaker's account. Stalnaker can be sure that if, in some world w, an agent is in optimal conditions and believes p as a result of being in some underlying state x, then p is the case. But it is far less obvious that, in the possible world most similar to w in which p is not the case, the agent is not in x, for that possible world might well be one in which the agent is not in optimal conditions. But if this does not hold, then, on the counterfactual theory of causal explanation under consideration, we would get the result that x does not indicate p after all. In general, responding to the conjunction problem by placing strong constraints on explanation does not appear to be a promising strategy for the causal-pragmatic account, since such constraints will likely rule out other explanatory claims needed for the account to be broad enough to cover a large class of our beliefs.

Thanks to Gideon Rosen for pressing me on this point.

shows that taking this class of possible worlds to be those in which the agent is in optimal conditions leads to an absurd conclusion; a different problem arises if we turn our attention from the specifics of this class of possible worlds to the very idea that the contents of our actual belief states are fixed by goings on in possible worlds very different from the actual world.

The problem with using facts about such possible worlds to fix the contents of actual belief states may be brought out by considering a kind of counterexample to which the indication theory of content is open. According to Stalnaker, an agent can believe p in the actual world only if she is in some state x which indicates p; that is, only if, in the nearest possible world w in which she is in optimal conditions and in x, she is in x because p. What is it for an agent to be in some belief state x both in the actual world and in some possible world w? There seems no alternative to spelling out sameness of belief states across possible worlds in terms of sameness of physical type. It cannot, after all, be done in terms of sameness of what the states indicate; the definition of indication presupposed a prior notion of sameness of belief states across possible worlds.[30] Now, the claim that a state x indicates p trivially entails that there is some possible world in which an agent is in x and p is the case. But, for appropriate values of p, there will be no such possible world. Suppose, for example, that belief states are brain states, and that an agent A believes that his brain is made entirely of silicon.[31] The proponent of the indication theory is then committed to the claim that there is some possible world in which A is in one of his actual brain states and yet it is true that his brain is made entirely of silicon. But, given that our brains are not made of silicon, this is not possible. In general, the indication theory entails that it is impossible for an agent to believe any proposition p which entails something false about the physical states that underlie our beliefs.[32]

More important than these sorts of counterexamples, though, is the underlying problem which leads to them. In the example above, we derived a contradiction from the causal-pragmatic theory along with the obvious claim that it is possible for an agent without a silicon brain to falsely believe that his brain is made of silicon. But note that we did *not* arrive at the result that it is impossible for such an agent to be in optimal conditions and to be in the state x which actually underlies his false belief about his brain; rather, we arrived at the result that there could be no such world in which the actual content of x — namely,

[30] An alternative which is theoretically open to the causal-pragmatic theorist is to spell out sameness of belief states across possible worlds, not in terms of sameness of physical type, but in terms of similarity in functional role. The problem is that specification of the functional role of a state itself presupposes some notion of sameness of belief states across possible worlds, since functional relations between states are most plausibly thought of as counterfactual relations between those states; so it seems that the same problem arises here. For more discussion, see "What is a functional role?", §4.5.1 below.

[31] The objection could be reformulated if belief states were some other sort of physical state.

[32] This counterexample is similar to the cases of altering discussed in Appendix 2 of Johnston (1993).

A separate but structurally similar problem arises when we consider the belief of an agent that she is not in optimal conditions. According to the indication theory, she can only believe this if she is in some belief state x such that, in the nearest possible world in which she is in optimal conditions, she is in x only because it is the case that she is not in optimal conditions. But this is not a possible world, since the above description contains a contradiction.

that the agent's brain is made of silicon — is true, and hence no such world in which this state of affairs is the cause of his being in x. Presumably it *is* possible for the agent to be in x and to be in optimal conditions.[33]

What, then, is the status of this state in the nearest possible world in which the agent is in this state and is in optimal conditions? It's difficult to say. Perhaps in that possible world it underlies a different belief; perhaps in that possible world it is not a belief state at all. The one thing we do know is that it is not caused by the state of affairs which, by hypothesis, is its actual content. What this shows is that there is a tension between the claim that the contents of belief states are fixed by their causes in possible worlds in which we are in optimal conditions, and the fact that properties of one's brain — including, presumably, facts about one's belief states — are among the things that would differ between the actual world and worlds in which agents are in optimal conditions. Examples of agents with false beliefs about their brains illustrate this tension in a dramatic way; but the fundamental problem is that, because reference to a belief state in the actual world and in a certain possible world must in the context of the causal-pragmatic theory be taken as reference to a certain physical state in the two worlds, there is no reason to believe that the similarities between the agent in one world and that agent in the other should be such that the causes of a state of that agent in one world should be a reliable guide to its content in the other.[34]

4.4.4 The objects of belief

A third problem for the causal-pragmatic theory arises when we shift our attention from its account of the facts in virtue of which beliefs have certain contents to its account of what sorts of things the contents of beliefs are.

Stalnaker takes his account of belief to show that the objects of belief cannot be more fine-grained than sets of possible worlds. He writes,

> . . . however we make precise the propositional relations of indication . . . in terms of which the analysis explains belief and desire, it is clear from the general schemas

[33]Note that the causal-pragmatic theory requires that, for any belief state, it be possible for an agent to be in that state and be in optimal conditions; were some state not to meet this requirement, it would indicate every proposition. But, given that belief states are physical states, this claim of the causal-pragmatic theory is very plausible; I presume that it is possible for agents to be in optimal conditions while being in virtually any physical state, given appropriate changes in other physical states and the agent's environment.

[34]This point is reminiscent of Kripke's 'infinitude' objection to idealized dispositional theories of meaning in *Wittgenstein on Rules and Private Language*. He imagines a proponent of such a theory endorsing the claim, "If my brain had been stuffed with sufficient extra matter to grasp large enough numbers . . . then, given an addition problem involving two large numbers m and n, I would respond with their sum." Kripke asks: "But how can we have any confidence of this? How in the world can I tell what would happen if my brain were stuffed with extra brain matter? . . . We have no idea what the results of such experiments would be. They might lead me to go insane" (Kripke (1982), 27). One way to see the argument of this section is as pointing out that, just so, we are entitled to wonder what we would be like were we in optimal conditions; would there be any continuity at all in our belief states? One way to see the continuity required by the causal-pragmatic theory is to note that, if any state indicates a proposition p at a world w, it must have p as its content both at w and at the world nearest to w in which the agent in question is in optimal conditions and in that state.

> for the definitions of those relations that the following will be true: if the relation holds between an individual and a proposition x, and if x is necessarily equivalent to proposition y, then the relation holds between the individual and y.[35]

Recall that, on the causal-pragmatic account of belief, an agent believes p just in case she is in some state which indicates p and is disposed to act so as to satisfy her desires in a world in which p and her other beliefs are true. Stalnaker's idea in the passage above is that it follows from this account of belief that, for any necessarily equivalent propositions p, q, an agent believes p just in case she believes q. This is because, first, if an internal state indicates p, then it also indicates every proposition true in the same states of the world as p, and, second, if an agent is disposed to act so as to satisfy her desires in a world in which p and the rest of her beliefs are true, then the fact that p and q are true in just the same worlds is sufficient to ensure that she will also be disposed to act so as to satisfy her desires in a world in which q and the rest of her beliefs are true. Stalnaker regards this as an argument for the view that the objects of belief are no more fine-grained than sets of possible worlds; I shall argue that it is better regarded as a further argument against the causal-pragmatic account.[36]

The *prima facie* problems for this thesis about belief are well known; here I'll rehearse them briefly.[37] First, note that any proposition is necessarily equivalent to the conjunction of itself and any of its necessary consequences. Hence, if Q is among the necessary consequences of P, it follows that

$$\Box\ (a \text{ believes } P \equiv a \text{ believes } P \ \&\ Q)$$

from which it follows, given the distribution of belief over conjunction,[38] that

$$\Box\ (a \text{ believes } P \rightarrow a \text{ believes } Q)$$

So, given the thesis about belief that Stalnaker derives from his indication theory of belief, it follows that belief is closed under necessary consequence: if one believes p, then one also

[35] Stalnaker (1984), 24.

[36] It's worth noting that, depending on one's view of explanation, there may be room within a causal-pragmatic theory for the view that an agent can believe a proposition p while disbelieving a necessarily equivalent proposition. It is not clear that it follows from the fact that, for two propositions p, q, p because q, that for a proposition r necessarily equivalent to q it will be true that p because r. (Some uses of 'because' in expressing the reasons why a mathematical proposition is true, for example, seem not to allow this sort of substitution.) If this is right, and 'because' sentences create hyper-intensional contexts, then Stalnaker is wrong to infer that the indication theory of belief entails that, for any necessarily equivalent propositions p, q, an agent believes p if and only if she believes q. I don't consider this possibility, since it is clear from the text that Stalnaker does not intend the 'because' in his statement of the theory to be understood in this way. If some such interpretation of 'because' could be made out, though, this might provide a way of circumventing the objection to indication theories of belief which follows in the text. (More would be required than the claim that 'because' creates non-intensional contexts; we would need the claim that it creates a context which allows all and only those substitutions permitted in the complements of belief ascriptions.)

[37] The objections are drawn from Soames (1985), (1988). To state these objections, I assume that beliefs are relations to propositions; this is common ground with Stalnaker. I do not have to assume the naive relational theory of attitude ascriptions; more on this below.

[38] For a brief discussion of the distribution of belief over conjunction and its relation to the indication theory of content, see p. 71 above.

believes all of p's necessary consequences. From this two particularly damaging consequences follow: (a) No one believes any necessary falsehoods since, all propositions being necessary consequences of a necessary falsehood, if one believed a necessary falsehood one would thereby believe every proposition; and no one believes every proposition. (b) Everyone who has any beliefs at all believes every necessarily true proposition, since all necessary propositions are necessary consequences of every other proposition. From (a) it follows that, for example, no one has ever held a false mathematical belief or believed that water is not H_2O; from (b) it follows that every creature with any beliefs believes that arithmetic is incomplete, and that water is H_2O. These conclusions seem clearly to be incorrect.

Stalnaker has, however, constructed a defense of the view that belief is closed under necessary consequence. His strategy consists in two claims, the first of which is a claim about belief ascriptions. Though he takes beliefs to be relations to propositions, Stalnaker denies the naive relational theory of attitude ascriptions: the view that an ascription $\ulcorner\alpha$ believes that $\sigma\urcorner$ is true just in case the referent of the value of 'α' bears the belief relation to the semantic content of the value of 'σ' (in the context of the ascription). Instead, Stalnaker thinks, such ascriptions sometimes report a relation to a meta-linguistic proposition about the truth of the sentence in the complement clause of the ascription. Because this proposition will always be contingent, and the possible worlds account of the objects of belief runs into trouble precisely with necessarily true and necessarily false propositions, this meta-linguistic reinterpretation promises to deliver a more intuitive assignment of truth-conditions to attitude ascriptions than the unmodified possible worlds theory.[39]

The second part of Stalnaker's strategy is a way of limiting the scope of the closure of belief under necessary consequence. Suppose that an agent believes two propositions, p, q which jointly entail a third proposition r. One might think that, by virtue of believing p and believing q, the agent believes the conjunctive proposition p & q. From this along with the closure of belief under entailment, it would follow that the agent believes r. Stalnaker replies that the agent's beliefs may be compartmentalized; the agent may believe p and believe q without ever integrating the two beliefs, and so without ever coming to believe the conjunctive proposition p & q. In this situation, Stalnaker rightly notes, we are not licensed by his theory to infer that the agent believes r.

The main problem with these two strategies is not so much that they are implausible as that they do very little to palliate the counter-intuitive consequences of Stalnaker's theory. Consider the sentence, "No whole number raised to a power greater than two is equal to the sum of two other whole numbers, each raised to that power." This is an example of a sentence which poses problems for the view of possible worlds as the objects of belief, because (i) since it expresses a necessary proposition, it follows from the closure of belief under necessary consequence that any agent who has any beliefs at all believes what it says, and yet (ii) there is no difficulty in finding an example of an agent A such that the sentence

[39]Note that Stalnaker does not deny that, for example, anyone who has any beliefs at all bears the belief relation to the (one and only) necessary proposition, expressed by, among many other sentences, "Arithmetic is incomplete"; what he denies is that, in all such cases, an ascription $\ulcorner\alpha$ believes that arithmetic is incomplete$\urcorner$ will be true.

[1] *A* believes that no whole number raised to a power greater than two is equal to the sum of two other whole numbers, each raised to that power.

is clearly false.[40] Intuitively, many agents have beliefs without believing Fermat's last theorem. The meta-linguistic strategy is designed to block our having to treat [1] as true in these cases by interpreting this sentence as attributing to *A*, not belief in the necessary proposition expressed by

[2] No whole number raised to a power greater than two is equal to the sum of two other whole numbers, each raised to that power.

but rather belief in the contingent meta-linguistic proposition expressed by the sentence[41]

[3] "No whole number raised to a power greater than two is equal to the sum of two other whole numbers, each raised to that power" is true.

Since the proposition expressed by [3] is contingent, closure under necessary consequence doesn't entail that *A* believes it; hence Stalnaker's semantics for belief ascriptions seems to provide the wanted result that [1] is not true.

A problem with this strategy of systematically reinterpreting attitude ascriptions is that, as Hartry Field has pointed out, among the beliefs possessed by agents are meta-linguistic beliefs; and this is enough to negate any advantage gained by the appeal to meta-linguistic propositions.[42] Let us suppose for purposes of the example that *A* understands the sentence which expresses Fermat's theorem; he has learned enough arithmetic to know what a whole number is, and what exponentiation is. If he understands this sentence, we may suppose that he believes the meta-linguistic proposition expressed by

[40]For simplicity I've been a bit loose with corner quotes here; strictly, the above should say that there is no difficulty in finding a name "*A*" of an agent such that the sentence ⌜*A* believes that no whole number raised to a power greater than two is equal to the sum of two other whole numbers, each raised to that power⌝ is false.

[41]There is some question what the nature of the meta-linguistic proposition is supposed to be. Sometimes, Stalnaker takes them to be about "the relation between a proposition . . . and its content" ((1986), 21). In this case, it seems, a meta-linguistic interpretation of the above ascription would attribute to *A* belief in the proposition expressed by "'*S*' means *p*." But this version of the meta-linguistic strategy will not serve Stalnaker's purposes. We are assuming that *A* understands "*S*"; hence we can assume that *A* knows what "*S*" means. But from this it follows that the ascription, so interpreted, is true. (Moreover, this sort of meta-linguistic interpretation would make true all sorts of ascriptions which are clearly false; e.g. "John believes that 2+2=5" would come out true, so long as John knows that "2+2=5" means that 2+2=5.) For this reason I shall stick with the interpretation in the text, which lets the proposition be about the truth of the representation rather than its meaning. The apparent difference between the two formulations is likely due to the fact that Stalnaker identifies meanings with truth-conditions.

[42]Field first made this point in his (1978), 38-9; he develops it further in his (1986), 111.

[4] "No whole number raised to a power greater than two is equal to the sum of two other whole numbers, each raised to that power" means that no whole number raised to a power greater than two is equal to the sum of two other whole numbers, each raised to that power.

The problem is that we already know that Stalnaker is committed to the claim that A believes the necessary proposition expressed by [2]; the meta-linguistic strategy is not a denial of the claim that there is one necessary truth and everyone who has any beliefs at all believes it, but rather a claim about the interpretation of belief ascriptions. But the conjunction of the proposition expressed by [2] with the proposition expressed by [4] has as a necessary consequence the meta-linguistic proposition expressed by [3].[43] It then follows by the closure of belief under necessary consequence from the fact that A believes the propositions expressed by [2] and [4] that A believes the proposition expressed by [3]. But, since the meta-linguistic strategy takes [1] to attribute to A belief in the proposition expressed by [3], it seems that the proponent of this strategy is forced to treat [1] as true after all. And this was the result the strategy was designed to avoid.

Stalnaker sees that this sort of response to the meta-linguistic interpretation of belief ascriptions can sometimes be made;[44] indeed, this sort of argument is one of the motivations behind the second part of Stalnaker's strategy: the compartmentalization thesis. The argument of the above paragraph moved from the claims that A believes the propositions expressed by [2] and [4] and that the conjunction of these entails the proposition expressed by [3] to the conclusion that A must believe the proposition expressed by [3]. But this sort of argument may be blocked by claiming that A's beliefs in the propositions expressed by [2] and [4] are not integrated, and hence that A does not believe their conjunction.

There is, however, an extension of Field's objection which the compartmentalization thesis seems powerless to block; for, in cases like the one we've been discussing, there is no need to integrate two beliefs. The above argument turned on the claim that

(S means p) & p

entails

S is true.

But, in cases where p is a necessary proposition, the claim that a sentence S is true is a necessary consequence of the claim that S means p alone. So the case of A's belief in Fermat's last theorem does not require any belief integration after all; all that is required for an agent to believe the theorem is for him to know the meaning of a sentence which expresses it. But this is surely a mistake; whether A is a student learning about exponentiation or a mad mathematician searching for a counterexample to Fermat's theorem, [1] must, contra Stalnaker's account, be regarded as false.[45]

[43] This is an instance of the general fact that the conjunction of the propositions that S means p and p entails that S is true.

[44] See, for example, Stalnaker (1984), 76.

[45] It should be noted that the compartmentalization strategy does rule out some problematic cases. For example, consider an agent who believes each of the axioms of some formal system; the compart-

Since the indication theory of content leads to a view of sets of possible worlds as the objects of belief, and since this view commits one to an implausible treatment of belief ascriptions, this constitutes a further argument against the causal-pragmatic theory of belief.

4.4.5 Indeterminacy and the pragmatic account of belief states

So far we have been focusing on the causal half of the causal-pragmatic theory; this half was meant to give an account of the contents of internal states. It is now time to turn our attention to the pragmatic half of the account, which tries to answer the question, 'What is it for an internal state with content p to be a belief state?'

As noted above, the causal-pragmatic strategy gives an account of what it is for an agent to have a given belief partly in terms of facts about that agent's desires; similarly, the strategy suggests an account of what it is for an agent to have a given desire partly in terms of facts about that agent's beliefs. As Stalnaker points out, there is a *prima facie* problem with accounts of belief and desire which are interrelated in this way; namely that, because both belief and desire are defined in terms of a single class of facts, there will be many different ascriptions of beliefs and desires to agents — obtained by varying attributions of beliefs and desires together — which are consistent with the theory in question.

This problem emerges if we consider a purely pragmatic theory which makes no use of facts about what internal states indicate, but instead analyzes belief and desire together in terms of an agent's dispositions. Such a theory might, following the pragmatic half of the causal-pragmatic theory, run as follows:

> An agent believes p iff he is disposed to act in ways which would tend to satisfy his desires in a world in which p and all of his other beliefs are true.
>
> An agent desires p iff he is disposed to act in ways which would tend to bring p about in a world in which all of his beliefs are true.

Suppose that an agent is disposed to ϕ. On this purely pragmatic account, this disposition is consistent with her desiring X, and believing that ϕing will bring it about; her desiring Y, and believing that Y will be realized by ϕing; and so on for any number of other such

mentalization thesis does seem to block the result that the agent must also believe all the consequences of those axioms. Even in this kind of case, though, there is some question as to whether the compartmentalization thesis might be undercut by the fact that Stalnaker's account of belief seems to imply that, in many cases in which an agent has two beliefs p and q, the fusion of the belief states which underly these two beliefs will count as a belief state with the conjunctive content $(p \,\&\, q)$. Were this the case, it would be sufficient to show that the agent not only believes p and q, but also believes their conjunction; and this all that is required to show that the two beliefs are, in the relevant sense, integrated. Consider a case in which both p and q are true, and in which the agent believes p because p and believes q because q. Then the fusion z of the states in virtue of which she believes these two propositions will be a state she is actually in because $(p \,\&\, q)$. It seems likely that, in such a case, in the nearest possible world in which the agent is in optimal conditions and in z, she will also be in z because $(p \,\&\, q)$; but this is all that is required for z to indicate this conjunction. So while Stalnaker is certainly right to claim that an agent can believe two propositions without believing their conjunction, it is an open question whether his theory really makes room for this possibility.

possibilities. Given any disposition or set of dispositions to behavior, it takes little imagination to conceive of many different sets of attributions of beliefs and desires which would fit both those dispositions and this sort of pragmatic theory. Because it seems clear that such indeterminacy would remain even given a specification of all of the agent's dispositions, and because the pragmatic theory says nothing to resolve this indeterminacy, it is, according to the pragmatic account, indeterminate which of them is true. As Stalnaker rightly says, this sort of widespread indeterminacy regarding mental states is very implausible, and is sufficient to show that the purely pragmatic theory is false.

According to Stalnaker, the causal-pragmatic account avoids this problem. Because it does not restrict itself to behavioral dispositions in giving an account of belief and desire, but also makes use of causal facts about what belief states indicate, it gives us a "fixed point with which to break into the circle that is responsible for the relativity of content."[46] Intuitively, the idea is that the causal aspect of Stalnaker's account gives us an extra constraint on attributions of beliefs and desires to agents. Since the causal-pragmatic theory is equivalent to the conjunction of the pragmatic account with the addition of a necessary condition on beliefs — the requirement that to believe p an agent must be in a state that indicates p — the question is whether this extra constraint is enough to eliminate the indeterminacy which plagues the pragmatic account.

To see that it does not, recall that, as Stalnaker notes, one can be in a state that indicates p without believing p; as he says, the reflectance properties of a bald man's head indicate features of his environment, but are not plausibly belief states. Given this fact, the causal aspect of the causal-pragmatic account does not provide an independent account of belief with which to break the circle that led to the relativity of content. Rather, it delivers an inventory of the states that indicate something; and it is then the job of the pragmatic half of the theory, given as input these facts about indication and facts about the agent's dispositions, to deliver an account of both belief and desire.

With this in mind, a slightly absurd extension of Stalnaker's 'bald man' example is sufficient to show that the causal-pragmatic theory leads to roughly the same sort of indeterminacy as the pragmatic account alone. Suppose that a bald man is playing center field in a baseball game, and that he is running toward the outfield wall with his glove outstretched. Suppose further that one of his internal states indicates, in the above sense, that a batted ball will land somewhere near the fence. In fact, though, the ball is about to hit him on the head; and, because it is a sunny day and his hat has fallen off, a state of his head indicates that the ball is about to hit him on the head. These two indicating states — one of his brain and the other of the surface of his head — are both candidate belief states. Given the action he is performing, we then have two candidate ascriptions of a belief and a desire to the center fielder. He may believe that the ball will land at a certain point near the fence, and desire to catch the ball; alternatively, he may believe that the ball is about to hit him in the head, and desire to be hit in the head with the ball while running toward the outfield fence. Of course in this case one wants to say that the first ascription is correct; but, so far as the causal-pragmatic account of belief and desire is concerned, there is no fact of the matter as to which is correct.[47]

[46] Stalnaker (1984), 19.

[47] We can also give the same sort of schematic example as was given above in the discussion of the purely pragmatic theory. Suppose that our agent is disposed to ϕ, and that she is in states which

One response to this problem on Stalnaker's behalf might be to bind the causal and pragmatic halves of the account more closely; rather than just conjoining them, a causal-pragmatic theorist might require that, for an agent to believe p, that agent must be in some state x which both indicates p and *is the causal basis* of the agent's disposition to act so as to satisfy his desires in a world in which p and her other beliefs are true. This might rule out the above example, since the reflectance properties of the outfielder's head do not seem to cause him to run toward the outfield fence.

But this is at best a partial solution. For consider a slightly modified example, in which the temperature of an agent's skin indicates to a high degree of accuracy the temperature of the surrounding air; suppose that it indicates that the surrounding air is 97.6°F. Suppose further that the agent believes that it is hot outside, but has no beliefs about the exact temperature of the air; and suppose that the agent dislikes hot temperatures, and desires to remain cool. Outside in the hot air the agent might be disposed to go find some air conditioning, and the temperature of his skin might be among the causes of his being so disposed. But then it seems that the modified causal-pragmatic theory will still deliver the unwanted result that the agent believes that it is 97.6°F outside. The root problem is the same as above: virtually every state of an agent will indicate something, and every action will have many causes; hence the overlap between states which indicate something and states which are causes of actions seems sure to include, contra the causal-pragmatic account, states which are not belief states which have as content what they indicate.

If such cases were limited to examples as recondite as the reflectance properties of the heads of bald men chasing fly balls, this result could be dismissed as theoretically unimportant. But in fact counterexamples like these will be very widespread, since, of all the states of an agent which indicate something, very few will be belief states. For consider any property F of an agent. In the nearest world in which that agent is in optimal conditions and is F, there will be *some* explanation for the fact that the agent is F in that world. But this is all that is required for the agent's being F to indicate something, and hence for F to be a candidate belief state.[48] The pragmatic aspect of Stalnaker's theory thus must rule out an enormous number of states when determining the beliefs and desires of an agent; the example discussed above shows that the constraints placed on states by this pragmatic account are not nearly strong enough. The pattern here is the same as in the case of the purely pragmatic account; because belief and desire are interdefined, we can arrive at different ascriptions of beliefs to an agent consistent with the causal-pragmatic account by making compensatory changes in the desires ascribed, and vice versa. This sort of indeterminacy is no more plausible in this case than it was in the case of the pragmatic theory.

---

respectively indicate p and q; we then want the pragmatic account of belief states to tell us which of these are belief states. In every world we may suppose that the agent's ϕing brings about *some* result; let us say that in the nearest world in which p is true it brings about $p*$ and that in the nearest world in which q is true it brings about $q*$. We then have two different possible ascriptions of beliefs and desires to the agent in question: one according to which he believes p and desires $p*$ and one according to which he believes q and desires $q*$. The causal-pragmatic account does not help us to decide between these; hence, according to this account, there is no fact of the matter as to which is correct.

[48] Further, every such state will indicate a vast number of things, since presumably the agent's being in that state in the nearest world in which he is in optimal conditions will, like most any state of affairs, have very many causes.

Lessons of the causal-pragmatic theory

These four objections are enough, I think, to show that the causal-pragmatic theory of belief should be rejected. The question then is: where should the mentalist turn next?

Stalnaker's theory of belief is roughly divisible into two halves: the causal/ indication theory of content, and the pragmatic account of belief states.[49] The problems we have encountered with each might fairly be taken by a mentalist not to count against mentalism *per se*, but rather against Stalnaker's version of mentalism, and in favor of what I called above a functional role theory of content.

To see why a functional role theory of content might be thought an improvement over the indication theory of content, consider again our solipsistic theory [T], discussed in §4.2 above. This theory took relations between belief states to be among the determinants of the contents of those states; it ran into trouble because it took *only* these relations to be relevant to content. There is no reason, though, why our choice should be all or nothing; it is worth asking whether a theory which takes both inter-belief state relations and relations between belief states and the world to be relevant to content might be more promising than one which focuses on either sort of relation to the exclusion of the other. In trying to build a theory of the contents of belief states, we might as well appeal to all of the facts about those belief states which we have at our disposal.

One might be similarly hopeful about the prospects for a functional role theory solving the problems we found with indeterminacy and the pragmatic account of belief states. In a functional role theory, the very same second-order properties which qualify a state as having a certain content also qualify that state as a belief state. So one might think that this could avoid the problems we found when we tried to combine the causal and pragmatic halves of Stalnaker's story.

4.5 Content and functional role

The functional role of an internal state is some combination of the causal relations in which that state stands to other internal states, 'input factors' like perception, and 'output factors' like action.[50] To defend a functional role theory of belief is to say that an internal state is a belief state with a certain content in virtue of its having some such functional/causal role. In trying to defend such a theory, though, we meet with a problem right at the outset in spelling out the nature of these causal relations.

4.5.1 What is a functional role?

There is a problem with the very notion of a functional role, which can be illustrated by considering the outlines of a functional role theory. The basic question is whether *actual* causal relations between belief states and other internal states and the world, or certain

[49] It is only 'roughly' divisible because the two halves cannot be separated in practice. Stalnaker's aim was not to give a content-independent account of what it is for an internal state to be a belief state, but rather to give an account of belief content and belief states which, when conjoined, would yield the right results.

[50] Recall that we are now concerned only with 'nonsolipsistic', or wide, functional role theories.

possible such causal relations are relevant to determining the content of a belief state. Neither, I shall argue, are very well suited to the purposes of the functional role theorist.[51]

It is, I think, fairly obvious that actual causal relations cannot do the trick. For beliefs, like most anything, can be caused by a wide variety of things; and only some of their causes are very relevant to their content. My belief that Princeton is a sleepy town — and hence also my instantiating a belief state which has the content that Princeton is a sleepy town — might equally be caused by seeing Nassau Street at night, by a bout of self-pity, or by a friend's remark to that effect. Just so, my belief might cause any number of things; an even more maudlin bout of self-pity, a desire to walk to the train station, etc. The things which happen to cause or be caused by a belief are not the determinants of that belief's content.[52]

This is why, I think, functional role theorists usually mean by "causal role," not actual causes and effects, but rather *potential causes and effects*. But if "potential" means "possible", this is not very much help either. Remember that we are talking about the potential causes and effects of belief states, and that these are being construed as physical states, presumably of the brains of agents. Now, one of the motivations for functionalism about mental states generally was the phenomenon of multiple realizability: the fact that a single mental state type may be realized by many different physical states. The problem is that the converse seems just as obvious: a single physical state type might, in different situations, realize many different mental states. Hence the state of my brain which in me realizes the belief that Princeton is a sleepy town might, had the world been different, have realized a different mental state: say, the belief that Trenton is a sleepy town, or that grass is green, or whatever. But if this is so, then the possible causes and effects of a given belief state of mine do not look to be very plausible candidates for determining the contents of my beliefs; how could the causes and effects of this state in a world in which it has the content that grass is green be relevant to determining the fact that, actually, it has the content that Princeton is a sleepy town?

The functional role theorist might respond by rejecting the converse of multiple realizability, on the grounds that, when we think that we are conceiving of a single physical belief state having a content other than its actual content, we are in fact imagining two distinct physical states. But this is not very plausible. The goal of the functional role theorist is to give an non-circular, explanatory account of the contents of beliefs in terms of the properties of certain internal states. If this is to be accomplished, then there must be some way of individuating internal physical states which does not rely on their supposed intentional properties. And it seems clear that, given any such way of individuating physical states, it will turn out to be metaphysically possible that these states play a quite different role in the cognitive economy of — and hence, on a functional role understanding of mental representation, have a different content for — the agent in question.

[51] This problem is often obscured by presentations of functional role theories which begin by talking about a certain kind of theory, which is then Ramsified to yield a 'definition' of belief, desire, or whatever term of the theory one is interested in. (See, e.g., Lewis (1970).) Asking what a functional role is supposed to be is equivalent to asking what sort of claims the theory to be Ramsified makes: whether claims about actual causal relations, possible causal relations, etc. I avoid the 'theory' presentation because it seems to me to introduce unneeded complications.

[52] Note that this is the case *even if we imagine the contents of all my other mental states to be fixed*. So the problem is not one which turns on the fact that we are considering the causal role of a single belief state in isolation. This is enough to show that appealing to a 'holistic network of actual causal relations' can't do the trick.

It seems clear, then, that neither the actual causes and effects nor the possible causes and effects of a belief state are likely to yield a constitutive theory of the contents of beliefs. But there is a different interpretation of causal role available: the causes and effects of a belief state in a given possible world, or set of such worlds. That is, we might consider what the causes and effects of a certain physical state might be, had certain counterfactual conditions obtained. We need neither restrict ourselves to actual causes and effects nor be so undiscriminating in our consideration of possible worlds as the alternatives considered so far suggest.

What might the relevant causes and effects be? It will be useful to begin by considering the simplest case: beliefs which are usually, or often, gained as a result of a perceptual experience.[53] Let us take as our example of a observational belief an agent A's belief that there is a red wall in front of A. Then we might hope to find some counterfactual conditional of the form

> (A satisfies such-and-such conditions & there is a red wall in front of A) $\Box\!\!\rightarrow A$ is in x

which we may take as expressing part of the causal role, in the relevant sense, of x. We might then hope to employ this claim about the causal role of a belief state on the right-hand side of an account of what it is for A to have this belief, as follows:

> A believes that there is a red wall in front of $A \equiv \exists x$
> (i) A is in x, &
> (ii) (A satisfies such-and-such conditions & there is a red wall in front of A) $\Box\!\!\rightarrow A$ is in x

This is, of course, only an account of what it is for one agent to have a single belief; nevertheless, one can see how, if something along these lines were true, it could be generalized to an account of observational beliefs for any agent. And this might be a promising start to a functional role account of belief more generally.

However, this counterfactual account suffers from essentially the same problem as its predecessor. The idea was supposed to be that in order to have a observational belief p, one must be in a physical state x such that, were one in ideal perceptual conditions with regard to p — in this case, the fact that a red wall is in front of one — one would be in x. But is there any reason to think that this is true? Recall yet again that belief states are just physical states of the brain, and consider the following case: an agent is blindfolded, and finds out that there is a red wall in front of him by being told that there is by a trustworthy friend. The version of functional role theory we are considering claims that what it is for an agent to believe that there is a red wall in front of him is for that agent to be in the same belief state he would be in, were he to be in ideal observational conditions relative to the red wall. Hence, since our blindfolded agent has accepted the testimony of his friend and believes that there is a red wall in front of him, he must, if this functional role theory is correct, *be in the same physical state as he would be in were he looking directly at the wall in good lighting conditions.* But is there any reason to believe that this must be so? Why should we think that an agent *must* come to be in the same physical state as a result of hearing the words "There's a red

[53] This way of setting up a functional role theory follows Loar (1981).

wall in front of you" as he would have come to be in by turning his gaze to a red wall in front of him?

It seems to me that to maintain that this must be so is to deny the possibility of one version of a phenomenon which was one of the original motivations for functionalism: the phenomenon of the multiple realizability of mental states. The above case plays on an intuition about multiple realizability: that there is nothing incoherent in the idea that the physical state I come to be in when hearing the words "There is a red wall in front of you" may well differ from the physical state I would have come to be in as a result of turning my gaze to that red wall, even if in both circumstances I come to believe that there is a red wall in front of me. But the functional role theorist must deny that this kind of multiple realizability is possible, because she accepts the following three propositions:

1. The blindfolded agent believes that there is a red wall in front of him.
2. An agent a believes that there is a red wall in front of him iff he is in some belief state x which has the content that there is a red wall in front of him.
3. A state x of an agent a is a belief state with the content that there is a red wall in front of a iff (a is looking at a red wall in observationally ideal conditions $\Box\!\!\rightarrow a$ is in x)

And these three propositions jointly entail that the blindfolded agent is in the same physical state x as he is in at the nearest possible world in which he is looking directly at a red wall; hence the functional role theorist must, implausibly, deny the possibility of the sort of case described above.

At this point, the functional role theorist might respond that I've mischaracterized his view. The above argument, he might point out, turns on requiring that sameness of belief states across possible worlds be a matter of sameness of physical type; but, he might continue, there is no reason for the functional role theorist to accept this. The functional role theorist should, instead, understand sameness of belief state in terms of sameness of *functional* type. With this in hand, the functional role theorist will then be in a position to accept the possibility that the belief of our blindfolded agent that there is a red wall in front of him in the actual world is realized by a different physical state than is his belief that there is a red wall in front of him in the nearest possible world in which he is ideally observationally located relative to a red wall.

This line of response, though, rests on a confusion. We began by asking what it is for a state to have a certain functional role, and arrived at the answer that for a state x to have a given causal role is for *that state* to cause and be caused by certain things in a limited set of worlds, to be determined by the relevant counterfactual. One cannot then claim that what makes a certain internal state in the relevant possible world the same belief state as x is sameness of functional type; sameness of belief states across possible worlds was *presupposed* by our account of what functional roles *are*.

Nor does this problem depend on the simplicity of the example. Here I took the relevant causal role to be exhausted by a given "input condition." A functional role theorist may think that, even for observational beliefs, the relevant functional role must be more complicated, so as to contain relations between that belief state and other internal states or dispositions to action as well as connections to observable states of affairs. But this only exacerbates the

problem because, by adding conditions to the antecedent of the counterfactual which specifies the causal role of the state, one only renders the relevant possible world more remote, and thereby decreases the plausibility of the claim that to have the relevant belief, one must be in an internal state of the same physical type as as one would be in at the possible world in question.

But what if we complicate the functional role, not by adding conditions to the relevant counterfactual, but rather by making the functional role consist of a disjunction of counterfactuals? That is, we might weaken the requirements of the theory so that an agent can believe that there is a red wall in front of him in virtue of being in a state which satisfies one of several different causal roles. For example: the functional role of being a belief state which would be caused by fixing attention on a red wall, or being a belief state which would be caused by hearing the words "There is a red wall in front of you," or .... This may seem to be just the response that the present objection requires, since it avoids the commitments to sameness of physical realizations of beliefs across possible worlds which above seemed so implausible.

The main problem with this idea is that it makes the conditions on having a belief with a certain content too weak. Given the long list of causes and effects which might be associated with any given belief, the list of functional roles used to fill in the ellipsis above will not be a short one; hence the claim, to which the current version of the functional role theory is committed, that an agent cannot be in *any* of these belief states *without* thereby believing that there is a red wall in front of him is a very strong one. There is, so far as I can see, no reason to think it plausible. (It's difficult to give a counterexample in the absence of a worked out theory; but, if the conditions are weakened in this way, then it seems likely that someone looking at an illusion of a red wall who withholds belief, knowing that he is looking at an illusion, will be counted as believing that there is a red wall in front of him.)[54]

[54]The functional role theorist might try to answer this challenge by turning to the writings of David Lewis, in which "functional role" is spelled out, not in terms of facts about an individual's belief states, but rather in terms of the typical causal relations of belief states across members of a species; to believe p is, on this interpretation, to be in a belief state x such that, usually or typically, among members of one's species, x is caused by and causes certain things. The members of one's species relevant to defining this functional role may include possible members of one's species, but "only to an extent that does not make the actual majority exceptional" (Lewis (1980), 232). But this appeal to a species seems unlikely to help, at least if the goal is to give a functional role theory of the contents of beliefs. A belief may be held by few members of a species, or by only one; in this case, the appeal to a species gains nothing. (It is not unusual for a belief to be held by only one member of a species; I suppose that many of my beliefs about myself have this characteristic.)

Further, the considerations advanced above show that there is no reason why the actual causes and effects of a belief state among members of a population should have very much in common. Suppose that every member of a group believes that gluttony is a virtue; on the species-wide version of functionalism, there must be some belief state of each member of this group which, by virtue of its usual causes and effects within the group, has the content that gluttony is a virtue. But there need by no interesting similarity between the causes and effects of their coming to have this belief; one might believe it because his older sister told him so; another on the basis of observing a rapacious and very happy diner; another because she believes the opposite of everything her teachers tell her. Moreover, there's no reason why their beliefs must have any effects in common; it does not make this story incoherent to suppose that each was struck down by God immediately upon acquiring their belief. So the appeal to a species — which is perhaps apt in the case of pain, to which Lewis applies

To these objections to an interpretation of functional role in terms of counterfactual causes, it might seem that the functional role theorist has one last reply. She might claim that the functional roles of internal states consist neither or actual nor of certain possible causal relations, but rather of certain *dispositional* properties of internal states. In general, the response might go, being disposed to ϕ when p is not equivalent to having the property of ϕing in the nearest possible world in which p is the case.[55] So we might say, for example, that an internal state x is the belief p (of some agent) iff the agent is disposed to be in x when certain conditions are satisfied without thereby committing ourselves either to these conditions actually being satisfied or to anything in particular being the case in certain far-off possible worlds.

As applied to the present case, this move from counterfactuals to dispositions might take the following form:

> a believes that there is a red wall in front of him $\equiv \exists x(a$ is in x & a is disposed to be in x when ideally observationally located relative to a red wall)

But it seems that the same considerations which counted against the counterfactual version of functional role theories count against this one; if I am blindfolded and told that there is a red wall in front of me, there is no reason to think that I must then be in a state with the above dispositional property. The only difference between this and the counterfactual version is that we cannot now defend this intuition by spelling out its implications in terms of possible worlds.[56]

It goes without saying, moreover, that the belief that there is a red wall in front of one is supposed to be an easy case for the functionalist; about the easiest there is. The case is much more clear cut if beliefs further removed from experience are used as examples.[57]

---

this idea — is not of any help to the functional role theorist we're now considering. One might think that appeal to a linguistic community rather than a whole species might be more effective. It seems to me that many of the same criticisms would apply; but, more importantly, this would amount to the abandonment of mentalism, since it would be a way of taking facts about linguistic meaning as relevant to the determination of the contents of beliefs.

[55] See Appendix E for some brief discussion of the implications of this point.

[56] Could it turn out that the right account of dispositions would count against this intuition? So far as I can see, nothing rules this out; but there seems no reason at present to think that it will.

[57] These problems with the notion of a functional role are made considerably worse, moreover, by the problem of error as it arose for the solipsistic theory discussed above: the problem of making room for faulty inferences. Just as agents often have false beliefs, so they often come to have beliefs for bad reasons. This is problematic for the functional role theorist because she holds that the content of a belief state is fixed by its causal relations to perceptual input, other internal states, and action; but if agents are capable of mistaking the evidence of the senses, incorrectly forming beliefs on the basis of other beliefs to which they bear no special relation, and and acting irrationally, it seems clear that such an agent could believe p *without* having any belief state with the functional role associated with p. What this shows is that functional role theories of content, like solipsistic theories and causal theories, need to find a way to make room for the possibility of certain kinds of error. And, as noted above, the usual way to do this is by appeal to some class of optimal conditions: conditions under which an agent is not disposed to make such errors. This makes matters worse for the counterfactual version of functional role theory because adding the requirement that the agent be in optimal conditions to the antecedent of the counterfactual which defines the functional role for some belief p renders the possible world w relevant to determining whether some actual agent

At this point, facts about belief may begin to appear a bit mysterious. We seem to be able to hit on *nothing* that everyone who believes p has in common. But how could this be? Mustn't there be *something* in virtue of which an agent has certain beliefs? Later I'll argue that this air of mystery arises from the fact that the mentalist is simply looking in the wrong place for facts which determine the contents of an agent's mental states; for now, the important point is the negative result that there is no interpretation of "functional role" which will serve the purposes of the functional role theorist. This is an argument against functional role theories of the contents of beliefs which is different in character from most other arguments which have been levelled against functionalism. It does not depend upon contentious assumptions about "qualia," or upon any difficult to imagine thought-experiments. The argument is simply that any way of making the notion of a functional role precise makes the theory false.[58]

4.5.2 Commonsense functionalism and psychofunctionalism

This is a general problem about what sort of thing a functional role is; we know that it has *something* to do with causal relations, but it seems that no combination of the causal relations between belief states and other things can be necessary and sufficient for a belief state to have a given content. A further problem comes when we try to say what the relata of these causal relations should be. In this section, I shall argue that, even setting aside the problems of the previous section, there are no connections between the belief states, other internal states, perceptions, and actions of agents systematic enough to determine the contents of belief states. In doing so, I'll loosely follow the most developed version of the functional role theory available: that proposed by Brian Loar in his *Mind and Meaning*.

The goal of a functional role theory of content is, for any proposition p, to specify facts about a functional role F such that it is necessary and sufficient for an agent to believe p that that agent be in some internal belief state which satisfies F. This is to be accomplished via counterfactuals like the input conditions for observational beliefs sketched above; these counterfactuals specify the functional role which a belief state must have for it to be constitutive of an agent believing a given proposition.[59] The general method for arriving at these counterfactuals is to begin with a counterfactual claim relating facts about the *beliefs* of an agent to other facts — perhaps facts about other propositional attitudes, the agent's dispositions to behavior, or objects in the world causally acting on the agent. This claim about beliefs is then transformed into a claim about belief *states* by replacing each mention of a belief by an existential quantification over the belief states of the agent.[60] This latter claim then specifies

believes p more remote from the actual world, and in so doing renders the claim that the actual agent be in the same physical state as he would be in at w even more implausible.

This shows the similarity of the present argument against functionalism to the argument raised against causal theories of content in 4.4.3 above.

[58]The argument is, however, restricted to functional role accounts of the contents of mental states; I am unsure whether a version would work against functionalism about 'phenomenal states' like feeling pain.

[59]As discussed in the previous section, there are problem to do with this appeal to counterfactuals; I use it here because it is, I take it, the most promising of the interpretations of "functional role" sketched above.

[60]Again, this is often discussed in terms of a 'theory' of belief and other propositional attitudes, which is just a conjunction of claims about counterfactual relations between beliefs and other facts,

the functional role a belief state must have to have a certain content: it must be such as to make the claim true, when it is the value of the relevant variable.

Thus, in the case of an 'observational' belief p, we might begin with the claim about belief that, were p the case and an agent in such-and-such conditions relative to p (e.g., 'observationally ideal conditions,' whatever these might be), then the agent believes p. Replacing mention of the belief p with existential quantification over belief states, we get the following counterfactual claim which tells us what functional role a belief state must have to have the content p:

> a believes $p \equiv \exists x$ ((p & a satisfies such-and-such conditions relative to p) $\Box\!\!\rightarrow a$ is in belief state x)

It seems clear that any functional role theory of content will have to include some such clause for observational beliefs; as Loar notes, any explication of the notion of causal role must relate belief states "to something *external* if suitable asymmetries among beliefs are to be found which, directly or indirectly, give each belief a functional role."[61] The idea behind a functional role theory which takes observational beliefs as primary will be to come up with similar counterfactuals which relate observational beliefs to other beliefs, and, transforming these counterfactual claims about the relations between beliefs into claims about the relative functional roles of belief states, to thereby define, for each possible belief, the functional role a belief state must have in order to be that belief. In other words, we work our way up from observational beliefs to non-observational beliefs.

The route Loar suggests seems the most plausible one. His functional role theory involves specifying two kinds of counterfactual claims to derive the relevant causal roles: rationality constraints and meaning constraints. The rationality constraints are claims like

> a believes p $\Box\!\!\rightarrow \neg$(a believes $\neg p$)[62]
>
> (a believes ($p \rightarrow q$) & a believes p) $\Box\!\!\rightarrow a$ believes q

The meaning constraints rest on the idea that the ability to make certain inferences is a necessary condition on possession of concepts of a certain type;[63] they include claims like

> (a believes that (x is north of y & y is north of z)) $\Box\!\!\rightarrow a$ believes that x is north of z
>
> a believes that x is next to y $\Box\!\!\rightarrow a$ believes that y is next to x

which is then Ramsified to yield claims about functional roles. But this talk of theories is dispensable, and the formulation in the text seems to me a simpler way of putting what is going on in functional role theories.

[61] Loar (1981), 65.

[62] This is ill-formed. Since 'p' is being used here, as elsewhere, as a variable over propositions rather than over sentences, the formula inside the parentheses has the same syntax as "a believes not that S." It should be read as shorthand for $\ulcorner \alpha$ believes that $\neg\sigma \urcorner$.

[63] Loar ((1981), 83 ff.) takes meta-linguistic M-constraints to be a key part of his functional role theory. Discussion of this would introduce unneeded complications at this point; I don't think that any of the objections I raise below will turn on omission of this aspect of Loar's theory. The main difference which results is that without these sorts of M-constraints, we won't be able to achieve a goal which Loar claims his theory accomplishes: that the functional role theory should say something unique about the functional role constitutive of each belief. This failure of 'systematic uniqueness' won't be important for what follows.

It is important to remain clear how these sorts of 'platitudes' about belief figure in a functional role theory. These constraints give us supposed facts about belief which internal states must 'satisfy' to be the relevant beliefs. The sense in which internal states must satisfy these constraints is that they must be related to each other (and input and output) in the same way the platitudes claim that the relevant mental states are related.

The obvious way to use these constraints in giving a constitutive account of belief, then, is to use them to define functional roles satisfaction of which is necessary and sufficient for belief states to have a certain content. Again, this is done by the method of transforming these counterfactuals about beliefs into claims about the causal roles of belief states illustrated above. If we take as given the claims about the relations between causal roles of belief states and their contents provided by the input conditions for observational beliefs, these counterfactuals give us ways of deriving facts about the relations of causal roles of non-observational belief states and their contents. Suppose, for example, that we already know that some agent A believes p iff she is in belief state x, and that she believes $(p \rightarrow q)$ iff she is in belief state y. Then the second rationality constraint above will tell us what it is for A to believe q; she believes q iff, for some belief state z of A, were A in x and y, then A would be in z.

But, as Loar has noted, it seems quite clear that these sorts of rationality and meaning constraints will not be informative enough to give us sufficient conditions for an agent to believe an arbitrary proposition p. This is, I think, intuitively clear from the constraints listed above; how *could* these constraints do this much work? But the point can be illustrated by considering the above example of the second rationality constraint. We are told that A believes q iff she is in a belief state z such that were A in x and y, then A would be in z. But note that, in the counterfactual scenario being considered, A is in *many* such belief states. If we take this condition on believing q to be sufficient, then, we get the result that A actually believes q if she is in *any* of these belief states. But this is clearly incorrect.[64]

A second, better way for the functional role theorist to proceed is to give up the idea that these sorts of rationality and meaning constraints will by themselves provide us with necessary and sufficient conditions for belief, and instead make the more modest claim that they give us necessary conditions on belief. This is Loar's idea; we should take the input conditions for observational beliefs, rationality constraints, and meaning constraints as jointly determining a meta-condition which any theory of content for an agent must satisfy.[65] Loar's thought is that a functional role theory of the sort sketched above will tell us several conditions that any answer to this question must satisfy; in terms of the example of the previous two paragraphs,

[64] We might, following the statement of Stalnaker's indication theory, revise the account so as to say that a believes p iff there is some belief state x of a such that, were a in such-and-such conditions, a would be in x *because* p is the case. But this would not solve the problem; and, as applied to the rationality and meaning constraints, makes the conditions far too strong. Take the example of the second rationality constraint, discussed above. It is surely possible that an agent come to believe p, later come to believe q, and, as a result of this latter belief, come to believe the conditional $p \rightarrow q$. But the rationality constraint, if revised in accordance with the current suggestion, would require that the belief state x which is, for some agent, constitutive of believing q be such that, in the nearest possible world in which that agent believes p and $p \rightarrow q$, that agent comes to be in state x *as a result of* believing p and believing $p \rightarrow q$. Since the nearest such world is the actual world and the agent's beliefs do not display this pattern of causal relations, the constraint gives the wrong result.

[65] Loar (1981), 78.

it will tell us that any belief state z which realizes A's belief q must satisfy the following condition: it must be the case that, were A in x and y, then A would be in z. But the theory, by giving only necessary conditions, avoids having to say that *any* belief state which satisfies this condition is the belief q.

Suppose that we grant this point: input conditions, along with rationality and meaning constraints, give us necessary conditions for a belief state having a given content. This still leaves us with a problem: how are we to give sufficient conditions for a belief state having a given content? Loar's idea is that this is to be done, not by coming up with more 'commonsense' constraints, but by appealing to the deliverances of theoretical psychology.[66]

Intuitively, the idea is that we have so far been proceeding using only armchair psychology. We have certain 'commonsense' views about how certain mental states are related, and can use this to get some information about the internal states which realize those mental states; but, as we have seen, armchair psychology does not give us enough information to give a full constitutive account of belief. No surprise, one might think; all this shows is that we need to make use of the results of theoretical and experimental, as well as armchair, psychology. But, while this psychofunctionalist proposal is hard to evaluate without a theory on the table, I think that there is a kind of incoherence in this appeal to theoretical psychology.

To see this, recall how we arrived at the counterfactuals which gave us links between the functional roles of belief states and their contents in the first place. We began with counterfactual relations between *beliefs* of agents and other facts about that agent; only given such counterfactuals could we derive claims about the contents of belief *states* by removing the references to beliefs and replacing them with existential quantification over belief states. So where did we get the claims about beliefs in the first place? We got them from the place where functionalists have traditionally looked: from generalizations about belief implicit in our ordinary talk of belief. But we have just seen, and Loar admits, that this 'commonsense functionalism' does not have the resources to give us these counterfactual relations; this was the point of the claim that our functional role theory — the input constraints, meaning constraints, and rationality constraints — is not strong enough to give us sufficient conditions for have a certain belief. So it is clear that we need to find some more of these generalizations about beliefs; and we know that we can't get them from 'commonsense' psychology.

The present problem is that it is difficult to see how theoretical psychology could help us in this case. In the best case scenario, theoretical psychology might provide us with an inventory of the belief states of an agent, along with a list of counterfactual causal relations involving these belief states. Some will be of the form, if the subject were in state x, then the subject would be in state y; others will be of the form, if the subject were attentive and in the presence of such-and-such an object, then the subject would be in state z; and so on. But these sorts of claims simply do not tell us *anything* about the conditions under which an agent believes some proposition p; they tell us about the causal relations between belief states, without telling us how these relations are to determine the contents of those states. Facts about the counterfactual relations between belief states yield information about content only given facts about counterfactual relations *between beliefs* of the agent, so that we can assign propositions to belief states on the basis of a correlation between the two sets of counterfactual relations. But in the task of coming up with such relations between beliefs, it is obscure how theoretical psychology could succeed where commonsense functionalism has

[66] Loar (1981), 79-81.

failed. It will not, to be sure, get them from counterfactuals relating occurrences of belief states; without a theory of the contents of those states, no such generalizations about beliefs will be forthcoming, and without such generalizations, we cannot get the needed theory of content. So 'commonsense functionalism' cannot succeed; and it is mysterious how an appeal to a scientific 'psychofunctionalism' could offer any assistance.[67]

Conclusion

The mentalist picture of intentionality consists of two main tenets: (i) the fundamental kind of intentional or representational fact is that agents believe, desire, and intend certain things, and (ii) the meanings of expressions in shared, public languages are derived from these propositional attitudes. As such, mentalism is one of two ways of thinking about the relationship between language and the mind consistent with the thesis of the priority of the mental. I have now argued that both (i) and (ii) are false.

Thesis (i) is false because, as argued in the first section of this chapter, (i) entails functionalism: the view that what it is for an agent to believe p is for that agent to be in an internal state x with a second-order property in virtue of which x is a belief state and has the content p. But, as argued in the remainder of this chapter, there is no second-order property acceptable to the mentalist which can do this work.

Thesis (ii) is false because, as argued in Chapters 2 and 3, all attempts to reduce public language meaning to the propositional attitudes of agents meet both persistent counterexamples and problems of principle. Among the latter were that the mentalist seems either to vastly over-intellectualize language use or to run afoul of the problem that facts about meaning are more fine-grained than facts about beliefs.

A negative verdict on the mentalist picture of intentionality raises the question of where this picture went wrong. At this stage of the argument, I think that there are two plausible answers to this question.

First, one might draw a communitarian moral from the failure of mentalism: one might think that the falsity of (i) and (ii) show that the order of explanation between mental states and public language is precisely the reverse of that endorsed by the mentalist. One might be inclined to think this on the basis of the various cases explored above in which the propositional attitudes of agents — what they assert, mean, and perhaps also, believe — seem to depend upon their linguistic behavior, and the meanings of sentences which figure in their linguistic acts.

But one might think that the falsity of (i) and (ii) point in the direction of quite a different picture of the relationship between language and mind. On this view, the falsity of (ii)

[67]Note also that the meta-condition jointly imposed by the input conditions, rationality constraints, and meaning constraints is not as strong as it might appear. These conditions, after all, will clearly not be exceptionless; it is possible for people believe contradictions, or believe a conditional and its antecedent without believing its consequent, or believe that x is north of y and y north of z without believing that x is north of z. It seems quite possible, moreover, for a member of some community to be in such a position fairly routinely. Suppose, then, that theoretical psychology delivers a theory about the psychology of an agent which entails that he routinely violates the rationality constraints, or meaning constraints. Are we then in position to reject the theory? It seems not, absent a reason to doubt that an individual can violate these constraints. But then it seems that the functionalist is forced to retreat to a position rather near the claim that functionalism is true, though we as yet do not have much of an idea about what the true functionalist theory will look like.

does show that propositional attitudes are not the most fundamental level of representation; but, on this view, the contents of these attitudes are derived from the contents of mental representations in private or individual languages, rather than the meanings of words in public languages. And this response to the falsity of (ii) suggests a complimentary view about the falsity of (i): if we need to posit meaningful private languages in order to give a constitutive account of belief, then perhaps we can also use these private languages, or idiolects, to give a more satisfactory story about public languages than the mentalist was able to offer. This *private language picture* of the relationship between mind and language is another way of substantiating the thesis of the priority of the individual; it is the topic of Part II of this essay.

Part II

THE PRIVATE LANGUAGE PICTURE

Chapter 5

Belief and Mental Representations

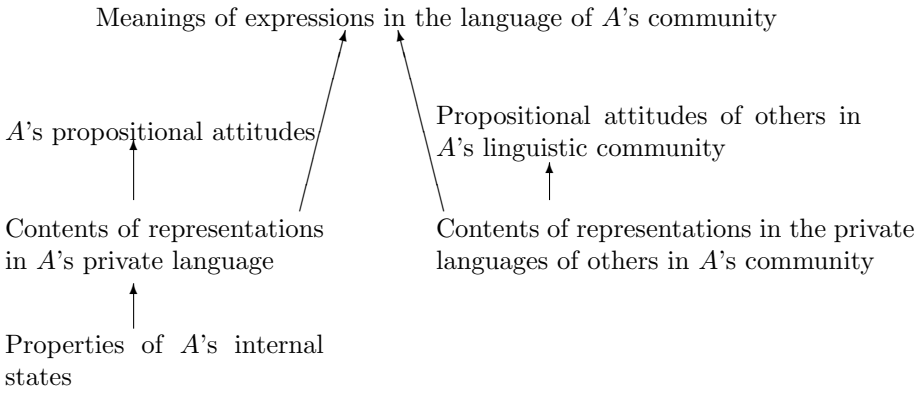
Contents

We concluded our discussion of the mentalist picture of intentionality with the claims that the mentalist can give adequate constitutive accounts of neither belief nor public language meaning. For a number of reasons, the private language picture appears well-positioned to remedy this lack.

Recall that, on the private language picture, the most fundamental facts about intentionality are about the contents of representations in the individual languages — depending on one's terminology, these are idiolects, languages of thought, or I-languages — of agents.

Schematically, the relationship between various sorts of facts about mind and language presented by the private language picture is as follows:

THE PRIVATE LANGUAGE PICTURE



One of the problems with mentalist views of public language meaning was its tendency, in appealing to various kinds of convention, to over-intellectualize language use by requiring that speakers of a language have implausibly complex beliefs about that language. It is plausible to think that the proponent of the private language picture can avoid this problem without much trouble: if she can give an adequate account of the contents of representations in private languages, she can treat public languages shared in a group as mere abstractions from those private languages without appealing to complex meta-linguistic beliefs of speakers. I discuss the possibility of giving such an account in Chapter 6, "Public and Private Languages."

One of the problems with mentalist views of the mind was that, on some views, the contents of beliefs ended up being no more fine-grained than sets of possible worlds. We discussed some of the problems with this view above.[1] Private language theorists are well-positioned to avoid this result; if belief states are thought of as structured states whose parts have certain contents, the contents of beliefs can be structured propositions, whose constituents are the contents of the constituents of the relevant belief state. In addition, as we'll see below in this chapter, the private language theorist has more resources for responding to the objections levelled in Chapter 4 against mentalist theories.

Perhaps most importantly, the private language theorist can, like the mentalist, endorse the thesis of the priority of individuals over social groups in the explanation of intentionality. The intuitions behind this kind of individualism are strong; individualism along with the falsity of mentalism are enough to provide a kind of *prima facie* argument for the private language picture of intentionality.

In this chapter and the next, I consider the question of whether this *prima facie* plausibility can be substantiated by an account of the contents of representations in private languages suitable for giving constitutive accounts of belief and public language meaning.

[1] See §4.4.4, pp. 75 ff.

5.1 Theories of belief and theories of content

What form might a private language account of belief take? Recall that in the previous chapter I argued that any plausible account of belief consistent with mentalism would have to be some form of nonsolipsistic, or externalist, functionalism. It seems clear that that result carries over to the private language picture. The premises needed to rule out behaviorist and identity theories of belief — crucially, the priority of belief over public language — are endorsed by the private language theorist.

I mentioned above that any functionalist account of belief owes some story about what it is for an internal state to have a given content, and what it is for such a state to be a belief state. The main difference between the sorts of functionalist accounts developed by the private language picture and by the mentalist come in the specification of the contents of belief states. Private language theorists, after all, take the fundamental units of meaning to be sub-sentential expressions in a private language — i.e., *mental representations* — which are constituents of belief states. This means that a foundational account of mental content of the sort which a private language theorist will propose will define content first for mental representations, and next for belief states in terms of the contents of the mental representations which are constituents of that state.[2]

This means that, while the mentalist's account of belief was of the form

a believes $p \equiv \exists x$ (x is a belief state of a & x has the content p for a)

the kind of functionalism proposed by the private language theorist will be something more like

a believes the structured proposition $< F_1, \ldots, F_n > \equiv [\exists \mu_1 \ldots \mu_n$
(1) $< \mu_1, \ldots, \mu_n >$ is a belief state of a,
(2) μ_1 has the content F_1 for a,
. . . . . ,
. . . . . , &
(n+1) μ_n has the content F_n for a]

In order to fill out this schema, a private language theorist must do two things: say what it is for a mental representation of an agent to have a certain content, and say what it is for an agent to have a belief state consisting of some set of mental representations.[3]

[2] I intend this use of 'constituent' to be fairly loose; there is no reason to pin on the private language theorist any particular view of how mental representations relate to the belief states in whose contents their contents figure.

[3] Actually, there is one more task for the private language theorist, which I will ignore for simplicity in what follows. In order to give an account of belief, the private language theorist must say not only what it is for a mental representation to have a given content and what it is for an agent to have a belief state consisting of some set of mental representations, but also say what it is for an agent to have a belief state consisting of these mental representations *as combined in a certain way*. In general, a set of contentful mental representations will not determine a unique proposition. To take a simple example, an agent may have a belief state whose constituent mental representations have as their content John, Mary, and the relation expressed by "loves"; this will be consistent with the agent believing one or both of the propositions that John loves Mary and that Mary loves John. So the private language

Among private language theorists, work on the first question has largely proceeded independently of work on the second; thus there are many theories of 'mental content' in the field which are not explicitly attached to any theory of what it is for a state which, according to the theory, has a given content, to be a belief, or desire, or some other mental state. It seems to me that this is a flaw in the methodology of contemporary philosophy of mind, if only because it is hard to see what pre-theoretic data a theory of content *simpliciter* is supposed to answer to. After all, we have pre-theoretic convictions about which beliefs, desires, and intentions agents have; and these convictions would provide reasonably strong evidence for or against any given functionalist theory. But these pre-theoretic convictions do not count for or against theories which simply entail facts about the contents of states, without entailing anything about what agents believe or desire; and it is not obvious, absent these convictions, what data there is for the testing of such theories.

This presents something of a quandary. Our aim is to evaluate the prospects of private language theories of mental content; but it is difficult to know how this can be done without embedding these theories of content in theories of belief states, intention states, and so on which are often not provided by private language theorists. It will be worth bearing this in mind throughout this chapter. I will be testing various theories of mental content against what I take to be obvious facts about the beliefs of agents; but, in each case, this will involve going beyond those theories of mental content to derive facts about belief.

5.2 Mental representations and the constraints on constitutive accounts

Before turning to particular views of mental content, it is worth discussing a general objection to the use of accounts of mental representations in giving a constitutive account of belief.

If our aim is to give a constitutive account, then our account of what it is for an agent to have a given belief must apply to actual and possible believers, and hence must make no use of contingent psychological claims particular to some proper subset of those agents.[4] This presents an immediate problem for the private language theorist if, as some have thought, the thesis that beliefs are underwritten by complex internal states whose constituents must stand in certain specific relations to objects and properties in the world is just such a contingent psychological claim. Stalnaker, for example, writes of what I have been calling private language accounts of belief that

> It is important to recognize that the suggestion being made is not just a claim about what is going on in the believer; it is a claim about what a belief attribution says about what is going on in the believer. ... According to this suggestion, if I say

theorist owes an account of what makes it the case that a belief state comprised of a number of mental representations has as its content one function of the contents of those representations rather than another; i.e., she must define a function F from mental representations to propositions which makes the following true:

> a believes $p \equiv_{df} \exists x \exists \mu_1 \ldots \mu_n \exists q_1 \ldots q_n$ ((x is a belief state of a) & (x has as parts mental representations $\mu_1 \ldots \mu_n$) & ($\mu_1 \ldots \mu_n$ are mental representations with contents $q_1 \ldots q_n$) & ($F(\mu_1 \ldots \mu_n)$=p))

[4] For discussion of this point, see §1.2 above, especially pp. 6 ff.

> that x believes that P, my claim will be false if the form in which the informational content of *that P* is stored is relevantly different from the form of the clause "that P." I think this suggestion makes a belief attribution carry more weight than it is plausible to assume that it carries. If it were correct, belief attributions would be far more speculative, and believers far less authoritative about their beliefs, than they seem to be. While theoretical and experimental developments in cognitive psychology may someday convince me that I store my beliefs in a form that is structurally similar to the form in which they are expressed and described in English, I don't think that my ordinary belief attributions commit me to thinking that they will.[5]

Now, there are some grounds for skepticism about the intuitions Stalnaker expresses in this quote. In particular, a private language theorist is not likely to be moved by Stalnaker's implication that the private language view should be rejected because it is implausible to think that ordinary speakers have complex mental representations in mind when attributing beliefs. After all, the private language theorist under discussion is committed to giving a constitutive account of belief in terms of such mental representations, but need not make the further claim that this constitutive account provides an analysis of the *meaning* of belief ascriptions, or of what what speakers mean by uttering them.

But there is a better, and simpler interpretation of Stalnaker's main thought here, which comes out most explicitly in the last line of the quote: it is implausible to think that our ascriptions of beliefs to agents would all be false if it turned out that those agents failed to satisfy some fairly specific psychological theory.[6]

In response, the private language theorist is likely to accuse the proponent of Stalnaker's position of confusing epistemic for metaphysical possibility. Surely, she might say, we can endorse the claim that if it had been the case that actual agents did not fit some psychological theory, our belief ascriptions would not all have been false. But this is rather like saying that if the clear, drinkable, liquid in the lakes and rivers had been XYZ rather than H_2O, then our water ascriptions would not all have been false. True enough; but this does not show that water could have been XYZ. It only shows that, had the actual world been different, our word 'water' would have picked out a different kind. Just so, the objection continues, our intuitions about belief ascriptions do not show that it is really *possible* for agents to have beliefs without having mental representations which satisfy some psychological theory; all they show is that, if actual agents had failed to satisfy that theory, our word 'believes' would have picked out a different kind.

There are, I think, two replies to this objection on Stalnaker's behalf, neither of which is really conclusive. The first reply is to cast doubt on the assimilation of 'believes' to natural kind terms. It is natural to think that the model of natural kind terms invoked by the objector rests on the view that such terms have a certain property not shared by all expressions of English: the property of having their extension determined by the physical constitution of some paradigm sample of the kind, even if speakers who use the term know very little about

[5] Stalnaker (1990), 230.

[6] Given Stalnaker's own views about belief, it is implausible to think that he really had in mind a claim about the meanings of belief ascriptions; it is, after all, no more plausible to think that ordinary ascribers of beliefs have in mind facts about what internal states indicate than that they have in mind complex facts about mental representations.

what this physical constitution is. Now, we can ask: what is it about speakers of the language that determines whether a given term is a natural kind term or not? So far as I know, there is no accepted answer to this question.[7] But one partial answer has it that it is sufficient for a term to be a natural kind term for speakers to introduce the term with certain intentions, such as the intention that the term refer to all and only those substances of the same kind as the items in some initial sample. No doubt this is at best an idealized model of the introduction of natural kind terms. But, as an idealization, it does not seem altogether implausible; it might be, for example, that speakers always had linguistic dispositions with respect to natural kind terms which had something to do with the basic physical properties of the stuff. The question is whether this model is very plausible when applied to 'believes'. It seems to me that it is not; but, lacking an adequate foundational account for the semantics of kind terms, this can only be regarded as an intuitive doubt.

The second reply is just to directly challenge the claims about what is and is not metaphysical possible to which the private language theorist is committed. To see that this does have some force, it is worth being clear on what the private language theorist is committed to. Such a theorist is not committed just to the claim that any possible believer should have mental representations, where this is construed as the claim that any possible believer should have some internal states which have parts and are related in some way or other to the beliefs of the agent. It is plausible that this *is* a necessary truth.[8]

Rather, the private language theorist is committed to the much stronger claim that any agent capable of beliefs must have mental representations which are related in a certain way to the environment of the agent. Suppose for illustration that a private language theorist presents a constitutive account of belief which involves the claim that a mental representation has a property as its content just in case that representation bears R, a certain kind of causal relation, to the property. Such a theorist is then committed, by the constraints on constitutive accounts, to the claim that any possible believer must process information in this way: by having certain parts of her cognitive system be R-related to parts of her environment. But, on the face of it, this looks like a case of mistaking the contingent for the necessary akin to the mistake of the identity theorist. Just as different physical states can realize different mental states, why not think that different creatures might acquire and process information from their environment in quite different ways? If this is a real possibility, then private language theorists have no promising way of giving a constitutive account of belief (or of any other sort of intentional fact, for that matter).[9]

There is, then, some reason to be skeptical about whether one of the key notions needed to develop an account of belief in keeping with the private language picture should have any

[7]But for some plausible suggestions see Soames (2002), Ch. 10, "What do Natural Kind Predicates Have in Common with Proper Names?", especially pp. 281 ff.

[8]This is roughly the sense in which, e.g., Schiffer regards the thesis that there is a language of thought as fairly trivial (see Schiffer (1993)). The concession here is given modulo worries about simple and immaterial beings who have beliefs being metaphysically possible.

[9]It is worth noting that many accounts of the contents of mental representations do not purport to be giving metaphysically necessary and sufficient conditions; the present objection is not an objection to such accounts, just as it is no objection to the use of mental representations in cognitive psychology. The point is just that, if this objection is right, then one interested in questions like 'What is the nature of belief?' or 'What is it for an agent to take the world to be a certain way?' should not look to mental representations and their second-order properties for answers.

role in a constitutive account of belief. But this worry derives from modal intuitions which, trustworthy though they seem to me, are denied by proponents of the private language picture, and are difficult to argue for. So for the remainder of this chapter I will set these worries about appeals to mental representations aside, and ask: on the supposition that mental representations have a role to play in a constitutive account of belief, is there any way of giving an account of the contents of an agent's beliefs in terms of second-order properties of her mental representations?

This question is parallel to the question which we explored in the previous chapter; there we asked whether, on the supposition that the beliefs of agents are underwritten by internal belief states, we could give any account of an agent's beliefs on the basis of second-order properties of internal states. I argued then that the answer to this question was "No," and that this gives good grounds for rejecting the supposition that beliefs are constituted by second-order properties of belief states. In this chapter I shall argue for a similar negative conclusion which, I claim, will then license us in rejecting our supposition that the beliefs of agents are constituted by second-order properties of their mental representations.

There is a second similarity to the structure of the preceding chapter worth noting. There I pointed out that the mentalist might take the relational properties of belief states relevant to determining the content of a belief to be either (i) external relations between those belief states and the world, or (ii) such external relations along with internal relations between belief states and other states of the agent.[10] The same choice, as applied to mental representations rather than to whole belief states, arises here. Following the terminology introduced in §4.3 above, I shall call theories of type (i) informational accounts of belief, and theories of type (ii) conceptual role accounts of belief.

5.3 Mental representations and information

The general form of an informational account of mental content is the same as that of an indication account of mental content. In both cases, the goal is to define the content of some internal state in terms of relations between the state and the environment of the agent. The only difference is that in the case of indication theories of content, the states are belief states, whereas in informational theories of content, the states are mental representations, thought of as constituents of those states.

It may seem as though this difference between the two kinds of theories would be of little significance. But, as it turns out, it makes an enormous difference. As we will see below, informational theorists have many more choices for explicating the relevant relation between mental representations and the world than did indication theorists; and this puts them on stronger footing in responding to certain kinds of objections. But the shift from belief states to mental representations also opens the informational theorist to an objection to his whole enterprise. This objection is the topic of the next section.

[10] See p. 65. As before, I am assuming the falsity of solipsistic theories of belief, for the reasons discussed in §4.2 above.

5.3.1 What is 'tokening a mental representation'?

This general objection is a way of picking up on the objection discussed in §5.2, to the effect that using mental representations in a constitutive account of belief is a way of mistaking (at best) contingent truths about belief for necessary truths. We can see how this objection arises in a particularly difficult form for the informational theorist by asking about the form of an informational theory of mental content.

The informational theorist takes the contents of mental representations to be fixed by some relation R between those mental representations and objects and properties in the world. So one might think that the form of a informational theory, for any mental representation μ, agent A, and content F, will have the form

> μ has content F for $A \equiv \mu$ bears R to F

But a quick example of a state of affairs which should be consistent with informational theories shows that we need a bit more information about what it is for a mental representation to bear a certain relation to a property.

Suppose that an agent has a stockpile of mental representations in his brain, which correspond to words of English: he has a 'cow' mental representation, a 'horse' mental representation, and so on. The agent, being very simple, forms beliefs only when he has a perceptual experience of something, and always when he has a perceptual experience of something; and the agent, being very lucky, only has veridical experiences. As it turns out, whenever the agent has a perceptual experience, a set of mental representations in his brain 'lights up.' And, as it turns out, whenever the agent is presented with a cow, the 'cow' mental representation is among those that lights up, and so on for other mental representations. Noticing these facts about the agent, an informational theorist might simply take R to be a causal relation: a mental representation has F as its content just in case that mental representation bears a simple causal relation to F.

But, as the example makes clear, talk of mental representations bearing causal relations to properties is really elliptical: it isn't the mental representation itself which bears the causal relations to the relevant properties, but rather occasions of the mental representation 'lighting up'. Causal relations between properties in the world and mental representations are defined in terms of causal relations between instances of those properties and events of the agent in question being in some mental state involving that mental representation.

I do not think that this point is controversial. But it gives us reason to ask: what are the relevant events of coming to be in a mental state, and how must they involve the mental representation in the content of which we are interested?[11] The following quote from Jerry Fodor illustrates a representative answer:

> Cows cause "cow" tokens, and (let's suppose) cats cause "cow" tokens. But 'cow" means *cow* and not *cat* or *cow or cat* because *there being cat-caused "cow" tokens depends on there being cow-caused "cow" tokens, but not the other way around.*[12]

[11] I owe the idea that the notion of tokening may cause special problems for the informational theorist to Mark Greenberg.

[12] Fodor (1990c), 91; emphasis his.

In this passage, occurrences of "cow" in quotes refer to a mental representation type — one which has the property of being a cow as its content. The theory is stated in terms of what causes tokens of this type, or, for short, what causes *tokenings* of mental representations. Tokening a mental representation is evidently a kind of mental state. One would like, however, to know a bit more about what this mental state is. Given the centrality of the notion of tokening a mental representation to informational theories of content, it is surprising how little has been said about what tokening a mental representation is. But it is, I think, fairly clear that there are two options for the informational theorist in giving sense to this notion.

The first is to make use of other internal states in defining the notion. We have already seen that we can make sense of the notion of a belief state in a fairly straightforward way: an agent is in a belief state with content p whenever the agent believes p.[13] This is not to give an account of what it is for an internal state to be a belief state; that, as we have seen, is no easy task. But it is enough to give us an idea of what we are talking about when we talk about belief states. If one takes belief states to be complexes with mental representations as constituents, then we can define the notion of an agent tokening a mental representation in terms of belief states in a fairly straightforward way: for an agent to token a mental representation is for that agent to be in a belief state which has that mental representation as one of its constituents.

The second option for the informational theorist is to take the mental state of tokening a mental representation to be *sui generis*, and indefinable in terms of the internal states which the private language theorist takes to underly thoughts, beliefs, or other propositional attitudes.

It seems to me that this second option is a non-starter. If we are to give a constitutive account of belief in terms of relations between things in the world and tokenings of mental representations, then it had better be a necessary truth that any agent capable of having beliefs also is capable of being in the mental state of tokening a mental representation. But, without any further information about tokening a mental representation, this seems like an assumption that we should reject.

To spell out this objection a bit more, it is worth exploring an kind of picture of tokenings which is suggested by the writings of informational theorists. Consider the phenomenon of, so to speak, ejaculating a word in response to some perceptual experience. Imagine, for example, a child, upon seeing a horse, yelling out "Horse!" Suppose further that this utterance is not an elliptical expression of a propositional attitude like thinking that there is a horse in front of oneself; rather, it is just a tokening of this word in response to perception of a horse. The intuitive idea is that tokening a mental representation is supposed to be a bit like this, except that it is an internal event which need not result in an utterance and, presumably, need not be knowable by introspection. On this view of tokening a mental representation, it is a substantive psychological claim that human beings token mental representations, and an outlandish claim (I suggest) that any possible agent capable of having beliefs would token mental representations in this sense.

This is not another way of trying to pump the intuition that agents could have beliefs without a certain kind of complexity in their inner representations. This objection allows that complex belief states related to each other and the world in certain very specific ways

[13] It's not quite this simple, since on the indication theory it is sufficient for the agent to be in a belief state with content q, where p is a necessary consequence of q. I ignore this complication here.

may be necessary to have beliefs; it just denies that, in addition to these belief states, one must perform these acts of tokening mental representations.

So it seems that the informational theorist should try to define tokening in terms of occurrence in the complex internal states underlying beliefs or other propositional attitudes. But this option faces a problem as well. Suppose we define tokening a mental representation in terms of thought-states, where these, by analogy with belief states, are the states which underly agents occurrently thinking p:

> A tokens a mental representation μ (at t) $\equiv \exists x$ (x is a thought-state of A (at t), and μ is a constituent of x)

We can now translate our schematic account of the form of an informational theory of content using the relation R between mental representations and features of the world as follows:

> μ has content F for $A \equiv$ events of μ being in thought-states of A bear R to F

R will be some relation between tokenings of the mental representation, in the above sense, and instantiations of the relevant properties.[14] As with any broadly causal theory, false thoughts will pose a problem — if I think that the cat is white when the cat is really brown, there may well be no instances of whiteness in the vicinity to stand in relation R to the state underlying my thought. But another problem arises for the informational theorist even if we abstract away from the possibility of error. We need to restrict the thought-states relevant to fixing the contents of the mental representations to those which not only have true contents, but also require for their truth the instantiation of all the properties which figure in their content.

This is not a trivial requirement. There are many propositions, one of whose constituents is a property F, which are such that the truth of that proposition does not require that F be instantiated. Indeed, some require that F *not* be instantiated. I may believe, for example, that dodos are extinct; presumably the informational theorist will account for this by my being in a belief state, one of whose constituents has as its content the property of dodo-hood. But obviously there is no reliable correlation, whether in ideal conditions or not, between my being in such a belief state and dodo-hood being instantiated. More generally, the problem is that the informational theory, in the above form, is an attempt to give an account of the content of a mental representation in terms of its occurrence in a thought-state; but because there is no guarantee that if a property occurs in a proposition, then the truth of the proposition entails that the property is instantiated, there is no guarantee that, even if we restrict ourselves to true beliefs, it follows that there is a reliable correlation between the presence of a mental representation in a thought-state and the instantiation of any property at all.[15]

To see why this is more than a technical problem for the informational theorist, consider how she might respond to the worry by modifying the schematic account given above. She might define some condition C such that

[14] I simplify by focusing on mental representations which are the analogues of predicates in natural language.

[15] This is a problem particular to causal theories which make use of mental representations; the indication theorist can say that, in the case of any true thought, the truth-making state of affairs is always actual. (Whether or not it can always be regarded as a cause or explanation of the relevant thought-state is a different question.)

> μ has content F for $A \equiv$ events of μ being in thought-states of A *which meet condition* C bear R to F

The aim in defining condition C will be to make it such that it is only met by thought-states whose contents are such that their truth requires that every property in the content be instantiated.

The real problem here is that the informational theorist cannot specify condition C — which is a property of thought-*states*, not of their contents — in terms of the contents of those states. This would contradict the order of explanation required by the informational theory. An informational theory is not a theory which takes the contents of certain mental states, such as belief states, as primary, and then goes on to abstract from these the contents of the mental representations which figure in these belief states. *This* sort of theory would be a variant on indication theories, and would face exactly the same problems with the fine-grained character of the objects of belief. Rather, one of the motivations behind informational theories is to take the contents of mental representations as prior to facts about the contents of belief states; by explaining content in this direction, the informational theorist then has a hope of accounting for the fact that the contents of belief states must be more fine-grained than sets of possible worlds. As a result, in specifying a condition C of thought-states which meets constraints (i) and (ii) above, *the informational theorist cannot help himself to facts about the contents of those mental states, if he is to use* C *in explaining what it is for a mental representation to have a given content.* And, as we'll see, this fact makes identifying such a class of mental states impossible.

It is no problem to restrict the class of *propositions* to those such that their truth entails that each property in the proposition is instantiated; this sentence is sufficient to define such a class. But the problem facing the information theorist is the problem of defining the class of *thought-states* which are such that they have propositions of this sort as their contents, and to do so without, on pain of circularity, assuming anything about the contents of those belief states. This means, in effect, that the informational theorist must find some purely syntactic property of thought-states which is a sufficient condition for such a state to have as its content a proposition such that its truth entails the instantiation of each of its constituent properties. Call this class of propositions *I-type* propositions. I think that a quick examination of some sentences which express I-type propositions alongside their non-I-type neighbors is enough to convince that there is no reason to believe that there need be any syntactic difference of the sort the informational theorist under consideration needs:

I-type	**Not I-type**
Dodos are plentiful	Dodos are extinct
John knows that Bob is bald	John believes that Bob is bald
Bob is bald and athletic	Bob is bald or athletic
There are two apples in the barrel	There are zero apples in the barrel
Harry is a bachelor	Harry was a bachelor

The foregoing argument shows that, in order to give even a rough criterion for agents tokening mental representations, the informational theorist must assume that there is some syntactic

difference between the way that I-type and non-I-type propositions are represented by belief states.[16] Moreover, since we are interested in answering the question, "What is it for an agent to believe p?", and not in giving a contingent explanation of part of the human cognitive system, our informational theorist is committed to a syntactic difference of this sort being a metaphysically necessary condition on an agent having any beliefs at all. But this is surely a mistake.

To sum up: the argument presents a dilemma. On the one hand, the informational theorist may not take tokening a mental representation to be defined in terms of states underlying certain propositional attitudes; in this case, it is implausible to think that tokening a mental representation should be a necessary condition for having beliefs. On the other hand, the informational theorist may try to define tokening a mental representation in terms of the occurrence of mental representations in states underlying certain propositional attitudes. But then it is implausible to think that the existence of a syntactic distinction in one's inner language between states which have I-type propositions as their contents and those which do not is a necessary condition on having beliefs.[17]

I think that this problem about the notion of tokening a mental representation is a fundamental one which shows that informational theories of content are misconceived. But, once again, one might take this argument to rest on questionable modal intuitions: in this case, the intuition that it is not a necessary truth that any agent capable of having beliefs also be such as to be in the *sui generis* state of tokening mental representations in the right way and at the right times. So it will be useful to show that, even setting this problem about tokening mental representations aside, no way of working out the informational theory of content can be correct.[18]

The basic question for an informational theorist to answer is: given that the contents of mental representations are fixed by relations between events of tokening those representations and instantiations of properties in the world, what, exactly, are the relevant relations? Here a number of different proposals have been made. None of them, I will argue, are successful.

5.3.2 Information as causation

As with the indication theory, the natural initial thought is that the contents of mental representations are the things which cause them to be tokened. But, as was the case with the simple causal version of indication theory, this falls immediately to the problem of making room for false mental states: in this case, false tokenings of mental representations. Whatever tokening a mental representation is, it should be possible to have false tokenings: cases in which, to use the hackneyed example, an agent tokens a mental representation which has the

[16]It is important to be clear about what condition C must do. There is no requirement that the condition give an exhaustive distinction between thought-states which do and do not have I-type propositions as their content; it is enough that the condition state a sufficient condition for a state having an I-type propositions as its content. One also assumes that the condition must be weak enough to let in a substantial number of such states.

[17]I owe to Mark Greenberg's seminar in the Fall of 2000 the idea which underlies this dilemma: namely, that there are cases other than cases of error in which the covariation required by the informational theorist will be absent.

[18]In what follows, I shall use the term 'tokening' to state the theories in question, and ignore the problems which, as we've seen, attend this notion.

content *horse* in response to seeing a cow on a dark night. But an informational theory like the present one which identifies information with causation yields the result that this sort of case is not possible, since the mental representation in question would the have the disjunctive content (say) *horse or cow on a dark night.*

5.3.3 Information as counterfactual dependence

The natural next move to make is the information-theoretic analogue to Stalnaker's appeal to optimal conditions. Given some non-circular specification of optimal conditions, we can say that the content of a mental representation is whatever, under optimal conditions, would have caused it to be tokened (in the relevant agent). So, along these lines, one might identify information with a certain kind of counterfactual dependence, as follows:

> μ has content F for $A \equiv ((A$ is in optimal conditions & A tokens $\mu)$ $\Box\!\!\rightarrow A$ tokens μ because F is instantiated)

But this meets the same problem as did the indication theorist's appeal to optimal conditions: the conjunction problem. By reasoning parallel to that pursued at length in "The conjunction problem" above,[19] it follows from the fact that a property F is the cause of a tokening of μ under optimal conditions that the conjunctive property composed of F and the property of being in optimal conditions is also a cause of the tokening of this mental representation. But then we get a result parallel to that reached above: for any property F, whenever an agent believes that x is F, that agent also believes that that x is F and in optimal conditions. But this is clearly false.

5.3.4 Information as covariation

But, it might seem, the informational theorist has a response to the conjunction problem which was not available to the indication theorist: he might appeal not to dependence of a tokening of a mental representation on a property under optimal conditions, but rather to *covariation* of tokenings of a mental representations with a property. That is,

> μ has content F for $A \equiv \Box(A$ is in optimal conditions $\rightarrow$ (A tokens $\mu \equiv F$ is instantiated))

This is not a move open to the indication theorist, who endorses mentalism rather than the private language picture. For if this idea of covariation under optimal conditions were put to use by an indication theorist, the resulting theory would say that an internal state has content p just in case, were the relevant agent in optimal conditions, she would be in that state iff p was the case. But this would change the nature of optimal conditions: they would, on such a theory, have to be conditions under which agents are not only infallible, but also *omniscient.*[20] And to hope for a non-intentional specification of optimal conditions in this

[19] §4.4.2, pp. 70 ff. There I discuss the reasoning behind the following argument, and a number of possible responses, at greater length.

[20] If the agent were not omniscient under optimal conditions, then there would be some such conditions in which the agent did not believe p even though p was the case. But such cases (according to the indication-theorist's view of belief) would then be cases in which the agent failed to be in some belief state x (which actually has the content p) even though p was the case. But such cases would be enough to disqualify x from having the content p.

sense is to hope for too much.

The appeal to covariation under optimal conditions does solve the conjunction problem; sometimes the agent will be in optimal conditions without tokening μ, and this will be enough to prevent this mental representation from having the property of being in optimal conditions as part of its content. This is all to the good.

But this solution to the conjunction problem raises another problem, to which there seems to be no non-circular solution. According to the view of information as counterfactual dependence, the content of a mental representation was fixed by its cause in the nearest possible world in which the relevant agent is in optimal conditions. The salient point is that the class of worlds relevant to content determination will be delimited by the constraint that they be the worlds most similar to the actual world in which the agent is in optimal conditions, and tokens the mental representation in question.

But in the case of the present theory, there is no such constraint. As it is stated, the content of the mental representation μ is fixed by covariance with properties *in every possible world in which the agent is in optimal conditions*. But this will not do. This class of possible worlds is, I presume, very large, and there will be many differences with respect to the cognitive systems of agents between some worlds in which they are in optimal conditions and other worlds in which they are in optimal conditions. In particular, it seems clear that some of their mental representations will have one content in some of these optimal conditions worlds, and another content in other such worlds. But this possibility is already enough to entail the falsity of the formula stated above.

The natural response is to find some way of delimiting the class of possible worlds relevant to the covariational account. That is, one wants to find some condition C on worlds such that only worlds in which the agent is in optimal conditions and which meet condition C can figure in the determination of the content of the mental representation. But this faces two problems: (i) It is hard to see what C might be, other than the circular 'a world in which μ has the same content as it actually has.' (ii) This lets the conjunction problem back in. By reasoning parallel to the above, it will follow that every mental representation will have a conjunctive content, part of which will be the property of being in a world in which C is satisfied.

5.3.5 Information as asymmetric dependence

The basic idea behind the counterfactual dependence and covariational versions of the informational theory of content was to respond to the problem of error by shifting the focus of the theory from actual causal relations to certain counterfactual causal relations. An alternative strategy has been proposed by Jerry Fodor: the contents of mental representations are fixed by certain special laws connecting mental representations and their contents. Because the laws can hold without the properties being instantiated, this too shifts the focus away from actual causal relations; but because the laws are not analyzed in terms of counterfactuals, one might reasonably think that Fodor's version of the informational theory can avoid the problems we have discussed so far. Beginning with his *Psychosemantics*, Fodor has pursued this strategy to arrive at a version of the informational theory which does not rely on finding conditions under which a certain sort of mental state always has as its content a true proposition, and so makes no use of counterfactuals involving optimal conditions.

Given the question in which we're interested — What is it for an agent to believe p? — we

shall have to rely on a version of Fodor's theory that he would not endorse. As I noted above,[21] Fodor is interested in providing merely sufficient conditions for a mental representation having a given content, rather than necessary and sufficient conditions. But, as argued above in §1.2, this restriction is unmotivated. Either one is interested in saying what it is for an agent to believe p, in which case necessary and sufficient conditions seem to be called for, or one is interested in showing that the existence of beliefs is consistent with a version of physicalism stated in terms of supervenience. If the latter, then provision of a complex supervenience base for beliefs seems to be of little interest, since it is unlikely to carry more conviction than the simple thought-experiment that no two worlds alike with respect to all of their physical properties differ with respect to the beliefs of the agents in those worlds. In any case, the important point at present is that, if Fodor's version of the informational theory is to answer the question to which this essay is devoted, we shall have to revise it to be a statement of necessary and sufficient conditions for a mental representation having a given content.[22] It should be clear from what follows, though, that it fails to provide either a necessary or a sufficient condition.

Fodor's theory is based on there being an asymmetric dependence of some laws connecting properties and tokenings of mental representations on others; considered as a constitutive theory of the contents of mental representations, it may be summarized as follows:

μ has content $F \equiv$ (1) it is a law that instantiations of F cause tokenings of μ, &
(2) if it is a law that instantiations of G cause tokenings of μ & $F \neq G$, then the G-law asymmetrically depends on the F-law[23]

[21] See p. 7, note 11.

[22] In one sense, there is nothing objectionable about the restriction to sufficient conditions. Fodor seems to think that the sort of account he gives will be well suited for mental representations which have properties as their semantic content, but not for logical vocabulary or proper names. There is, of course, nothing wrong with a piecemeal approach to the contents of mental representations; the objection in the text is to the restriction to sufficient conditions for even the favored class of mental representations. All the examples I use will be mental representations which play the role of predicates in the language of thought.

[23] Two quick points of clarification. First, the laws in question may be (and always are) *ceteris paribus* laws: laws which hold, all things being equal. Second, laws connecting properties may be true at a world even if one of the properties is not instantiated at that world; this makes room for mental representations which have uninstantiated properties as their content. But note that laws connecting properties cannot be true if one of the properties is *necessarily* not instantiated. So, as Fodor notes ((1990c), 101), it follows from his theory that there can be no semantically simple representations in the language of thought that have as their content necessarily uninstantiated properties. But, given that we have as yet no way to see which mental representations are semantically simple, this doesn't give us much of a test of Fodor's theory.

Note also that, for the theory to have any plausibility at all, values of 'μ' must be restricted to mental representations. Presumably there are very many laws connecting various sorts of properties, and many of them asymmetrically supervene on others. If any value of 'μ' were permissible, then no doubt we would end up with rocks and telephone poles having content. The restriction to mental representations is enough to avoid this sort of "pansemanticism"; but one might think that it is a rather *ad hoc* measure. If asymmetric dependence really were the basis of intentionality, then why

Understanding this account is complicated by the fact that Fodor gives two nonequivalent glosses on the asymmetric dependence:

[AS_1] Law L_2 asymmetrically depends on law $L_1 \equiv_{df}$ there is a possible world in which L_1 is a law but L_2 is not which is more similar to the actual world than any world in which L_2 is a law but L_1 is not.

[AS_2] Law L_2 asymmetrically depends on law $L_1 \equiv_{df}$ ((L_1 is not a law $\Box\rightarrow L_2$ is not a law) & $\neg$(L_2 is not a law $\Box\rightarrow L_1$ is not a law))

Fodor seems to switch between these characterizations throughout his work.[24] Of these two interpretations of asymmetric dependence, the latter is the stronger: if one law asymmetrically depends on another according to [AS_2], then it also asymmetrically depends on it according to [AS_1]; but the converse entailment does not hold.[25]

On neither interpretation, though, does Fodor's theory provide necessary or sufficient conditions for a mental representation's having a given content. In what follows I shall first develop a simple example which illustrates this point, and then go on to argue that Fodor's theory does not really address the simple cases of error which provide the motivation both for the appeal to optimal conditions and for Fodor's theory.[26]

First, the example. Suppose that an agent A thinks that a cow is in front of him whenever a cow is before him. Whenever he has this thought, by hypothesis, he is in a thinking-state which has the content

<*IN-FRONT-OF*<COW,*A*>>.

Suppose, for simplicity, that whenever he has this thought he is in the same thinking-state, and that this thinking-state consists of three mental representations, combined in a certain way. Suppose further that one of these representations, μ, is never a constituent of one of A's thinking-states unless a cow is in his presence, and that analogous facts hold for the other

shouldn't objects figuring in laws which exhibit the right pattern of asymmetric dependence have content? I set this aside in what follows.

[24]For a statement of asymmetric dependence in accord with [AS_1], see the first full paragraph on p. 113 of (1990c); for a clear statement of [AS_2], see his (1987), pp. 108-9. Sterelny ((1990), 120) takes Fodor to be committed to the claim that if one law asymmetrically depends on another, then there is no possible world in which the former is a law and the latter is not; but the text does not seem to me to support this interpretation.

[25]*AS_2 dependence entails AS_1 dependence*: Suppose that [AS_1] does not hold between L_1 and L_2. Then either (i) the nearest world in which L_1 fails to hold is more similar to the actual world than the nearest world in which L_2 fails to hold, or (ii) the nearest world in which L_1 fails to hold is also the nearest world in which L_2 fails to hold. If (i), then the first conjunct of the right-hand side of [AS_2] is false; if (ii), then the second conjunct is false. (For simplicity I ignore the possibility of distinct worlds equally similar to the actual world. Even if there are such worlds, this entailment holds.)

AS_1 dependence does not entail AS_2 dependence: Suppose that the nearest world w_1 in which L_2 fails to hold is one in which L_1 does hold; that the nearest world w_2 in which L_1 fails to hold is one in which L_2 does hold; and that w_1 is more similar to the actual world than w_2. Then [AS_1] dependence holds between L_1 and L_2, but [AS_2] dependence does not.

[26]Mark Greenberg has noted the possibility of counterexamples of roughly the sort I give. Much of this section is owed to discussions with Mark about Fodor's views.

two representations and their contents. We can suppose, then, that there is a law connecting COW instantiations to tokenings of μ. I presume that any proponent of an informational semantics will want to count μ as having as its content the property COW. The case is, after all, designed to be as unproblematic as possible; we are supposing the agent in question to be infallible, and focusing on a property instantiated by middle-sized material objects of the sort which stand in lawlike relations to mental representations, if anything does.

Now, while Fodor never gives explicit conditions for a law holding, it seems clear that these conditions will have to be fairly relaxed if there are to be any laws at all connecting instantiations of properties to tokenings of mental representations; after all, mistakes are not uncommon things. Recall also that, however we are ultimately to understand tokening a mental representation, to do so is just for one's brain to be in a certain state. Presumably, having one's brain be in a particular state is, in important respects, similar to having one's heart be in a particular state. More specifically, it seems that, just as taking certain kinds of drugs can lower heart pressure, taking certain kinds of drugs can alter one's brain states. But, if this is true, and tokening a mental representation is just another kind of brain state, then it seems very likely that there is some drug D — which may, of course, be merely possible — which would cause A to be in a state of tokening μ.

How plausible is it to presume that there is a law connecting some possible drug D and A's tokenings of μ? Consider how implausible the negation of this claim sounds: there is no metaphysically possible substance such that, were you to ingest that substance, you would be in a state of tokening a given mental representation. Many actual drugs stimulate different brain states; I presume that hallucinogens, in particular, cause tokenings of mental representations. And there are, I presume, infinitely many possible hallucinogens. Note further that nothing in what follows requires D to be a drug which is ingested; for the purposes of the argument, any way of causing a tokening of a mental representation unrelated to its content will do, whether it involves drugs or the manipulations of brains using as yet unimagined neurological techniques.

So we have two laws connected to the tokening of μ by A: one involving COW, and the other D. Is there reason to believe that the nearest world in which the former law is broken will be more similar to the actual world than the nearest world in which the latter is broken? The opposite seems just as likely: perhaps breaking the D-to-μ law would require changing fundamental facts about the chemistry of the brain. In this case, the [AS_1] interpretation of Fodor's conditions yields the incorrect result that μ does not have COW as its content; hence these conditions are not necessary. It does entail the result that μ has D as its content; since this too is incorrect, the conditions are not sufficient either.

Now consider this case in light of the [AS_2] interpretation of asymmetric dependence. The foregoing is enough to show that these conditions are not necessary; μ has COW as its content, despite the fact that it is not the case that, were the COW-to-μ law broken, the D-to-μ law would be as well. To show that these conditions are not sufficient either, one further assumption is required: were the D-to-μ law broken, the COW-to-μ law would be as well. And this is, I think, a plausible assumption: given that breaking the D-to-μ law would require changing fairly fundamental facts about A's brain, it is no surprise that changing such facts should also break the connections between μ — whose identity across possible worlds, to reiterate, cannot on pain of circularity be fixed by its content — and instantiations of COW.

So these conditions are not sufficient either.[27]

These cases show that, on either interpretation, Fodor's conditions are neither necessary nor sufficient, even if we idealize away from mistakes and focus on the sorts of properties which should be easiest for causal accounts. But one can also show that Fodor's account fails to give a satisfactory treatment of the simple cases of error which are supposed to motivate the theory.

Vary our example so that our agent A tokens μ whenever a cow or a dappled horse is in front of him. This is the sort of prosaic example of error for which Fodor's theory was designed to account. In this case we have three laws in place, each of which connects one of the properties COW, DAPPLED HORSE, or COW OR DAPPLED HORSE to μ. Fodor's idea is that the latter two asymmetrically supervene on the first. If one broke the law connecting COW instantiations to to tokenings of μ, one would thereby break the laws connecting these tokenings to our other two properties; after all, dappled horses cause tokenings of μ only because A mistakes them for cows!

But, though this line of thought may sound plausible at first, there is really no reason to believe that the counterfactuals in question should come out this way. Consider the first two of our three laws:[28]

L_1 Instantiations of COW cause tokenings of μ.

L_2 Instantiations of DAPPLED HORSE cause tokenings of μ.

When we imagine a world in which one of these laws fails to hold, what change to the actual

[27]Much the same problems arise from weakening slightly the assumptions of the example. Suppose that the D-to-μ law and the COW-to-μ law are in place by virtue of very different underlying mechanisms; it might then be the case then that neither law asymmetrically depends on the other. In particular, it might be the case that in the nearest world in which the former law is broken the latter law remains in place, and vice versa. What, then, should Fodor say about the content of μ?

His theory seems to entail that, in such a case, μ has no content at all, on the grounds that there is no law connecting instantiations of a property to tokenings of μ which is such that all other such laws asymmetrically depend on it. But, on second glance, there is such a law: the law that connects tokenings of the disjunctive property COW OR D to tokenings of μ. If the situation is as described above, then both of the laws with which we began asymmetrically depend upon this one; and, if every law other than these two asymmetrically depends on one of them, then every such law asymmetrically depends upon this new, disjunctive law. So it seems that Fodor must say that μ has the disjunctive content COW OR D, which it doesn't; so this gives us another counterexample to the sufficiency of Fodor's conditions.

One might reply by adding a proviso ruling out disjunctive laws; but this would not be much help, for two reasons: (i) Presumably, if there are such things as mental representations, and some of them have contents, some of them have as their content a disjunctive property; and this move would show that the contents of these representations are not explicable by Fodor's theory. (ii) This move would yield the result that, in this example, μ has no content; and this is hardly more plausible than its having a disjunctive content.

Here and in what follows, I use the [AS_2] formulation of asymmetric dependence; this seems the most charitable interpretation of Fodor's theory.

[28]For simplicity, I ignore the need to relativize these laws to an agent.

world are we imagining? Fodor's thought seems to be the following.[29] There are certain, no doubt complex, mechanisms of A's brain which underly L_1: a complex set of links between cows impinging on A's nervous system in various ways and tokenings of μ. Just so, there are certain, no doubt complex, mechanisms of A's brain which underly L_2: a complex set of links between dappled horses impinging on A's nervous system in various ways and tokenings of μ. Suppose that mechanism M_1 underlies L_1, and that M_2 underlies L_2. To imagine a world in which L_1 does not hold is to imagine the world most similar to the actual world in which mechanism M_1 is disrupted, and analogously for L_2 and M_2.

The key question, which Mark Greenberg has pressed, is: how are M_1 and M_2 related? One possibility is that they are precisely the same mechanism. It might be that A's visual system represents cows and dappled horses in similar enough ways that the same mechanisms for tokening mental representations come into play. But in this case, neither law will asymmetrically depend on the other: to break M_1 is to break M_2, hence to break L_1 is to break L_2, and vice versa. So in this case, since there is *no* law connecting instantiations of a property to tokenings of a mental representation such that all other such laws asymmetrically depend on it, Fodor's theory counts μ — incorrectly, I assume —as having no content.[30]

On the other hand, M_1 and M_2 might be completely distinct mechanisms. In this case, the result is just as bad for Fodor. Again, neither law asymmetrically depends on the other; if there is no necessary connection between M_1 and M_2, then we can break either of L_1 and L_2 without breaking the other. The difference between this case and that of the preceding paragraph is that, in this case, both L_1 and L_2 asymmetrically depend on a third law: the disjunctive law connecting instantiations of the disjunctive property COW OR DAPPLED HORSE with tokenings of μ. So Fodor's conditions assign this disjunctive property as the content of μ; and this, I take it, is as bad as saying that it has no content at all.

Indeed, the only way that Fodor's cases *could* work for these simple cases of error is if there is a partial overlap between M_1 and M_2 of a very specific kind: if they share just enough in common that breaking M_1 would break M_2, but not the converse. But there is just no reason to think that the facts which underly laws connecting properties to tokenings of mental representations should typically be like this; and it is clear from the foregoing that it is neither necessary nor sufficient for a mental representation having a certain content that they should be.[31]

[29] See, e.g., Fodor (1987), 107-8.

[30] One might reply on Fodor's behalf as follows: we should imagine L_1 and L_2 broken, not by disrupting the mechanisms which underly them, but by changing the appearances of cows and/or dappled horses so that they are more clearly distinguishable by A. But this is no help. In a possible world in which cows look like dogs and dappled horses look just as they do, L_2 will presumably hold, though L_1 will not. Hence there is no asymmetric dependence. (One could retreat to the weaker interpretation [AS_1]; but there seems no reason why the intuitively 'right' world will always be the more similar to the actual world.)

[31] In developing these objections to the asymmetric dependency theory of content, I have relied on the theory of content advanced by Fodor in the first half of his (1990c). He has considered modifying the account by adding either or both of the following conditions, which concern laws instantiated at the actual world rather than laws which merely hold at the actual world:

(1) If μ has content F, then the μ-to-F law must be instantiated.

(2) The only laws required to asymmetrically depend on the content-constituting law are those which are instantiated.

5.3.6 Information as teleology

These three responses to the failure of the simple causal informational theory — the views of information as counterfactual dependence under optimal conditions, of information as covariance under optimal conditions, and of information as asymmetric dependence — all have something in common. They all respond to the problem posed by the possibility of false beliefs by de-emphasizing the importance of actual causal relations. The first two responses do this by appealing to non-actual worlds, while the third does it by appealing to laws which make only minimal demands on actual happenings. A last option for the informational theory of content breaks with this strategy: a *teleological* theory of the contents of mental representations tries to make room for error by focusing on a broader class of actual causal relations than do most informational theories.

On teleological theories of content, the content of an internal state is determined by the function of the mechanism which produces that state: on one rough formulation, if an agent has a mechanism for coming to be in an internal state x, and if the function of that mechanism is to produce x when p, then x has p as its content. The notion of a *function* is to be understood in terms of natural selection: the function of a mechanism of an agent A is to produce x when p iff in A and A's ancestors, the actions caused by being in state x normally require for their success that p obtain.[32] The notion of *success* is to be understood in terms of adaptive advantage. In this sense, teleological theories appeal, not to what internal states agents would be in under certain circumstances, but to the actual history of the mechanism which produces that internal state.

(Fodor considers including these in his (1990c), 121, but seems to reject the idea in Appendix B (116 ff.) of his (1994).) The point of (1) is to account for the contents of natural kind terms, which seem to be fixed by the actual environment. This would not affect any of the cases discussed above. (2), however, would seem to be a way to answer the objection involving our imagined drug D. But this is only a superficial solution. Consider a possible world in which D is actual, and an agent once tokened μ in response to D. Would this change the contents of the thoughts had by the agent involving μ? It seems not. Neither (1) nor (2) go any ways toward solving the case of error discussed above; and, moreover, both create serious problems for mental representations which have uninstantiated properties as their content.

[32]This is, following Millikan (1989), a "consumer-based" account of function: the content of a belief state is p iff p is required for the actions of the consumer — in this case, the agent who is in the belief state — to satisfy his needs. The same sort of account is defended by Papineau (1987) (66-67) and Whyte (1990); the main difference between the views of these two and the view of Millikan is that, according to the former two, a belief state has content p iff p is *sufficient*, rather than necessary, for the success of actions based on this belief state. The objections to Millikan's account discussed below apply also to these variants on the consumer-based teleological approach.

This consumer-based approach should be distinguished from the teleological theory defended by someone like Dretske (see his (1988), (1990)). Dretske combines an indication theory of meaning with the teleological theory, so that an internal state has content p iff that state has the function of indicating p. For the purposes of the objections above against indication theories, Dretske's version of the teleological theory may be considered a version of the indication theory.

For purposes of exposition, I gloss over many complications in Millikan's account. She notes, for example, the objection that many standard background conditions — e.g., the presence of oxygen or gravitational forces — are necessary conditions for the success of many actions, but usually are not part of the contents of the internal states which cause these actions. The simpler version in the text is enough, I think, to see the intuitive problems with the approach.

The standard test cases for teleological theories of content often involve creatures like frogs and bacteria; there is some reason to doubt whether any theory of content tied so closely to biological needs could serve as an account of the contents of the beliefs of adult persons. Though I share this doubt, in what follows I shall restrict myself to a very simple case of (fictional) animal behavior, for two reasons: this will enable us to focus on the simplest version of teleological theories of content, allowing us to ignore the accounts of learning to which they have been attached; and, if the theory fails even for primitive cases of representation, this will rule out not only a teleological theory which attempts to account for all beliefs, but also a more modest theory which attempts to use teleology to account for the contents of the simplest representational states and build from these, by non-teleological means, to more complex representational states.[33]

On the sort of teleological theory we're considering, then, the explanation of the content of a belief state will involve the conditions under which actions based on that belief state are successful. The fundamental problem with this idea is, I think, that it misconstrues the relationship between the contents of mental states and the success of actions based on being in those states. Roughly, the success of an action based on an agent being in a given mental state x may sometimes be explained by the world being as x represents it to be; teleological theories, however, reverse this order of explanation, so that the content of a mental state — the way that state represents the world as being — is explained by the success of actions based on being in that state.

That this reversal of explanatory priorities is a mistake can be shown by consideration of cases in which the truth of an agent's belief is reliably, but only accidentally, correlated with the adaptive success of actions based on that belief. In such cases, we are inclined to count the cause of the relevant belief state as its content, and not that state of affairs — reliably but accidentally correlated with the cause — which explains the adaptive success of actions based on that belief state. A good example of this kind has been suggested by Paul Pietroski:

> The kimus live near a large rocky hill. Their only predators are snorfs, carnivores who roam past the hill each morning. Kimus used to be "color-blind." But in virtue of a genetic mutation, one particular kimu — call him Jack — came to have an internal mechanism M that produced tokens of a physically specifiable state type B in the presence of certain wavelengths of light. Each morning, something red on the hilltop caused Jack to form a B-token when he looked up. And Jack (like his descendants) turned out to have "fondness" for red things; i.e., other things being equal, Jack would move toward the distal causes of B-tokens when such tokens were produced. So each morning, Jack trudged up the hill and thereby avoided the snorfs. Natural selection took over; and Jack's mechanism type proliferated throughout the species. There was no other reason (e.g., detection of food) for the selection in favor of having the "color mechanism."[34]

On a teleological theory of content, the content of a token of B will be something along the lines of "snorf-free territory that way;" after all, the fact that the direction in which Jack and his descendants walk is snorf-free is the explanation for the adaptive success of actions based on tokens of B. But this is very implausible. We can suppose that snorfs are quite

[33]This more modest version of a teleological theory is endorsed by Sterelny (1990), §6.6.

[34]Pietroski (1992), 273. A similar example is suggested in Stich (1990).

stealthy and speedy predators; perhaps they nab the dumb and slow-moving kimus so quickly on their morning trek that no kimu has ever seen a snorf. In this case, we should be inclined to say that kimus have no beliefs at all about snorfs, let alone that their tendency to gravitate toward red things is in fact such a belief.[35]

The problem stems from the fact that the contents of beliefs in these cases seem to track the explanation of the agent's coming to be in the relevant belief state, and *not* the explanation of the success of the actions caused by this belief state. This is a strong indication that a causal theory of content should focus on the causes of an agent's coming to be in a given mental state, rather than on explanations of adaptive behavior; hence it is a strong indication that teleological theories, like the other informational theories we have considered, are on the wrong track.

5.3.7 Informational theories and belief states

All these problems, moreover, have come in the discussion of the first half of the task for the informational theorist: the task of saying what it is for a mental representation to have a given content. We have not yet asked how the informational theorist answers the question of what it is which makes certain sets of mental representations, but not others, qualify as belief states.

Because informational theories of content are usually pursued as 'theories of content' in isolation from the task of saying what it is for an agent to bear a certain relation, like belief or intention, to a content, little is forthcoming from informational theorists on this score. The most plausible form for such a theory to take would be that given in the pragmatic half of Stalnaker's causal-pragmatic theory; but we have already seen that this account of belief states is unsatisfactory,[36] and there is, for now, no reason to expect that the informational theorist can do better.

5.4 Mental representations and conceptual role

In a sense, conceptual role theories of belief stand to functional role theories as informational theories stand to indication theories. Just as informational and indication theories focus solely on mind-world relations for the fixation of content, so conceptual role and functional role theories focus on both mind-world and intra-mental relations. And just as informational theories are the private language theorist's version of indication theories, conceptual role theories are the private language theorist's version of functional role theories.[37]

But conceptual role theories are more intimately related to functional role theories than are informational theories to indication theories, as can be seen from the discussion of the notion of 'tokening a mental representation' above. Mental representations are constituents of belief states; but, according to the informational theorist, the informational content of a mental representation is not derived from what belief states indicate. Rather, the contents

[35] As Pietroski suggests (276), this intuitive judgement could be further supported by behavioral tests; we might, e.g., place a kimu in front of a snorf and notice that the kimu does nothing at all, or place the kimu in front of a red light and observe the kimu head toward the light.

[36] See above, §4.4.5, pp. 80 ff.

[37] See the diagram representing these four versions of functionalism on p. 65 above.

of mental representations are, for the informational theorist, fixed independently by events of tokening mental representations. Informational content is supposed to be prior to indication.

But the conceptual role theorist takes the contents of mental representations to be fixed at least partly by their role in reasoning. For this reason, she must take the conceptual role of a mental representation to be derivative from the functional roles of states in which that representation figures; there is, after all, no such thing as inferring one mental representation from another, just as there is no such thing as inferring 'dog' from 'cat.' Because the objects of inferential transitions seem always to be whole propositions, the notion of a functional role must be the more fundamental one.

Inasmuch as the notion of tokening a mental representation caused more than its share of problems, this is a point in favor of conceptual role theories of content, as is the fact that the conceptual role theorist has both internal and external relations between mental representations among her resources for the formulation of a theory of belief. But, as we shall see, the kinship between conceptual role and functional role theories is not without its problems for the conceptual role theorist.

5.4.1 Conceptual role semantics & conceptual role theories of content

This motivates the move toward what is commonly known as "conceptual role semantics". On the face of it, this is a strange use of the term "semantics". After all, as discussed above,[38] there is a distinction between semantic theories, which state axioms which entail pairings of expressions of a language with their meanings, and foundational accounts of content, which say what it is, or what constitutes, an expression having such a meaning. It seems, on the face of it, that "conceptual role semantics" provides an answer to the latter question: it says that mental representations have their content in virtue of having a certain conceptual role.

The word "semantics" is often used when what is really being put forth is a kind of foundational theory of content (e.g., informational semantics, indication semantics, teleological semantics ...). Usually, this use of the term is harmless. But in the case of conceptual role semantics it conceals the distinction between two very different views.[39]

Sometimes by this term, philosophers do intend to be setting forth a properly semantic theory: on this sort of theory, the meaning of a term *is* its conceptual role. (By contrast, no proponent of 'teleological semantics' thinks that the content of a term *is* its evolutionary function in a species.) Other times, conceptual roles are thought of as the facts in virtue of which mental representations have their contents; these contents might be, say Fregean senses, but the mental representations come to have these senses as their contents by having a conceptual role which, according to the theory, is matched with that sense.

Since we are engaged in trying to answer foundational questions about content, it might seem that we should only be interested in the second interpretation of "conceptual role semantics." But this is not quite right. After all, if the contents of expressions were just conceptual roles, it would be hard to see how the facts in virtue of which those expressions come to have those contents could be anything other than, well, their having those conceptual roles. So if conceptual role semantics were to provide a satisfactory semantics for mental representations,

[38] See §1.2.

[39] Here and in what follows I am indebted to Mark Greenberg.

this would give us a very good reason to adopt a conceptual role theory of the facts in virtue of which expressions have the contents that they do.

But it does not seem to me that conceptual role semantics, construed as a semantic theory, is very plausible. The idea that meanings *are* conceptual roles — perhaps construed as sets of inferential transitions — rather than Fregean senses or properties — seems like a kind of category mistake. One way to bring this out is to note that, standardly, meaning has been thought to determine reference; but it is hard to see how something like a conceptual role (as opposed to a Fregean sense determined by a conceptual role) could determine the reference of a mental representation.

The conceptual role theorist might reply by simply abandoning the thesis that content determines reference, and adopting a kind of disjunctive theory which identifies content with conceptual role but lets reference be fixed by causal relations between mental representations and their referents. But to make this move is to appeal to some version of an informational theory of reference for mental representations; and we have already seen in the preceding section that such theories are unsatisfactory.[40]

So in what follows I shall focus on versions of "conceptual role semantics" which take the conceptual roles of mental representations to be the facts in virtue of which they have a certain content, rather than the contents themselves.

5.4.2 The relationship between conceptual role and functional role

As noted above, the conceptual role of a mental representation is derived from the functional roles of the mental states of which it is a constituent; talk about the causal role, or conceptual role, of a mental representation must be construed as talk about what can be abstracted from the functional roles of the class of mental states of which the mental representation in question is a constituent. As noted above, this means that conceptual role theories, unlike informational theories, do not need to define a special notion of 'tokening a mental representation.' But this also means that conceptual role theories inherit all of the problems of functional role theories.

Above in "Content and functional role,"[41] I argued that functional role theories faced unanswerable problems on two scores. First, there is no satisfactory way of saying which causal relations (actual or possible) are relevant to the determination of the functional role of a belief state. Second, neither of two possible versions of functional role theories — commonsense functional role theories and psychofunctionalist theories — have the resources to deliver an account of the contents of belief states. In the case of commonsense functionalism, this was because ordinary platitudes about belief simply do not encode enough information for us to

[40] Another option is to move to the kinds of "two-factor" theories discussed in, e.g., Field (1977), Loar (1982), and Block (1986). On such theories, the conceptual role of a mental representation provides a function from informational links to contents. These theories are not, however, best understood as theories on which the contents of mental representations are identified with their conceptual roles. Rather, they can be considered equivalent to a kind of nonsolipsistic conceptual role semantics according to which both internal links between mental representations and external links between mental representations and the world have a role to play in determining the content and reference of a mental representation. Thanks to Daniel Rothschild for pointing out the difference between two-factor theories and the kind of disjunctive theory discussed in the main text.

[41] §4.5, pp. 83 ff.

extract a theory of content; in the case of psychofunctionalism, this was because it is obscure how empirical psychology could deliver the connections between beliefs (as opposed to belief states) needed for an assignment of contents to belief states.

But, since conceptual roles are just abstractions from functional roles, these results carry over to versions of conceptual role theories of content which take as their starting point either commonsense talk about belief or the results of empirical psychology.

5.4.3 Possession conditions & conceptual roles

But the conceptual role theorist, unlike the functional role theorist, has a way of avoiding the dilemma of relying either on commonsense platitudes about belief or upon the findings of empirical psychology. For the conceptual role theorist may, so to speak, proceed "concept by concept." That is, the conceptual role theorist may claim that the conditions which a mental representation must satisfy in order to have the content F are given neither by commonsense generalizations nor by psychology, but rather by a priori philosophical reflection on the concept F. This was not a move open to the functional role theorist, since the idea of proceeding "proposition by proposition" is not an attractive one.

This version of conceptual role theory has been defended at length and with great ingenuity by Christopher Peacocke.[42] However, it takes a bit of explanation to see how Peacocke's approach to mental content is related to the foundational questions about belief we have been considering. Peacocke's central focus is on providing *possession conditions* for concepts. But it is not immediately obvious how stating the conditions an agent must satisfy in order to possess a given concept might yield a constitutive account of what it is for an agent to have a belief whose content involves that concept.

Roughly, the account works as follows.[43] To say that an agent possesses some concept C is to say that the agent has a certain dispositional property: (i) that he is disposed to find certain transitions between contents compelling, (ii) that he finds them primitively compelling, rather than compelling on the basis of other premises or principles, and (iii) that he is so disposed in virtue of the form of those transitions.[44] Which transitions are relevant will depend upon the concept for which possession conditions are being given; the key here is that possession conditions are given in terms of dispositions to make certain *transitions*.

In general, given any proposition p in which a concept C occurs, there will be some transitions either from or to another proposition (or perceptual content) q which are included in the statement of the possession conditions for C. Call the set of these transitions T. We can then say what it is for an agent to be in a belief state the content of which involves C:

> An agent is in a belief state x the content of which includes $C \equiv$ the agent is disposed to make transitions T to and from x (and is primitively compelled to

[42] See especially Peacocke (1992).

[43] Peacocke's account is very complicated, and I abstract from some details of it here. I do not think that the simplifications affect the argument which follows. In particular, I ignore the fact that possession conditions will standardly have several clauses; for the relevance of this, see Peacocke (1992), 110-111.

[44] The details won't matter much here; for discussion of these requirements, see Peacocke (1992), especially Chapter 1.

make these transitions in virtue of their form)[45]

Letting 'T_c' stand for the transitions constitutive of possessing the concept denoted by 'c', we can then give the following account of the content of a belief state:

> A belief state x of an agent A has the structured proposition $<C_1, C_2, \ldots C_n>$ as its content $\equiv$ A is disposed to make transitions T_{c_1}, T_{c_2} $\ldots T_{c_n}$ to and from x[46]

We can then use the by now familiar principle that

> A believes $p \equiv \exists x(x$ is a belief state & A is in x & x has content $p)$

to arrive at an account of the beliefs of agents.

The key question, of course, is what transitions will figure in the possession conditions for concepts. There is no general answer to this question; but Peacocke has stated possession conditions for a number of concepts throughout his work. Perhaps the simplest is the possession condition for conjunction: roughly, to possess the concept of conjunction is to find the introduction and elimination rules for conjunction primitively compelling in virtue of their form.[47]

Perhaps it is true that we would not ascribe beliefs involving the concept of conjunction to agents who failed to meet this condition. The problem is that it seems that the number of concepts other than conjunction for which something similar is true seems vanishingly small. This can be shown by considering even a simple example like the concept of addition. According to Peacocke, an agent possesses the concept of addition just in case she finds transitions from

> $(m\ C\ k)$ is n

to

> $(m\ C\ s(k))$ is $s(n)$

primitively compelling in virtue of their form.[48] But imagine an agent who consults a calculator when presented with addition problems, and finds the answers only on the basis of the result delivered by the calculator. Or imagine an agent who computes addition problems intuitively: when presented with an addition problem, she simply sees the answer immediately.

[45] This is a condensed verson of the discussion in Peacocke (1992), 106-108. Again, I am ignoring the complications introduced by multiple-clause possession conditions.

[46] This is of course an oversimplification, since a set of propositional constituents does not in general determine a unique concept. Peacocke also discusses the question of what it is in virtue of which a belief state which includes among its contents some set of concepts has a content which is one mode of combining these concepts rather than another. This will not be important to what follows. As elsewhere, I am also abstracting from the problem of whether the possession conditions provide an account of what it is for a state to be a belief state as well as of what it is for such a state to have a given content.

[47] Peacocke (1992), 6.

[48] Peacocke (1992), 135-6.

She might be able to make transitions like these, but only on the basis of seeing first that the premise and conclusion of the inferential transition are both truths. Neither agent finds these transitions primitively compelling; but we should certainly be inclined to ascribe beliefs about addition to each.

And addition is, of course, an easy case; consider the only slightly more complicated case of multiplication. Following the example of Peacocke's treatment of addition, it is natural to think that the possession conditions for multiplication will involve finding the transitions which figure in a recursive definition of multiplication primitively compelling in virtue of their form. But either the relevant recursive definition will make use of the addition function, or it will not. If it does not, then it will be too complicated to expect ordinary users of the concept of multiplication to grasp it; if it does, then the possibility of an agent who learns multiplication without first learning addition is enough to show that it is false.[49] And things only get more difficult when we leave logical and mathematical examples behind.

This shows that Peacocke's account of concept possession is best construed as an account of a very specific kind of mastery of a concept. In general, the capacity to have beliefs involving a concept does not require having very specific dispositions to make certain transitions involving the concept, and to take some of these transitions rather than others as primitive.[50] So it seems that Peacocke's theory of possession conditions will account, at best, for a very small proportion of the beliefs of agents.

Though he might disagree about the proportions, Peacocke agrees that his account does not provide necessary conditions for belief.[51] Below, when we turn to the discussion of deference, we shall discuss the way in which he thinks that his account may yet have a role to play in a constitutive account of belief.[52] For now, the important point is that, setting aside the appeal to deference, Peacocke's account of possession conditions does not yield a plausible constitutive account of belief.

Conclusion

We have now reached a position with respect to the private language picture analogous to our position with respect to mentalism at the end of Chapter 3. There, I argued, we were in a position to conclude that mentalism is false, on the grounds that there is no true account of what it is or an expression of a public language to have a certain meaning in terms of the beliefs and intentions of users of the language. But, I argued, this still left the central claim of mentalism unchallenged: the claim that propositional attitudes like belief are more fundamental than the contents of linguistic expressions, whether in languages public or private.

We can now conclude that the private language theorist's project of giving a constitutive account of belief in terms of the contents of mental representations in an agent's language of thought is unsuccessful. But this still leaves the central intuition behind the private language picture unchallenged: the intuition is that we need language to explain a number of phenomena, but that public languages are shadowy abstractions incapable of playing any genuine

[49] This point about multiplication was suggested (in conversation) by Mark Johnston. The general point is owed to discussion in Mark Greenbeg seminar on Mental Content in the Fall of 2000.

[50] Compare the role of definitions in use of a concept; typically, switching the canonical definition of a scientific term does not involve changing the meaning of the term.

[51] See, among other places, Peacocke (1992), §1.4.

[52] See below, 7.2.2, pp. 147 ff.

explanatory role. One might preserve this intuition in the following weakened version of the private language picture:

> Perhaps we tend to talk as though the beliefs of agents are determined in part by the semantic properties of public languages; maybe this explains why attempts to give an account of belief in terms of properties of private languages fails. But the foundation of intentionality is still to be found in private languages; after all, talk about the meanings of expressions in public languages is just an abstraction from the meanings of expressions in the idiolects, or private languages, of agents. If we use these abstractions to ascribe beliefs to others, well, so much the worse for taking our ordinary ascriptions of beliefs and other propositional attitudes seriously. In any case, the basic level of intentionality *must* be at the level of private languages, since problems in individuating public languages show that they cannot play any genuine explanatory role.

We should, I think, distrust any picture of intentionality so willing to jettison our ordinary intuitions about the beliefs and other mental states of agents. But, fortunately, we needn't rely on these intuitions to reject this weakened version of the private language picture. As I'll argue in the next section, the private language picture fails to give an adequate picture of language, and the arguments against public languages often associated with the private language picture are not convincing.

Chapter 6

Public and Private Languages

Contents

6.1 The thesis of the priority of idiolects

One fundamental question about the study of natural language is whether public languages — languages typically shared amongst members of a group — or idiolects — languages specific to individual agents — come first in the order of explanation.

Note that, as stated, the issue is one about explanatory order and not, as is often said, about whether *there are* such things as public languages or idiolects. This is because, in my view, either sort of theorist should recognize the existence of both idiolects and public languages. Even a proponent of the priority of public languages should admit that an agent is capable of deciding to use a word in a nonstandard way, to mean something which no one else means by it; in at least this minimal sense, then, we know what it means to say that this agent speaks a private language, or idiolect, peculiar to her. (The proponent of the priority of public language may want to argue that her capacity to construct this idiolect is derived from her ability to speak some public language; but this is still not to deny that she speaks what is in some sense her own language.) Just so, even a proponent of the priority of idiolects should admit that there is some sense in which my sister and I speak the same language in which a monolingual Mongol and I do not. (The proponent of the priority of idiolects may want to argue that this is just loose talk for what is, strictly speaking, overlap in two idiolects; but this is still not to deny that there is *some* sense in which my sister and I share a language in which I and a monolingual Mongol do not.)

Though I take it that the dispute about the relative priorities of public and private languages is a fundamental question in the philosophy of language, it has received comparatively

little systematic discussion; or, rather, it has received discussion almost solely by proponents of the private language picture, most prominently Noam Chomsky. A recent exception is Jonathan McKeown-Green's *The Primacy of Public Language*.[1] Though McKeown-Green's project is centrally concerned with the philosophy of linguistics rather than with the relationship between mind and language, many of the arguments which follow are derived from McKeown-Green, and are discussed more fully in his work.[2]

What role are idiolects supposed to be play, according to the private language picture? Idiolects play three roles. First, they explain the meanings of utterances of agents: a sentence S as uttered by an agent A in a context will mean p in the context just in case the correct semantics for A's idiolect assigns p as the semantic content of S relative to the context. Second, they explain the contents of the thoughts of agents; that is, the proponent of the private language picture will say that what it is for an agent to believe p is for that agent to bear some relation to a sentence in her idiolect (I-language, language of thought) which means p in her idiolect. Third, insofar as the private language view countenances talk of public languages, these public languages will be explained as abstractions from idiolects of agents counted as members of the public language community.[3]

In the last chapter I argued that idiolects are not capable of playing the second of these roles; this chapter is addressed to the question of whether they might yet be able to play the first and the third. First, that is,

> Can we give an account of what it is for an utterance of a sentence by a speaker in a context to have a given meaning (relative to that context) in terms of that speaker's idiolect?

In §6.2 below, I argue that the arguments of the preceding chapter show that we should answer this question in the negative.

But this chapter must also answer the question of whether the private language theorist's view of public languages as abstractions from idiolects is the correct one. This is particularly important because the current critical examination of the mentalist and private language pictures of intentionality is part of an argument against the thesis of the priority of the individual; its role is to clear the way for a communitarian picture of mind and language. But proponents of the private language picture of mind and language often rely on 'in principle' arguments against communitarianism of roughly the following form: communitarianism gives public languages robust explanatory roles in accounting for, among other things, the beliefs and thoughts of agents; but, since there is no principled way of individuating public languages,

[1] McKeown-Green (2002).

[2] Special thanks to Jonathan for discussion of many of the issues covered in this chapter; the usual disclaimer that he should not be held responsible for places where I extend his views to slightly different areas of course applies.

[3] It is worth noting that there is also room for a view of idiolects which assigns them a much weaker role. On this view, idiolects play only the third of these roles; the contents of thoughts of agents, and the contents of their utterances, might then be explained in terms of the public language constituted by the relevant idiolects. But this view is really just a disguised version of communitarianism; it amounts to the view that public languages are constituted by some facts or other about the internal language processing system of speakers of the language. Regarding states of this system as having *meaning*, and hence as constituting a private *language*, is just terminological: these meanings play no explanatory role. I discuss this kind of view a bit in §6.3 below.

we should regard these 'public languages' as mere fictions. Treating public languages as capable of playing any sort of explanatory role must therefore be a mistake. Using the Chomskyan language of I-languages (idiolects) and E-languages (public languages), Peter Ludlow states the idea behind the argument well:

> An I-language is not a spoken or written corpus of sentences, but is rather a state of an internal system which is part of our biological endowment. ... From the E-language perspective, on the other hand, a natural language is a kind of social object ... and persons may acquire varying degrees of competence in their knowledge and use of that social object.
>
> I gather that, on Chomsky's view, such social objects do not exist and would be of little scientific interest if they did exist. To see why Chomsky's basic idea is right, consider the problem of trying to individuate such social objects. For example, simply consider the linguistic situation in Italy. We speak of "the Italian language," and we say that it is distinct from Spanish, but why? In large measure, Castilian Spanish and standard Italian are mutually intelligible when read or spoken slowly. Why don't we say that they are regional variants of the same language?[4]

Examples like these are common in the literature; the background thought is, of course, that questions like the above are not answerable in any non-arbitrary way, and thus that public languages should be regarded as, at best, abstractions from real languages — i.e., idiolects. If communitarianism is to be a defensible position, this challenge must be answered.

6.2 Idiolects and the meanings of utterances

First, though, I turn to the question of whether the arguments of Chapter 5 count against the attempt to provide an idiolect-based account of the meanings of utterances. It takes little work to show that they do.

A constitutive account of utterance meaning will have the same form as a constitutive account of belief. Just as we can ask, 'What is it for an agent to believe p?,' we can ask, 'What is it for a sentence uttered by an agent A in a context C to mean p (relative to C)?'[5] As above, to answer this question we need a completion of a formula of the form

> An utterance of a sentence S by an agent A in a context C means p (relative to C) $\equiv \ldots$

For a communitarian, who thinks that public languages are more fundamental than idiolects, an answer to this question will come in two parts: a constitutive account of what it is for a sentence in a public language to mean p (relative to the relevant context), and an account of what it is for an agent to be speaking that language. It is natural to think that a proponent

[4] Ludlow (1999), 17.

[5] Note that this is not the question discussed in Chapters 2 and 3; there we were asking what it is for a *speaker* to mean something by an utterance. Now we are asking what it is for an *utterance* of some speaker to have a given meaning. These are different topics, as the possibility of meaning something by an utterance other than what the sentence means shows.

of the private language picture of intentionality will have a different strategy for answering the question. Since she takes the languages spoken by agents to be their idiolects (rather than public languages), she will think that question should be answered by giving a constitutive account of what it is for the sentence uttered to mean p in A's idiolect.[6]

How might we go about saying what it is for an utterance to have a meaning in the idiolect of an agent? Chomsky's view is representative:

> We can perhaps make sense of Lewis's notion, "the language L is used by a population P," but only indirectly, through the medium of a specific characterization of L, a grammar (or perhaps even more indirectly, through a class of weakly equivalent grammars). This notion unpacks into something like: each person in P has a grammar determining L in his mind/brain.[7]

The idea here is that which language we should take a person to be speaking depends upon the syntax and semantics of their internal states. (And that which language we should take a group of such agents to speak depends upon overlap, or similarity, in the syntax and semantics of their internal states.) If this is right, then the meanings of expressions in the language spoken by an agent depend upon the contents of their internal states. But then we seem to be stuck with the same (by now, familiar) list of choices for the properties of those internal states which might determine their contents. The private language theorist might take the meanings of representations in an idiolect to be fixed either by relations between those representations and other internal states, by relations between those representations and facts external to the agent (whether these relations are spelled out in terms of causation, covariation, counterfactual dependence, asymmetric dependence, or evolutionary function), or by some combination of internal and external relations (as in conceptual role theories of content).

But if it is true that the only resources available for a private language account of utterance meaning are those available for a private language account of belief, then the failure of the former shows the failure of the latter. For take some counterexample to a private language account of belief, in which the theory fails to account for some belief p of an agent. In such cases, the theory of content yields the result that there is no sentence in the agent's language of thought that has content p. But now suppose that the agent in question expresses this belief by uttering a sentence which means p. Presuming, as above, that the private language theorist should account for the meanings of sentences as uttered by an agent in terms of the semantics of that agent's idiolect or language of thought, it follows that the private language theorist will not be able to account for this speaker's utterance being one which means p. The

[6] I take it that there are two broad strategies for doing this; which one a theorist takes will depend on how seriously she takes the claim that I-languages are internal mechanisms of agents. If she takes this claim very seriously, then the answer will come in two parts: she will have to define some relationship between the outward sentences uttered by agents and states of this internal mechanism such that the former inherit their meaning from the latter. But one might take the relationship between an idiolect and the relevant internal mechanism to be a bit looser than this, and take outwardly uttered sentences to be, in some sense, part of the idiolect of the agent. The difference between these will not matter much for present purposes.

[7] Chomsky (1980), 84. For similar views, see Schiffer (1993), §IV-V, and Laurence (1996), 282-285. Both are explicit in taking the Chomskyan private language picture to amount to a reduction of sentence meaning to the contents of states of an internal language processor.

case imagined, after all, is precisely one in which the agent in question has no sentence in his idiolect or language of thought which means p.

6.3 The autonomy of public languages

So far in Part II of this essay we have been discussing the problems encountered by a private language theorist trying to give constitutive accounts of belief and of meaning. By now I hope that it is clear that such a theorist lacks the resources to give either sort of account.

But the problem with the private language picture of intentionality is deeper than a lack of theoretical resources. I also want to claim that the private language picture gives the wrong picture of the nature both of public languages and of thought. The former point — that the private language account gives the wrong kind of account of public languages — can be brought out by considering a much weakened version of the private language picture, which the foregoing does not refute.

Suppose that the private language theorist were to grant the point of the preceding sections and chapter, and admit that there is no constitutive account of idiolect-meaning which makes idiolects capable of serving an important role in a foundational account of either the beliefs of agents or the meanings of sentences uttered by agents. Such a theorist might still retreat to the following position:

> Public languages have a genuine explanatory role to play with respect to the meanings of utterances by agents who speak such a language, what those agents mean and assert by their utterances, and what those agents believe. But the meanings of public language expressions are derived from the meanings of expressions in the idiolects of members of the relevant linguistic community. So idiolects do determine the contents of the thoughts, other propositional attitudes, and sentences understood by agents; it's just that they do this indirectly, via a public language.

I do think that this is the natural next move for the private language theorist; but it is not a very stable position. For one, it involves giving up skepticism about the explanatory power of public languages, which is one of the main motivations for the private language picture. More importantly, though, it makes the point of talking about idiolects obscure; so much so that this sort of view seems almost to be a notational variant of a kind of communitarianism. One kind of communitarian might think that the meanings of public language expressions are determined by the internal states of speakers of the language; the only difference between this kind of communitarianism and the present weakened version of the private language picture is that the latter insists on saying that these internal states have meaning, and so constitute a kind of language, even though the meanings of these states play no direct explanatory role.

But the real problem with this view, and what makes it unstable, is that it tries to have things both ways. The heart of the private language theorist's view of public languages is that (again, insofar as talk about public languages is allowed to be intelligible) public languages are essentially abstractions from the linguistic capacities of individuals who speak the language. This is, in some ways, a very intuitive view. I have tried to put some pressure on it by arguing that public languages cannot *just* be abstractions; we need facts about public languages to explain a broad class of facts about meaning and the contents of propositional attitudes. The sense in which the present weakened private language picture tries to have it both ways is

that it claims both that public languages are mere abstractions, and that public languages can figure in genuine explanations.

If we accord public languages this kind of explanatory role, what content is left to the claim that public languages are mere abstractions from the linguistic activities of speakers? Maybe this: public languages, unlike things which are not mere abstractions, like tables and people, lack a certain kind of persistence through time. At any given time, the properties of a public language defined as the language of some group are fixed by the properties of members of that group; the language then automatically tracks the relevant properties of members of the group as those members change through time. To put the point metaphorically, we might say that, according to this private language theorist's view of public languages, public languages lack *autonomy*.

This hypothesis is borne out by recalling the theories of meaning advanced by proponents of the private language picture discussed in the preceding section. There the contents of sentences in a public language — as used by some group — was defined in terms of the contents of internal states of language users. It follows from this that languages lack autonomy, in the relevant sense: according to this picture, the semantic features of a language shift along with the content-determining properties of internal states of language users.

As noted above, the claim that public languages lack autonomy in this sense does not sit comfortably with the claim that properties of public languages can figure in genuine explanations. But there is no outright contradiction in holding both of these views; a better way to attack this view of public languages is to point out that public languages do, in fact, have a certain kind of autonomy. To see this, consider the following thought-experiment from Wittgenstein's *Remarks on the Foundations of Mathematics*:

> Let us imagine a god creating a country instantaneously in the middle of the wilderness, which exists for two minutes and is an exact reproduction of a part of England, with everything that is going on there for two minutes. Just like those in England, the people are pursuing a variety of occupations. Children are in school. Some people are doing mathematics. Now let us contemplate the activity of some human being during these two minutes. One of these people is doing exactly what a mathematician in England is doing, who is just doing a calculation.—Ought we to say that this two-minute man is calculating? Could we for example not imagine a past and a continuation of these two minutes, which would make us call the processes something quite different?[8]

To extend Wittgenstein's example for our purposes, imagine that the person doing what the mathematician in England did has uttered the sentence, "I can't believe that Clive challenged me with such a trivial problem." (Suppose that there is some salient Clive in the original English scene recreated in the wilderness.) Now, rather than asking whether we ought to say that this two-minute man is calculating, we can ask whether this two-minute man has said something which means that he can't believe that Clive has challenged him with such a trivial problem. Wittgenstein's implied answer, of course, is 'No.' If this is the right answer, then this seems to rule out any view of public languages as shadowy abstractions with no independent existence. Wittgenstein's example would then be a case in which we have a

[8]Wittgenstein, *Remarks on the Foundations of Mathematics*, Part VI, §34. Thanks to Tomas Hribek for bringing this passage to my attention

difference in meaning between two situations alike except with respect to the history of the linguistic community. If this were right, it would then be natural to think that the difference in meaning is to be explained by the persistence of the semantic properties of the public languages in question.[9]

Wittgenstein's implied answer is not meant to just rest on intuition; the last sentence of the passage suggests an argument. The idea here is that we can imagine many possible histories of a place during which this two-minute scene takes place. It seems that in these different possible histories, many things might be different during this two-minute stretch, even if we suppose a 'molecule-for-molecule' identity between the two-minute world and the relevant segments of these various possible histories. Among the things which might be different are the meanings of words in the sentence uttered by our two-minute man. The challenge for the proponent of the view that we should answer 'Yes' to our question — that we should say that the two-minute man *has* said something which means that he can't believe that Clive has challenged him with such a trivial problem — is to say why we should take the meaning of the sentence uttered by the two-minute man to have the same meaning as the sentence uttered by his duplicate in England, rather than either the same meaning as the one uttered by his duplicate in one of the other possible histories of a place that correspond to the two-minute world in the relevant sense, or the an altogether different meaning.

On the face of it, this seems like a difficult challenge to meet; it just seems arbitrary which of the possible histories we choose for our interpretation of the two-minute world. But it is worth noting that, for purposes of the present argument, it does not matter whether we answer Wittgenstein's question with a 'Yes' or with a 'No.' Even if the private language theorist clings steadfastly to the thought that the sentence uttered by the two-minute man means the same as that uttered by his duplicate in England, it follows that the sentence uttered by the two-minute man means something different than one of the other possible histories identical in relevant respects to the two-minute world. And this is enough to run the same argument as above.[10]

One can always deny the modal intuition which this form of argument requires: namely, that there are various possible histories of places which correspond to the two-minute world in every respect other than the meanings of expressions which figure in those histories, but differ in that respect.[11] But, on the face of it, this is not a very easy intuition to deny. Proper names of no longer existing people provide the easiest place to defend the intuition. Imagine a possible world just like ours except that the names of Socrates and Plato are reversed. Then imagine a two-minute history of that possible world (restricted, perhaps, to some specific linguistic community) which is identical in every detail to some two-minute history of our world. Suppose that the name 'Socrates' is uttered in both two minute stretches. This is

[9] Why can't it be explained by the persistence of the semantic properties of the idiolects of users of the language? An answer to this question requires a discussion of the way in which agents, by *deferring* to other speakers, might have idiolects whose contents depend upon the contents of expressions in the idiolects of those speakers. For critical discussion of this idea, see §7.2.2, pp. 147 ff. below.

[10] Wittgenstein's concern in this passage was with intentional action, not with linguistic meaning; it is interesting that the same kind of argument as given here with respect to linguistic meaning can be given with respect to intentional action types. For some further suggestions about construing the relationship between bodily movements and intentional actions on the model of the relationship between expressions and their meaning, see Chapter 9, note 19, p. 194.

[11] Of course, they must differ in their history; that's part of the point of the case.

enough to provide a case of the relevant kind, so long as you grant that the utterance of 'Socrates' in our world refers to Socrates, whereas the utterance of 'Socrates' in the possible world refers to Plato. Is it really plausible to deny that this is a possible scenario?[12]

Of course, one can take this point about the autonomy of public languages too far. Public languages are rooted, in some way or other, in the practices and linguistic abilities of speakers of the language and their predecessors, along with facts about the environment in which those speakers and their predecessors are embedded. Saying how these practices and abilities determine the properties of languages is a difficult task, and one which any fully adequate defense of communitarianism must address. The point of emphasizing the autonomy of public languages is only to argue that the problem with the private language picture of public languages is not that we have not found the right theory of content for mental representations, nor that we have not found the right way of giving an account of the meanings of the sentences speakers utter in terms of the contents of those representations, but rather that the basic thought of the private language theorist — that public languages are a kind of abstraction or idealization, a mere shadow of the real facts about the language faculty of members of some group — is a mistake.

6.4 The case for skepticism about public languages

At the outset of this chapter, I outlined two tasks to be accomplished: first, to show that idiolects, are not only unable to provide a constitutive account of thought, but also of utterance-meaning, and second, to respond to the case for skepticism about public languages. At this stage, the first task is behind us. But, without addressing the second task, it is fair to regard our current position as something of a dilemma. On the one hand, idiolects (I-languages, languages of thought) seem incapable of playing any sort of foundational role. This indicates that, at least in trying to explain the meaning of an utterance of a sentence in a context, we should appeal to the semantics of public languages. On the other hand, we have an apparently compelling argument from Chomsky and like-minded private language theorists that the absence of adequate criteria for individuating public languages shows that they are not the sorts of things capable of playing any kind of explanatory role. The purpose of this section is to resolve this dilemma by responding to the case for skepticism about public languages.

The argument from language individuation

The main argument against taking public languages seriously is due to Chomsky:

[12]McKeown-Green develops the following ingenious scenario, which makes the same point as Wittgenstein's: "A child is born; its parents hastily choose a name by asking that the child be named after the doctor (of whose name they and all the nurses are ignorant). The doctor, who knows nothing about the arrangement, is asked to type his name into a computer-generated form. All the relevant parties sign it, but without reading it carefully. Everything is legal. Tragically, everyone involved and the official records perish very shortly afterwards — before anyone has referred to the child by name. It is highly likely — though not certain — that all evidence of the naming has gone. Does the child have a name?" ((McKeown-Green (2002), §5.2). McKeown-Green's answer seems to me convincing: "I think it does, at least for a short while after the tragedy. If, a day or so later, someone was able to produce conclusive evidence that the child had been named 'Preston the Corrigible,' wouldn't we all say 'So that's what his name is!'[?]"

> the commonsense notion of a language has a crucial sociopolitical dimension. ... That any coherent account can be given of "language" in this sense is doubtful; surely, none has been offered or even seriously attempted. Rather, all scientific approaches have simply abandoned these elements of what is called "language" in common usage.[13]

Recall Ludlow's statement of this point, quoted above. The idea is that we are presented with a bunch of language users, and each uses languages in slightly different ways than most any of the others. We find a gradation from any one speaker to any other, through any number of intermediate speakers. The challenge is to separate this mass of language users into neat 'public language communities'. The problem is that there seems no principled way of doing this. It is as if we were presented with the grains of sand on a beach, and asked to put them into neat categories. Surely we could put them into categories; but is there any principled way of doing so?

To be sure, we *do* often manage to classify language users into 'public language communities'; we say that some people speak French, others Chinese, and so on. The problem is that, as Chomsky rightly points out, these 'folk' classifications do not seem to correspond to any important linguistic differences. Rather, they seem to be based on contingent 'sociopolitical' facts: "a language is just a dialect with an army and a navy." This seems a poor foundation on which to base a linguistics (or a philosophy of language).

The mutual intelligibility criterion

The natural first choice for responding to this argument is to appeal to the notion of mutual intelligibility: surely, after all, two language users speak the same language just in case they can understand each other.

But this mutual intelligibility criterion faces an apparently insuperable objection. Mutually intelligibility supervenes on similarity of language use; and, like any criterion founded on similarity, mutual intelligibility fails to account for the transitivity of '___speaks the same language as ___.' McKeown-Green asks us to consider a speaker of a low German dialect, addressing a native Dutch and a native Flemish speaker. It might well be that the speech of the first is intelligible to both members of her audience, without either member of her audience being intelligible to the other. The mutual intelligibility criterion implies that the speaker of low German must have been speaking two languages simultaneously. But this is surely absurd.[14]

Linguistic project relativism

Jonathan McKeown-Green has proposed a quite different response to the argument from language individuation. He calls his response *linguistic project relativism*, and writes

> Granted, human linguistic practice, regarded as a fabric of predominantly social facts, does not subdivide neatly and naturally into objects for scientific scrutiny

[13] Chomsky (1986), 15.

[14] For an excellent discussion of the nature of Chomsky's case against principled individuation of public languages and the reasons for the failure of answers like (this version of) the mutual intelligibility criterion, see McKeown-Green (2002), §4.1-4.2.

> any more than human society subdivides neatly into internally cohesive communities — any more, if it comes to that, than the material universe can be neatly subdivided into closed mechanical or biological systems. But notice that what counts as the object of study for a particular scientist in a particular context always depends on the phenomenon she happens to be investigating: she carves up her terrain in some way calculated to reveal distinctions that are relevant to her topic. Change the topic and with it the distinctions, and there will be a different carving ceremony. ...
>
> My principle of language individuation is this: *make like any other sort of scientist; carve up the terrain (rule-governed speech patterns) to suit the topic; individuate project by project and idealize away from noise (like informant mis-speaking and measurement error).* Remember, your noise is another linguist's primary data![15]

One way to bring out the force of McKeown-Green's response is to imagine an analogue of Chomsky's objection being brought against a scientist who studies ecosystems. A skeptic might object that the physical objects in the world form a continuum, with most every such object differing from most any other, and with a number of other objects bridging the gap in different respects; so it is surely arbitrary to baptize some collection of such objects as an ecosystem, and make it the object of serious scientific study. It seems clear that the scientist might justifiably grant the first part of the skeptic's point, and but deny that this makes her project illegitimate. She might say, "Surely the physical objects in the world do not separate themselves into neat groupings; but I am interested in studying property X, and, for these purposes, the distinctions I make are suitable." Just so, it seems, a linguist studying the language of a group of language users might respond to Chomsky's objection.

As a point about the proper methodology of linguistics, McKeown-Green's reply to the arguments from language individuation seems to me entirely compelling. But I do not think that linguistic project relativism can be the whole story about what individuates public languages. According to linguistic project relativism, any language user A may be considered a member of more than one of several distinct but overlapping linguistic communities.[16] The languages of these distinct communities may, and probably will, have a syntax and semantics which overlap in the following way: some expressions are expressions of the same grammatical type in both linguistic communities, but have (perhaps only slightly) different meanings in the two communities. Now consider a case in which A assertorically utters a sentence containing such an expression. The pressing question is then: which of the several distinct but overlapping linguistic communities of which A is a member is relevant to the evaluation of this utterance? This question is pressing because we need an answer to it to answer several other questions which any adequate picture of mind and language must answer: What is the meaning of this sentence as uttered in this context? What did the speaker mean, or assert, by uttering the sentence in this context? If he was sincere, what belief did the speaker express by uttering this sentence?[17]

[15] McKeown-Green (2002), §4.3. Emphasis in the original.

[16] I ignore, for now, the possibility of an agent in a linguistic community of one; perhaps such an agent could not be regarded as a member of more than one distinct linguistic community.

[17] It is tempting to reply to these problems by denying that they are really problems about *language* — by claiming, that is, that what speakers assert, mean, or believe are not to be explained in terms of facts about public languages. But I have already argued against this response by arguing against

These are not questions which, so far as I can see, linguistic project relativism is well-suited to answer. It would be absurd to say that what the speaker means or believes depend upon the interests of some linguist studying one or more of these linguistic groups. It seems equally absurd to say that, in general, there is no fact of the matter about what, precisely, the speaker asserts or believes in such contexts, and that all we can say is that the speaker may be considered as asserting one or another proposition depending on the sort of study of the speaker we are interested in doing.

This objection to these applications of linguistic project relativism is not, I think, a point of disagreement with McKeown-Green; he writes,

> There is room for confusion here. We are trying to identify the objects that linguists study and to do that, we have been trying to get a feel for the linguistic landscape. We are *not* trying to do conceptual analysis on the folk notion, *language*. My relativistic principle of 'language' individuation is intended to inform scientific taxonomy, not to legislate speakers' attitudes. I don't think that which language you should regard yourself as speaking depends on which project some linguist is undertaking. ... At any rate, if my claim that languages are project-relative offends, try replacing 'languages' with 'objects of study in linguistics.'[18]

So it seems that, even if linguistic project relativism succeeds as a response to the language individuation problem as applied to views of linguistics as a social science rather than a branch of cognitive psychology, it does nothing to defuse that objection as applied to the communitarian claim that facts about public languages are partially constitutive of various of the propositional attitudes of speakers.

It does, however, point the way to a more adequate response to this form of the argument from language individuation. One fundamental element in McKeown-Green's proposal is that it gives a useful way of meeting Chomsky's objection without offering context-independent criteria for saying when two speakers speak the same language. This is, in my view, the crucial idea behind linguistic project relativism. When one reads versions of the argument from language individuation, it is fairly clear that the reason why theorists presume that no response to the argument is possible is that they tacitly assume that any answer would have to issue in some context-independent criteria. And surely these theorists are right than when we consider cases of actual language users, the task of giving a non-arbitrary and context-independent completion of the formula

> For any two language using agents x and y, x and y speak the same language $\equiv$ ...

seems hopeless. McKeown-Green's insight is that we can solve the problem of the individuation of public languages without thinking that there is any satisfactory completion of this schema.

To see the importance of this point, consider an analogy. According to the view of public languages I have been encouraging, public language communities are social institutions which

mentalist theories of speaker-meaning and assertion in Chapter 2 and Appendix B, and against individualist accounts of belief in Chapters 4 and 5. Thus the present dilemma.

[18] McKeown-Green (2002), §4.3

have a kind of autonomy. In this, they are similar to entrenched social customs, like certain standards of etiquette. There is a sense in which membership in a certain community is sufficient for you to be bound by certain standards of etiquette, even if you do nothing in particular to ensure that you are bound by those standards, rather than other ones; there is also a sense in which standards of etiquette 'automatically' persist through time in a community, without anyone doing anything much to keep them in play. To be sure, you can opt out of the practice of paying heed to standards of etiquette; but one can also opt out (theoretically, at least) of the practice of speaking a certain public language. It is also true that standards if etiquette would cease to be in play if, for sufficient time, no one paid them any heed; but, again, the same is true of public languages. We can also raise skeptical worries here analogous to those raised by Chomsky with respect to the individuation of public languages: the standards of etiquette in different places do not fall into into neatly ordered bunches; rather, there is a gradation of rules of etiquette from one place into the next. But (I suggest), as in the case of public languages, it would be rash to conclude that there are no standards of etiquette.

So, by way of approaching the question about the individuation of public languages, we can ask: what makes it the case that a certain individual is bound by, or falls under, one standard of etiquette rather than another? It does not seem as though there is any very exciting answer to this question. There are certain practices in certain places of correcting certain kinds of behavior and encouraging others. In nearby places, these practices may differ slightly, but be close enough that for some purposes we regard them as having the same system of etiquette. For many people, it will be vague whether they fall under one rule of etiquette rather than another; for many people, it will be true that in some contexts they fall under one rule of etiquette, and in a different context another. But, given an individual and a context, we have some grasp of the rules required for determining which rules of etiquette are relevant; even if there is no principled context-independent criterion for the individuation of systems of etiquette, it seems plausible that there are criteria for saying which community's standards of etiquette are relevant to a particular activity of a particular individual at a particular time.

The same verdict seems to be in order for public languages. The speaking of a public language is a social practice which, like etiquette, is founded upon interaction between different individuals. The extent of the practice will, like the practices characteristic of etiquette, often be a vague matter, and there may be no non-arbitrary way of dividing up public languages apart from consideration of particular contexts of utterance. But rough guidelines as to the extent of the practice may be gleaned from the extent to which the standards of the practice are enforced among different agents in different places. Among the facts which might be relevant to determining which linguistic community is relevant to the evaluation of a particular utterance might be the intentions of the speaker to be understood by certain individuals, the expectations of speakers that they be held responsible to certain standards, and dispositions of speakers to address different audiences differently; perhaps facts about the audience in certain communicative situations might also be relevant.[19]

[19]Multilingual speakers show that we should relativize languages not to speakers, but to contexts of utterances by speakers; multilingual speakers may also raise a number of further problems to do with explaining what it is for a speaker who speaks two languages to be speaking one rather than another at a given time. These problems are exacerbated if Peter Ludlow is correct in claiming that many of us often switch from one linguistic community to another routinely in our daily lives; a philosopher

A Chomskyan is likely to object that the same sorts of objections may be raised against these criteria as against mutual intelligibility, or any criterion which supervenes on similarities between agents: (i) because there is a gradation between the expectations, dispositions, and intentions of different speakers, any way of dividing up these speakers into public language communities will be arbitrary, and (ii) for reasons similar to those given in the discussion of the mutual intelligibility criterion, this proposal will entail the failure of the transitivity of '___speaks the same language as ___.' But these points cut no ice against the present proposal. We have already granted the Chomskyan's point that the linguistic world does not come divided up into neat public language communities which may them be made to do theoretical work, any more than human society may be split into neat groups on the basis of rules of etiquette. The present proposal is that public language communities are relative to specific utterances of individual agents; all that is needed for public languages to do explanatory work is for there to be rough criteria capable of fixing the linguistic community relevant to the utterance of a particular individual at a particular time. The Chomskyan, so far as I can see, has no principled argument against the thought that the kinds of facts about speakers mentioned in the preceding paragraph will be capable of providing such criteria.

Now, it is important to be clear about the nature of this response to the skeptic about public languages. I have not given a worked-out account of the individuation of public languages (relative to agents and utterances); in this sense, I have not carried out what Jim Pryor calls the *ambitious skeptical project* of refuting the skeptic on his own terms.[20] The Chomskyan can still insist, for all I've said, that we can give no account of the kind I've gestured at. But I do take myself to have, following McKeown-Green's lead, given a *modest* response to the skeptic: to have shown how the believer in public languages can justifiably resist the force of Chomsky's argument against public languages.[21]

6.5 Four explanatory uses for public languages

A diehard opponent of public languages, though, might respond to these points by asking whether there is any *positive* reason to believe that we should take the existence of public languages seriously. After all, she might insist, the best current research program in linguistics has no use for public languages; so why should we posit such things, apart from a dubious

might call a colleague a "realist" in the context of a philosophical discussion, but deny the title of "realist" to the same colleague in the context of a dinner table discussion of that colleague's political views. The explanation is not that the philosopher changed his mind about whether his colleague is a realist; rather, the linguistic community relevant to he evaluation of the first utterance is not the same as the one relevant to the evaluation of the second utterance. (See Ludlow (1995).) It is unclear to what extent this case might be regarded as a simple case of ambiguity in a shared language.

[20]Pryor employs the distinction between ambitious and modest responses to skepticism in the context of a discussion of skepticism about the external world in Pryor (2000).

[21]A different argument against public languages has been given by Davidson in his "A Nice Derangement of Epitaphs." The argument is basically that public languages are dispensable for purposes of explaining communication between speakers, and so that they should not be posited. It is, however, a presupposition of Davidson's argument that literal meaning (what he calls 'first meaning') can be explained along roughly Gricean lines. But we have already seen in Chapter 2 that Gricean accounts of literal meaning are unsuccessful. For a nice reply to Davidson's argument, see Dummett (1986).

prior attachment to them?

This is not, I think, a very serious challenge to the view being developed. First, there is room to doubt whether the practice of current linguistics shows that idiolects are really the prime object of study.[22] Second, even if the phenomena studied by linguists do not require public languages for their explanation, that would hardly show that public languages are not needed for the explanation for some other phenomena, in which linguists happen not to be interested. Third, and most important, it is not hard to pick out phenomena most naturally explained in terms of facts about public languages. Indeed, we have already noticed a number of these.

Action-based propositional attitudes

In Chapter 2, we considered several attempts to give an account of what it is for a speaker to mean something by an utterance which did not make explanatory use of facts about public languages. Each of these failed, and did so in an instructive way: the cases for which each account failed were cases in which, intuitively, a speaker meant p by an utterance in virtue of the fact that the utterance was done sincerely and seriously, and that the sentence uttered meant p in the public language of the agent. I argued there that this result extended to other action-based propositional attitudes, and that this provided strong evidence that a constitutive account of such attitudes — of speaker-meaning, saying, and asserting, for example — must be a communitarian account, which explains these attitudes partly in terms of public language facts.[23]

Belief

Further, the arguments of Chapters 4 and 5 have pointed up a number of problems with various kinds of attempts to give a constitutive account of belief which does not make use of facts about public languages. These arguments, I think, give us good grounds for thinking that, if there is any true constitutive account of belief, that account will have to be given partly in terms of facts about the public languages spoken by agents. Indeed, in the next chapter I shall discuss an argument against functionalist accounts of belief of the sort available to proponents of the mentalist and private language pictures of intentionality which shows this more directly. Then, beginning in Part III of this essay, I shall begin the sketch of a positive communitarian account of belief. If successful, this account will show that facts about the beliefs of agents, like facts about what agents assert and mean by their utterances, like the autonomy of public languages, and (perhaps) like ordinary cases of communication, is a phenomenon the explanation of which requires taking public languages seriously.

The autonomy of public languages

A third explanatory role for public languages comes from the considerations advanced above in favor of the thesis that public languages possess a kind of autonomy. It is hard, for the reasons given above, to see how these phenomena can be reconciled with a view which takes public languages to be mere epiphenomena, or abstractions from more fundamental facts

[22] See, e.g., Soames (1984) and McKeown-Green (2002), esp. Chapter 6.

[23] For a sketch of such an account, see Appendix C below.

about idiolects. Rather, it seems that we should accept the result that the properties of public languages at a time can sometimes explain the properties of those languages at a later time.

Communication

A less obvious explanatory use for public languages has to do with our ordinary practice of communication using language. In his "Language and Communication", Dummett writes

> What, then, constitutes a subject's understanding the sentences of a language in a particular way? ... is it ... his having internalized a certain theory of meaning for that language? If this is the internalization, then indeed his behaviour when he takes part in linguistic interchange can at best be strong but fallible evidence for the internalized theory. In that case, however, the hearer's presumption that he has understood the speaker can never be definitively refuted or confirmed: he can only have evidence that he has done so, which falls short of being conclusive. So regarded, communication does rest ultimately on faith — faith that one has hit on the very theory of meaning that one's interlocutor has internalized.
>
> The only escape from the absurdity of this conclusion is to treat the supposedly internalized theory of meaning, not as constitutive of the speaker's attaching the meanings that he does to his words, but, indeed, as an empirical hypothesis to explain what enables him to use the language to express those meanings.[24]

Dummett here is focusing on the fact that, according to the private language theorist, each of us possess our own languages; communication between individuals, then, must be construed, not as the exchange of information using a shared medium, but rather a more or less fortuitous confluence of separate languages.

I think that it is hard not to share Dummett's thought that this is a strange picture of what happens in ordinary cases of communication. The phenomenology of communication seems to be such that when another English speaker says something to me, there is no guesswork involved; I simply take in the other person's words, and grasp their sense. No doubt this is most easily explained by our sharing a public language as a medium of communication. But this can only be construed as a *prima facie* case for communitarianism; it could be a fortuitous but contingent fact that our idiolects overlap and our language modules work so as to create the illusion of a shared medium of communication.

Dummett's main point, though, is not about the phenomenology of ordinary cases of communication. The crux of his argument is that the private language picture makes communication 'rest on faith'; and the idea seems to be that this opens the door to an absurd form of skepticism.

Here, though, we must be careful; on the face of it, this argument proves too much. After all, I *can* wonder whether an interlocutor is using some word in the same way that I use it; sometimes, she may not be, even if, in every ordinary sense of the word, we share a language. If Dummett's point were that public languages made this sort of thing impossible, then this would constitute an argument against, rather than for, a communitarian account of communication.

[24] Dummett (1989), 180-1.

But this is not, I think, the best interpretation of Dummett's thought here. His idea is not that this local form of skepticism about another meaning the same by a word as me is impossible, but that a certain global kind of such skepticism is impossible. And here, I think, he is on to something. Can we really imagine it turning out that no one with whom we had ever conversed meant the same by any of their words as we do? Certainly, we can imagine this being so because of some other form of radical skepticism being true; if, for instance, we were brains in vats and only seemed to ourselves to be interacting with others, or if those agents who seemed to be interlocutors were cleverly disguised robots. But can we conceive of this kind of skepticism under the supposition that our interlocutors are genuine human agents, much like ourselves? I am not at all sure; there does seem to be some kind of incoherence here not present even in the hypothesis that we are brains in vats. But absent an argument for this incoherence, we can only advance the following cautiously conditional claim: because the situation envisaged by this kind of skepticism seems to be possible on the private language picture of communication, if such situations are impossible, a fourth explanatory use for public languages is to explain this impossibility.

Part III

THE COMMUNITARIAN PICTURE

Chapter 7

Why Individualism Failed

Contents

7.1 Two presuppositions of individualism

In Parts I and II of this essay, we have discussed the two versions of individualism: the mentalist and private language pictures of intentionality. I have argued that the central claims of each about the nature of mental states and about the nature of linguistic meaning are false. The question I want to focus on in this chapter is: Why do individualist views of the mind fail?

Every plausible individualist view of mental states with which we have been concerned — most centrally, theories of belief — incorporates two claims. The first of these is the guiding principle of individualism:

> Constitutive accounts of mental states like belief need not make any use of social facts about the public languages of agents.

I argued in Chapter 4 that this language independence thesis entails another:

> Beliefs and like mental states are internal states of agents; i.e., they are constituted by second-order properties of internal states of agents.

I claim that both of these tenets of individualism are false. As against the language independence thesis, I claim:

> Public languages can be vehicles of thought; i.e., the correct constitutive account of many propositional attitudes will be given partly in terms of the meanings of sentences in public languages and the relations in which agents stand to those sentences.

As against the internal state thesis, I claim:

> Beliefs and like mental states are not internal states. Rather, they are constituted by dispositions to action on the part of the relevant agent.

In this chapter, I shall consider these two mentalist assumptions in turn, and argue for their replacement by these two communitarian theses.

7.2 Public languages as vehicles of thought

So far, we've surveyed a number of ways in which an individualist might attempt a foundational account of the contents of beliefs, and found decisive problems with each. Why is this? Is it because we simply have not hit upon the right mentalist theory yet? This seems unlikely; we have considered views which take the contents of beliefs to be constituted by the relations borne by belief states to each other, to the world, and to both, and found problems which seem endemic to each. At this stage, facts about the beliefs of agents may begin to appear a bit mysterious; look where we might, we seem unable to find a foundation for facts about belief.

Is this negative result, then, due to facts about belief being *primitive*, in the sense that there is no interesting story to be told about what it is for an agent to believe p? This would not be an absurd result; facts about belief might supervene on other facts without there being a constitutive account of the former in terms of the latter. Though not absurd, this diagnosis is, I think, a bit hasty. Rather, I think that the failure of individualist accounts of belief is to be explained by the fact that individualists are simply looking in the wrong place. In particular, individualists look to the properties of believers to the exclusion of properties of the social linguistic groups in which they are embedded. In slogan form, the individualist ignores the role that public languages can play as vehicles of thought.

7.2.1 Belief and language use

What does it mean to say that public languages can be vehicles of thought? To say that public languages are vehicles for a certain propositional attitude R is to say that, at least sometimes, *what it is* for an agent to bear R to a proposition p is for that agent to bear some relation R' to a sentence in her public language which means p.

A philosopher interested in answering the questions under discussion in this essay is interested in finding the nature of various mental states. Since facts about the natures of things hold necessarily, a first step in this project is uncovering necessary truths about the mental state in question; so an initial test for whether public languages are vehicles for a given propositional attitude is whether we can find necessary truths connecting those propositional attitudes with relations toward sentences of public languages with certain meanings. Just which attitudes toward sentences are relevant will depend upon which propositional attitude

is under investigation. Our example throughout has been belief; when we turn again toward this attitude, a plausible necessary truth of the kind sought is not hard to find.

The connection between language and belief I have in mind is just this: if one is disposed to sincerely accepts a sentence which means p, and one understands the sentence, one thereby believes p. This disquotational principle may be expressed as follows:[1]

$\Box\forall a\forall S\forall p$ ((a is disposed to accept S & S means p) $\rightarrow$ a believes p)

Might the individualist simply deny the disquotational principle? I don't see how; so far as I can tell, there are very few facts about belief more obvious and striking than the fact that one can infer from someone's saying "Grass is green" that she believes that grass is green.[2] Of course, the agent in question must understand the sentence, must be serious, and perhaps must be minimally reflective; but we can build these into the notion of "accepts" used to formulate the disquotational principle without affecting anything in the argument which follows.

Tyler Burge, among others, has made prominent the idea that this disquotational principle can pose a problem for certain accounts of mental content.[3] To see how this poses a problem for the individualist, recall our statement of the canonical form for an individualist functionalist account of belief:

[F] $\Box\ \forall a\forall p$ (a believes $p \equiv \exists x$ (x is a belief state of a & x has the content p for a))

The problem arises from the fact that the disquotational principle and [F] jointly entail

$\Box\forall a\forall S\forall p$ ((a accepts S & S means p) $\rightarrow \exists x$ (x is a belief state of a & x has the content p for a))

But, when we consider ways of filling in the '___has the content ___for ___' relation alongside ordinary examples of language use, this claim quickly comes to seem absurd.

Above, we've discussed several different ways of defining the 'has the content' relation. But on none of these is it plausible to say that whenever an agent accepts a sentence which means p, that agent is in a state such that this relation holds between the agent, that state, and p. Consider, for example, Stalnaker's indication theory. There the relevant relation was, roughly, counterfactual dependence, under optimal conditions, of a state of an agent on a certain fact. A simplified version of that theory may be stated as follows:

[1] Here I simplify by ignoring context-sensitivity, and ignoring the need to require that the agent in question understand the sentence, and accept it sincerely and reflectively. These may be understood as built into the notion of 'accepts' in play here and in what follows. I discuss this more in the next chapter, along with the question of how a communitarian use of this principle relates to animal beliefs.

[2] This kind of confidence in the disquotational principle in normal cases is consistent with thinking that it might need to be limited in some ways, or that the notion of 'understanding' required to make it true might require some refinement. I have in mind cases in which it is plausible to think that a speaker, using a sentence containing indexicals, is able to express a thought which he is not able to grasp, despite his, in some sense, understanding the sentence.

[3] See Burge (1979).

> a believes $p \equiv_{df} \exists x$ (x is a belief state of a & (($\mathrm{O}a$ & a is in x) $\Box\!\!\rightarrow$ (a is in x because p)))

Along with the disquotational principle above, Stalnaker's theory then entails the following:

> $\Box \forall a \forall S \forall p$ ((a accepts S & S means p) $\rightarrow \exists x$ (x is a belief state of a & (($\mathrm{O}a$ & a is in x) $\Box\!\!\rightarrow$ (a is in x because p))))

In other words, the proponent of the indication theory is committed to thinking that every time an agent accepts a sentence which means p, that agent is in an internal state which, under optimal conditions, he would be in only because of p.

But reflection on ordinary cases of language use shows that this claim is very implausible. Take, for example, the sentence, "The 14th president of the United States was Franklin Pierce." Is it the case that, in order to accept this sentence, an agent must be in a state which, under optimal conditions, he would be in only because of the fact that (in the optimal conditions world under consideration) the 14th president of the United States was Franklin Pierce? The state in question can't be a disposition to accept a sentence with a certain meaning; the point of mentalist accounts of belief is to give account of what it is for an agent to believe p without appealing to facts about the meanings of public language expressions.

Aside from this disposition, it seems to me very unlikely that there are any interesting similarities at all between the internal states of various competent speakers of English who accept this sentence. Consider, for example, the following example:

> Bob knows very little about the American political system; indeed he has many false beliefs about the office of the presidency. He thinks that "President" is a hereditary title; he knows that the president has significant political power, but is at a loss to say much about what this power is. He has heard the name "Franklin Pierce" before, but always thought (falsely) that Franklin Pierce was a prominent nineteenth-century baseball player. Then one day a trustworthy friend who, he takes it, knows more about politics than him, tells him, "The 14th president of the United States was Franklin Pierce." Bob reflects a bit on this new information; his friend has always told him the truth, so far as he knows, and certainly seems to be speaking seriously on this occasion. So he endorses the sentence. It seems that Bob thereby forms several new beliefs. He now believes that the 14th president of the United States was Franklin Pierce; that one of the former presidents of the United States was a prominent baseball player; and so on.

Is Bob now in an internal physical state which is such that, in the nearest world in which he is in that state under optimal conditions, he is in that state because, in that world, the 14th president of the United States was Franklin Pierce? Two kinds of arguments indicate that he need not be in such a state.

First, the example of Bob shows that, given the amount of mistaken beliefs plausibly compatible with being counted as understanding and accepting a sentence with its usual meaning, there may be very few interesting similarities between the internal states of the various agents disposed to accept some sentence of their language. Nevertheless, since they are all disposed to accept this sentence, they all have the same belief. Given the fact that they have so little in common other than their acceptance of this sentence, it seems odd to

try to explain the sameness of their beliefs by trying to find some similarity in properties of their internal states; indeed, it seems mere fancy to claim that each *must* be in some internal state with a certain second-order property. Rather, the natural explanation of their shared beliefs is that each is disposed to accept a sentence of their public language which has the same meaning for each. But this appeal to public language meaning is just what the mentalist picture was meant to avoid.[4]

A second way to make this point may be brought out by considering Bob*, an intrinsic duplicate of Bob who lives in a linguistic community identical to Bob's but for the fact that, in his community, the predicate "president" expresses a property coextensive with what we would express with the disjunctive predicate "president or vice-president."[5] Both Bob and Bob* accept the same sentence but, intuitively, acquire different beliefs by so doing. The problem for the causal-pragmatic theory is in accounting for this difference between the beliefs of Bob and his intrinsic duplicate. For, since Bob and Bob* are in the same physical state and have the same belief-forming mechanisms, it seems that the nearest world in which Bob is in optimal conditions will be *the same world* as the nearest world in which Bob* is in optimal conditions.[6] But, if this is so, then it follows from the causal-pragmatic account that Bob and Bob* have the same beliefs; and this runs counter to the intuition that the difference in the meanings of the sentences they accept is sufficient to give them different beliefs.

Nor is this point limited to terms of a special and highly restricted class, or to language users who are unusually ignorant. It is not hard to come up with an example of an expression which an agent understands, but with respect to which they are in much the same position as Bob in the example above. Consider, for example, the word "tracking." If you have used a VCR, you know that sometimes the screen will begin to flip, or show other kinds of interference. You may also know that, when this happens, it sometimes helps to press the button labelled "tracking" on one's remote control or VCR. I submit that, if you are in this position, you have beliefs about tracking; you might believe, for example, that the tracking sometimes goes off on your VCR, and that pressing such and such a button can help to fix the tracking. But, if you are like me, you can say virtually nothing about what tracking is. So, if you are like me, you are in much the same position with respect to "tracking" as Bob was with respect to "president."

Nothing is special here about the indication theory of content (though it is convenient to use, since it is more clearly statable than most theories of content). Similar points can be made about other individualist theories of content we've considered. According to informational theories, Bob acquires the belief that that Franklin Pierce was the 14th president of the United States by virtue of coming to be in a belief state, one of whose constituents is a

[4]Another way to dramatize this point is to imagine Bob before he acquired the relevant belief, and hence before he was in an internal state such that, were he in optimal conditions, he would be in that state only because Franklin Pierce was the 14th president of the United States. When Bob accepts the sentence, he acquires this belief; must he, by accepting this sentence, also come to be in a new internal state with this peculiar property? It seems unlikely.

[5]This is an extension of the well-known thought-experiments of Burge (1979).

[6]One might deny this, on the grounds that difference in the meanings of expressions of their respective languages might lead to the nearest world in which Bob is in optimal conditions being distinct from the nearest world in which Bob* is in optimal conditions. But to think this is to build facts about linguistic meaning into the foundational account of belief; and this is just what the mentalist denies.

mental representation which has the property of being president as its content. Hence, on one formulation, the informational theorist is committed to this representation having the following characteristic: under ideal conditions, Bob would token this mental representation only because of an instantiation of the property of being president. But again, consider how different Bob would be under optimal conditions; he would have many more beliefs about the presidency, and would not have many of those that he actually has. By virtue of what are we licensed to assume that, under these conditions, Bob would ever token the mental representation which figured in his current belief state, let alone that he would only token this representation because of an instantiation of the property of being president? It is easy to extend the point to functional role, conceptual role, and other informational theories.

So systematic connections between language use and belief do pose a serious challenge to individualist accounts of belief. But, one might still want to ask, what is going on here? How *could* something like language use cause facts about beliefs of agents to float free of facts about the properties of internal states of agents?

The answer is, I think, to be found in a fact about linguistic competence, which, in the recent literature, was first pointed out by Kripke in *Naming & Necessity*. The core point is that very little is required for an agent to be a competent user of an expression, and hence very little is required for an agent to be in a position to acquire new beliefs involving the content of the expression by accepting sentences in which the expression figures. Once you notice that all that is required for understanding an expression is satisfaction of minimal communal standards of use, it is not surprising that the class of speakers of a language who are competent with a given expression might not share any interesting similarities apart from their use of a shared language. But these speakers are, by virtue of their understanding this expression, in a position to acquire beliefs in which the content of that expression figures; hence, one might also find it unsurprising that the various agents who share a given belief might share no properties of a sort which can be exploited by an individualist to explain the fact that they all believe p.

The importance of this point about competence with expressions of a public language is, I think, one of the most important lessons of Kripke's discussion of the meanings of proper names in *Naming & Necessity*. One of the ways to view the import of Kripke's many examples of speakers who are not possessed of uniquely identifying information regarding the referent of a name — but who are still clearly able to use the name to refer to its usual referent[7] — is to see these examples as showing one of the characteristic faults of descriptivism to be its overestimation of the knowledge required for speakers to be competent users of the name. The present point is just a generalization of Kripke's claims about proper names to public language expressions more generally. This generalization shows one of the characteristic faults of individualism to be its overestimation of the knowledge required for speakers to be competent users of expressions of any linguistic category.

If confronted with this argument, a sophisticated individualist is not likely to simply roll over and give up. Rather, she will likely grant the point that the existence of the kind of 'language-dependent' beliefs we have been discussing requires a special clause in an individualist's theory of belief.

[7]See, for examples, the discussions of Schmidt and Gödel, and Peano and Dedekind, in *Naming & Necessity*, 83-85.

7.2.2 The appeal to deference

This defensive maneuver will likely take the following form: cases like these are special cases; in *these* cases, the contents of the beliefs of agents do, in some sense, depend on facts about the meanings of words. But these are exceptions; and, in these exceptional cases, there are mechanisms — which can themselves be explicated in terms of facts about mental content — which explain how agents come to have beliefs with the relevant contents. To make this move is to appeal to what is sometimes called *deference*, or *the division of linguistic labor*.

As Mark Greenberg has pointed out, the underlying idea is that it is the job of an account of belief or thought more generally to say what it is for an agent to have full mastery of a concept, or full grasp of a proposition.[8] We should, on this view, distinguish between cases in which an agent believes a proposition of which she has full grasp, and cases like that of Bob in which we can truly attribute belief in a proposition to an agent even though his grasp of that proposition is only partial. In turn, our account of what it is for an agent to have a given belief — i.e., our account of the facts in virtue of which we can truly attribute certain beliefs to agents — should be disjunctive, to reflect this distinction. It will then be the job of a theory like the causal-pragmatic theory to give an account of the first class of beliefs — those in which an agent has full grasp of the proposition believed — and the job of a theory of deference, or the division of linguistic labor, to give an account of the second class.[9]

This may initially seem like a promising strategy. Above, I argued that no individualist theory of content gives necessary conditions for an agent's mental representation having a given content, because of the examples issuing from the disquotational principle. But, one might conjecture, even if this result is accepted, we might be able to give a theory of content for mental representations which delivers the result that, for any agent A with a mental representation with content F, the theory can either explain that fact directly in terms of facts about A's internal states, or indirectly in terms of the internal states of some other agent to whom A bears some important relation. In the latter cases, the theory of deference could then come in to explain, by spelling out the relevant relation, why A's mental representation inherits the content F from a contentful representation of some other agent. In its essence, then, the appeal to deference is a way of trying to mimic the results delivered by the communitarian view that public languages can be vehicles of thought while restricting oneself to the resources of individualism.

To give a *theory* of deference is to explain how the contents of the mental representations of one thinker come to be inherited by another thinker. What is it about thinkers that makes it the case that some times, but not other times, the contents of a thinker's thoughts are fixed not by some individualist theory of content, but by the contents of the thoughts of another?

Deferential intentions

Often, what the deference-theorist has in mind is some kind of deferential intention on the part of a language-user;[10] this is the first form in which I'll consider the idea. The core idea

[8] See especially Greenberg (in preparation).

[9] This is roughly the position of Peacocke (1992), 27-33, though Peacocke focuses on possession conditions for concepts rather than conditions for full grasp of propositions.

[10] See, e.g., Loar (1991), 121-122.

here is that one individual's representations inherit the content of the another's just in case the former has certain intentions concerning the latter.

The first question facing the proponent of deferential intentions is: what are these deferential intentions like? An initial thought is that the intentions should concern the mental representations of the agent in question; if we could somehow affix the contents of those representations to the meanings of words in the public language — like "president" in the example above — then the mentalist might be in a position to give account of the contents of the beliefs of that agent in terms of the contents of those mental representations. Perhaps, then, the mentalist should say that the shape Bob's deference takes in the case above is that he intends that one of his mental representations has the same content as the English word "president."

It is hard to take seriously the idea that most people have intentions involving their mental representations at all, let alone in every case in which they acquire a belief by accepting a sentence. But, more importantly, even if this move did work it would be no great help to the mentalist. Right now the intention under consideration is the intention that a mental representation mean whatever some English word means; but this is just a roundabout way of saying that while the contents of beliefs are fixed by the contents of mental representations, the contents of the latter are determined by the meanings of expressions in the public language of the agent in question. And of course this is what the individualist is using deference to try to avoid.

Perhaps, then, we should switch the intention to make it the intention that, to use our example above, one of one's mental representations has the property of being president as its content. Here we appeal to the property which is the content of the English word directly, without relying on the fact that it is the meaning of this English word, and thus avoid appealing to any facts about the meanings of expressions in English.

But recall what we were trying to use these deferential intentions to do: explain Bob's ability to have thoughts involving the property of being president when one's theory of content seems to entail that none of his mental representations could have this property as their content. Now we're considering explaining this ability in terms of Bob's ability to have certain *intentions* in which the property of being president figures. But this is surely just to push the bump in the rug. This move assumes what it is trying to explain: the ability of Bob to have mental states in the content of which the property of being president figures.

This also points to a more general problem with the appeal to deference. We are trying to give an account of the contents of beliefs; but now we find ourselves appealing to facts about the contents of intentions in order to complete the story. But, if one doubts that a mentalist account of the contents of intentions which does not appeal to facts about belief is forthcoming, then this appeal to intentions does not seem to be a serious advance for the mentalist cause.

It is also worth noting that the very idea of this sort of intention-based explanation of mental content is a bit puzzling. Suppose that I desire a better understanding of Frege; I thereby form the intention that my beliefs about Frege have whatever content those of Paul Benacerraf have. Still, intending doesn't make it so.[11] The same goes in the case of words; my intending to defer to others in the use of some expression is not sufficient for me to understand

[11]For a similar point, see Fodor's response to Brian Loar in Loewer & Rey (1991).

the expression, or be able to have thoughts involving its content.[12]

Deference without intentions

Gilbert Harman has presented a different sort of account, from an individualist perspective, of how facts about mental content can depend on facts about the meanings of words as used by others; one of the virtues of this account is that it avoids the use of intentions which proved problematic above.

Harman, following Putnam, asks us to imagine a case in which you "cannot distinguish oak trees from elm trees and do not know any of the distinguishing properties of these two sorts of trees."[13] You do, however, speak English; and you do use the words "oak" and "elm." This is the sort of case which poses problems for mentalists, since, given that you cannot distinguish oaks from elms, it seems that there will be no two mental representations of yours which will, independently of their attachment to English words which you know how to use, have properties which can account for their having oaks and elms in their respective extensions. The task, then, is to explain how one of your mental representations might have come to have the property of being an oak tree (rather than the property of being an elm tree) as its content, and to do so without appealing to facts about the meaning of "oak" in English.[14]

Call the mental representation in question "μ." On Harman's view, μ comes to have the property of being an oak tree as its content by a fourfold process:

[12]Another sort of account, which should be kept distinct from the appeal to deferential intentions, is the sort of appeal to meta-linguistic beliefs endorsed in Loar (1981) (159-160). According to this sort of account, "gold" means "the kind called 'gold' by our experts," where this is a description which takes wide scope over modal operators. The idea then seems to be that we can give, in our example above, an account of the fact that Bob believes that Franklin Pierce was of the kind called "president" by our experts; and, if the claim about the meaning of "gold" given above is true, this is sufficient to give an account of the fact that Bob believes that Franklin Pierce was president. But there seem to me many reasons to distrust this claim of about the meaning of "gold." The most basic is that "is gold" and "is of the kind called 'gold' by our experts" do not seem to be substitutable in attitude contexts. One may believe that one's wedding ring is gold without having any beliefs about experts, or any meta-linguistic beliefs at all.

In addition, some of the arguments advanced against descriptivist attempts to capture intuitions about rigidity in Soames (2002), ch. 2 apply to Loar's account. One of these, in particular, is relevant. On Loar's account the proposition that gold is a metal is identical to the proposition that the kind called "gold" by our experts is a metal; hence if one believes the former, one believes the latter. The latter proposition makes essential reference to experts in the actual world. The problem with this is that it seems clear that possible agents can believe that gold is a metal without having any beliefs at all about the experts in the actual world. (This is the analog of one of Soames's arguments against descriptions rigidified using the "actuality" operator.)

[13]Harman (1987), 219.

[14]Harman uses the term "concepts" rather than the term "mental representations"; but it seems clear from the context that what he means by the former is much the same as what I have meant by the latter.

1 There is some other person B who has some mental representation μ^* which has the property of being an oak tree as its content in virtue of facts about μ^*'s conceptual role in B's psychology.[15]

2 The word "oak," as used by B, means oak because of its connection with μ^*, which has the same content.

3 The word "oak," as used by you, means oak by virtue of its connection with the same word as used by B.

4 Your mental representation μ comes to have the property of being an oak tree as its content by virtue of its connections with "oak," as used by you.[16]

From an individualist perspective, the virtues of the account are clear: it gives an account of the cases in question by appealing to dependence on the contents of the mental representations of others, rather than on facts about public language meaning; and it does so without any appeal to problematic intentions.

There is, however, something puzzling about the account. In step (3), the word "oak," as used by you, derives its content from "oak," as used by B; and this happens because of some connections between the word as used by you and as used by B. The first thing that needs explaining is: Why does your word derive its content from his, rather than the other way around? So far we have a word as used by you, and as used by your interlocutor, and the two bear some connection to each other; what we don't have is an explanation of why the transmission of meaning goes in one direction, rather than in the other. In order to explain this, one wants to appeal to the meaning of "oak" in English: "oak" as used by you acquires its meaning from "oak" as used by B because on the latter use, "oak" means what it means in English, the public language that both you and B share. But this is the sort of appeal to facts about public language meaning that the individualist is out to avoid. One might instead appeal to your intentions to mean the same as B; but then we're back in the camp of the intention-theorist, which was dismissed above. So we need some explanation of the direction of meaning-determination in this sort of conflict of idiolects; and it seems to me unlikely that any which does not appeal to public language meanings is available.

One might try to solve the problem by explaining the direction of meaning-transmission in terms of the behavioral dispositions of the relevant agents; perhaps one might focus on their dispositions to accept and receive correction from each other, or from other agents in their community. But an example from Tyler Burge seems to count against any such attempt. Burge asks us to imagine a case in which an agent develops a nonstandard theory about some kind of thing; for example, she might come to believe that sofas are not pieces of furniture intended to be sat upon, but rather works of art or religious artifacts.[17] We would attribute thoughts involving the concept of a sofa to such a person; we might say, after all, that she thinks that sofas are religious artifacts. But because she believes that others in her community are incorrect in their views about the nature of sofas, she will not be disposed to defer to their claims about sofas, nor be particularly disposed to accept correction on the matter of sofas from other language users. These cases pose no problem for the communitarian; but it

[15]The use of conceptual role here is not mandatory; the mentalist can substitute in her favorite account of the contents of mental representations.

[17]The example is from Burge (1986), though he puts it to a different use than I do here.

is not obvious how the proponent of deference without intentions might account for them.[18]

7.2.3 Why deference is a red herring

This shows that the task of providing a theory of deference is a difficult one. An individualist might respond to these objections to particular theories of deference by espousing a common view of deference which one might call the 'whatever it takes' view. On this view, we do not know exactly how deference works; but we do know that *something* is making a difference in cases like that of Bob and the concept of a president. Deference is just whatever this extra something is.

This is not a very comfortable resting place for the individualist. After all, the explanation in these cases might be that the relevant agents are members of linguistic communities in which certain expressions have meanings, and that the agents understand some of these expressions, and so are in a position to form beliefs in whose contents the meanings of the understood expressions figure. In other words, it might be that what it takes to explain these cases is a communitarian view of thought and language. We have already seen some reasons to endorse this conclusion; but there are further reasons for doubting that deference will be the answer to all the individualist's problems.

A logical problem with the appeal to deference

First, note that deference is only to be a useful tool for a theory which already gives sufficient but not necessary conditions for a representation to have a given content. Deference, after all, just provides a new way for representations to have a content; this provides no help to a theory which, by failing to yield sufficient conditions for representations having certain contents, is committed to incorrect content assignments. But the individualist views we have considered were all of this sort.[19]

Inheritance without deference

There is also a more basic reason to doubt the mentalist's division of the terrain into cases of full grasp and cases of partial grasp supplemented by deference: as Mark Greenberg has

[18]The account also requires that you come into contact with some such person as B, who has a mental representation with the requisite content independently of any facts to do with his public language. But surely this is not necessary; you can learn the word "oak," and thereby come to have beliefs about oak trees, without meeting this sort of expert. One might then claim that there is some causal chain leading back from your interlocutor to such an expert, whether or not it is a very short one. But why think that this must be the case? It seems sufficient that "oak" mean oak in English, whether or not you bear any causal connection to a privileged class of users of the expression. For some cases in which the causal chain is plausibly missing, see McKeown-Green (2002). Further problems stem from cases discussed by Mark Greenberg, in which there is no one member of the group that has full mastery of the relevant concept. See Greenberg (in preparation) and below for more discussion.

[19]This is not a problem for the deference-theorist who only wants to explain away some extra beliefs whose existence is not entailed by his theory of belief. The point is aimed at a theorist who thinks that the arguments of the preceding chapters were inconclusive because not directed at the conjunction of individualist theories of content with accounts of deference.

emphasized, there seem to be cases in which we are willing to attribute beliefs to agents which do not fit into either category.[20] Which cases these are depend upon what one's theory of deference is; but there are at least a few which seem to have broad generality.

One of these was already discussed above: in Burge's example of the agent with a non-standard theory of sofas, it is hard to see that the agent does anything which might be characterized as deference. Another case, which is in some ways more fundamental, has been developed by Greenberg. Greenberg notes that we can imagine cases in which there is no one person in a community with full mastery of a concept, but rather several different 'experts' who master various aspects of the concept. This case strikes at the heart of the appeal to deference, since the usefulness of deference is its ability to recast superficially communitarian facts about the thoughts of agents in terms of relations between individual agents. But it is hard to see how one could get an account of this sort off the ground if there is no one agent who masters the concept in the first place.[21]

Problems with the notion of full mastery

Perhaps the most fundamental problem with the appeal to deference, though, has to do with its motivations. We saw above that the idea of the individualist who appeals to deference is that we can solve cases of 'language dependent' beliefs by hiving them off from other cases of belief. On this view, the class of beliefs divides into central cases in which the agent in question has 'full grasp' of the proposition believed, and (theoretically) peripheral cases in which the agent acquires beliefs by some mechanism of inheriting the thought contents of others.

But do we have any pre-theoretic grasp of this notion of full mastery, or any reason to think that it should be a theoretically useful notion? Among the agents to whom we may truly attribute beliefs involving the concept expressed by "president" or "tracking," there is a continuum of knowledge, from very little to a great deal, about the nature of this concept. Surely if we are interested in the nature of belief, the fundamental distinction is not between those agents who fall on one or the other side of some line drawn in this continuum, but rather between those agents to whom we can truly attribute thoughts involving a concept and those to whom we cannot. It is only if one has already adopted a view of thought as essentially prior to and independent of public language that it seems as though there *must* be some such distinction.[22]

Further, there is really no sense in which 'language dependent' beliefs are exceptional or atypical. To be sure, it may have seemed as though these cases were atypical, because in contemporary philosophy of mind the main use of the disquotational principle has been in arguments for externalism: the thesis that the contents of an agent's beliefs do not supervene on her intrinsic properties.[23] To come up with clear cases in support of externalism, one must

[20]For an excellent discussion of issues involving the appeal to deference, see Greenberg (in preparation).

[21]For a discussion of other such cases, see Greenberg (in preparation).

[22]This is not, of course, to deny that one can define a coherent notion of full grasp of a content within one's theory. (See Greenberg (in preparation) for a few ways of doing this.) The worry is just that there may be no principled way of doing so.

[23]This is, I take it, the main target of Burge (1979).

imagine fairly recondite scenarios, simply because one must hold fixed *every* intrinsic property of the agent in question. But the present issue is not about externalism; most individualists are externalists. Rather, the issue is whether public languages are vehicles of thought. To provide cases which count in favor of this thesis, one need only imagine scenarios in which the beliefs of agents vary while we hold fixed the facts claimed to be relevant to the determination of mental content by some particular individualist theory of content. I hope that the foregoing has made it clear that cases like this are not so hard to find.

Conclusion

So we should conclude that the individualist's attempt to mimic communitarianism by appeal to deference is unsuccessful. Rather, we should take the disquotational principle at face value: it shows that one of the ways of having a belief is accepting a meaningful sentence of one's public language, and so shows that public languages should be a part of our constitutive account of mental content.

This conclusion is further strengthened by the arguments of Chapters 4 and 5 above. It is not as though individualist theories of content handle the basic cases well, and that this success motivates the search for a theory of deference which will allow us to preserve these results. Rather, each of the individualist theories of content we considered failed for many seemingly basic cases. Viewed in this context, the appeal to deference is a failed attempt to salvage a failed research program.

7.3 Beliefs, inner states, & behavior

Though not so explicitly a part of the individualist program as the denial that beliefs are constituted by social facts about public languages, many individualists have also thought of beliefs as internal states of agents.[24] In this chapter, I present some arguments against the thesis that beliefs are inner states; in the next chapter, I shall present a view of belief which rejects this thesis.

7.3.1 Functionalist accounts of belief states

One sort of argument against the claim that beliefs are inner states was already provided by the foregoing: if beliefs were inner states, then the contents of beliefs would have to be determined by the properties of those internal states.[25] But, as we've seen, there are no properties of the internal states of an agent which can account for the contents of that agent's beliefs; hence beliefs are not internal states.

To this sort of argument, the mentalist has a reply. Above I separated the tasks of a broadly functionalist theory of belief into two: saying what it is for an internal state to be a

[24] See §4.1 above for some of the reasons.

Throughout this section I rely on an intuitive distinction between internal states of agents and dispositions of agents. I'm not sure how far the distinction can be pressed; the point is just to draw attention to the distinction between views which take the properties of believers relevant to their mental states to be facts like their brain states, and those which take the relevant properties to be associated exclusively with the actions of agents.

[25] Of course, as above, these properties are not restricted to either intrinsic or 'solipsistic' properties of these internal states.

belief state, and saying what it is for a belief state to have a certain content. An individualist might reply that, while we can give no substantive account of what it is for an internal state to have a certain content, we can give an account of what makes certain internal states, but not others, belief states.

So what might make certain internal states of agents belief states? The only going option seems to be a kind of functional analysis: that certain states are belief states because of the way that they are causally related to the perceptual inputs and behavioral outputs of agents. Consider one schematic idea of how this might go:

> If a state x has content p then x is a belief state iff x is typically caused by perceptions with content p and typically causes actions which would satisfy the agent's desires if p were the case.[26]

No doubt, this account is oversimplified. But the important point is that this is simply an account of the wrong form. The individualist under consideration has already granted that there is no substantive account of mental content to be had; the problem is that the present account of what it is for an internal state to be a belief state presupposes an account of mental content.

So we need a functionalist account of belief states which does not presuppose facts about the contents of those states. But it is difficult to see how such an account might go; so far as I know, the task has never seriously been attempted. One might be inclined to try something like

> x is a belief state iff x is the kind of state which is typically caused by perceptions and typically plays a role in causing actions

but the idea that we could so delimit the kinds of causal relations involved that we would get the right results seems, at this stage, to be overly hopeful.[27]

So against the thesis that belief are internal states we have the following argument: (1) If beliefs are internal states, then there must be some account of what makes certain internal states belief states. (2) If there is no account of the contents of belief states, then there is no account of what makes certain internal states belief states. (3) There is no account of the contents of belief states; hence (4) there is no account of what makes certain internal states belief states. Therefore, (5) beliefs are not internal states.

A proponent of the claim that beliefs are internal states might respond to this argument by denying (3). He might grant that there is no individualist theory of the contents of internal states, but that there might yet be a communitarian account of the contents of such states. Moreover, he might continue, there is simply no alternative to the idea that beliefs are internal states; the only alternative, after all, is a kind of behaviorism which was discredited several decades ago.

---

[26] This is taken from Stalnaker (1984), 15-19, though this is not exactly Stalnaker's view.

[27] This point also holds, to some extent, against functionalist accounts of belief states like the one of the preceding paragraph which make use of facts about the contents of internal states. The point is only that disallowing this use of facts about content makes matters worse. See Stalnaker (1984) for a nice discussion of the problems in giving a functionalist account of belief without an account of content; he discusses this kind of account under the heading of the "purely pragmatic" theory of belief.

This second point requires an extended answer. In the next section, I argue that the objections typically taken to discredit behaviorism fail, and in the next chapter argue that we can give a plausible neo-behaviorist account of some mental states. So let's focus for now on the claim that we can save the idea that beliefs are internal states by giving a communitarian account of the contents of those states.

The natural way to construct such an account might exploit the (alleged) necessary truth that dispositions always have categorical bases which are internal states of the bearer of the dispositional property. One might then say: let us grant that dispositions to accept sentences of one's public language can be constitutive of having certain beliefs. We can then just say that the internal states which are the categorical bases of these dispositions are belief states with the content of the sentence accepted.

One can indeed say this; but there is some disingenuous in the claim. These internal states are relevant only insofar as they are the bases of the relevant dispositions to action; the latter are doing all the work. To mount this sort of defense of the claim that beliefs are internal states is a bit like claiming that actions are internal states, since actions always have a causal basis in the internal states of agents.

7.3.2 Why behaviorism fell out of fashion

But we have yet to respond to the objection that, since behaviorism is false, the claim that beliefs are internal states is the only game in town.

Largely under the influence of Wittgenstein's *Philosophical Investigations* and Ryle's *The Concept of Mind*, accounts of the nature of various sorts of mental states in terms of the behavioral dispositions of agents occupied center stage in the philosophy of mind of the 1950's and early 1960's. Since then, however, the idea that mental states are constituted by dispositions to action has increasingly come to be regarded, as the result of a series of arguments, as an unfortunate downturn in the history of Anglo-American philosophy and psychology. This change was not only because of the simple argument that in many cases there is no (non-linguistic) behavior which could plausibly be taken to be constitutive of many mental states. Rather, there have been a number of related arguments which together led to abandonment of the behaviorist program.

I think that, in the case of some mental states, behaviorism is quite a plausible option; in the next chapter, I shall argue that such a treatment of belief is defensible. To clear the way for this discussion, we will need to respond to the historically influential objections to behaviorism, among which are the following:

[1] There are, in general, no translations from talk about mental states to talk about dispositions to behavior; hence beliefs cannot be constituted by dispositions to behavior.

[2] Any list of an agent's dispositions is logically compatible with an indefinite number of ascriptions of beliefs and desires to that agent; for example, the disposition to ϕ is compatible with desiring p and believing that ϕing will realize p, desiring q and believing that ϕing will realize q, and so on. Hence the beliefs of an agent are not determined by their behavior alone.

[3] Talk about mental states is not talk about behavior; it is, if anything, talk about what is the source of that behavior. Hence belief cannot be constituted by behavior.

[4] "... each individual X-worlder may *think* to himself: 'This pain is intolerable. If it goes on one minute longer I shall scream. Oh no! I musn't do that! That would disgrace my whole family ...' But X-worlders do not even admit to *having* pains. They pretend not to know either the word or the phenomenon to which it refers. ... Only, of course, they do have pains, and they know perfectly well that they have pains. If this last fantasy is not, in some disguised way, self-contradictory, then logical behaviorism is simply a mistake."[28]

[5] It is bad scientific method to limit psychologists to overt behavior in the study of mental phenomena. If we limit psychological laws to stimulus-response correlations — rather than also taking into account facts about the internal states of organisms — then we must either make the notions of stimulus and response so broad as to lack any explanatory force, or restrict them so far as to divest them of any actual application.[29]

Given assumptions made by mid-century behaviorists, most of these may be turned into good arguments; that is, given the state of philosophy at the time when the debate between behaviorists and their opponents took place, their opponents had the better arguments. But each of the relevant assumptions should be rejected.

One of these assumptions was an identification of the notions of necessity, analyticity, and logical truth. It is widely taken to be a moral of Kripke's *Naming & Necessity* that these notions are distinct, and importantly so; and this is enough to block several of the above arguments. For example, the premise of [1] is surely correct: there are no meaning-preserving translations between any claims about mental states and any claims about the behavior of agents. On some understandings of analyticity, this might be enough to show that there are no analytic entailments either from claims about mental states to claims about behavior, or from claims about behavior to claims about mental states. But, once we dissociate analytic entailment from necessary consequence, this argument simply does not show that there are no necessary relations between mental facts and facts about behavior. Hence it does not show there are no constitutive relations between these two sorts of facts.

A related point shows one place where [2] goes wrong. Intuitively, one class of facts determines another if the former necessarily fix the latter. The argument of [2] moves from the obvious claim that facts about belief are not logical consequences of facts about behavior to the claim that facts about behavior do not determine the beliefs of agents; but this argument is only valid if one assumes that there are no necessary consequences which are not logical consequences.[30]

[28] This is Putnam's example of the super-super-spartans, from Putnam (1968), 155.

[29] This is one of the influential criticisms of B.F. Skinner's *Verbal Behavior* to be found in Chomsky's 1959 review.

[30] Actually, [2] embodies a worse mistake as well. Many opponents of behaviorism were also proponents of some sort of functionalism; but an argument like [2] may be run just as easily against

By now the reply to [3] should be clear. The proponent of the view that beliefs are constituted by dispositions to behavior should not regard her account of belief as revealing analytic truths about the concept of belief, and a fortiori should not regard her account as saying "what people meant by belief talk all along." Rather, she should take her account as stating necessary truths about belief which together say what it is for an agent to have a certain belief. This point does not, however, divest [3] of all force; it is certainly true that we sometimes explain the actions of agents in terms of their beliefs, and the proponent of any account of belief must say how beliefs, on her picture, can be the sorts of things which explain action. Part of the answer is that the behaviorist should not take beliefs to be constituted by actions, but by dispositions to action; the prospects of explaining behavior in terms of dispositional properties are not so hopeless as the prospects of explaining behavior in terms of that very behavior. But the challenge of making sense of belief-desire explanations of behavior is one of the main challenges facing the kind of account developed below; I return to it briefly toward the end of the essay.

[4] goes wrong in two ways. The first was just mentioned above; a behaviorist should take facts about belief to be constituted, not by behavior, but by dispositions to behavior. The characters in Putnam's example exhibit no overt pain behavior; but it is not clear that they have no dispositions to perform such behavior. Indeed, their having such dispositions seems to be one of the key sources of the intuition that they really do feel pains. The second point is more obvious: just because one class of mental phenomena is constituted by dispositions to action, it is not obvious that *every* mental phenomenon should be so constituted. Indeed, it seems clear that feeling a pain or an itch is a very different sort of thing than having a belief or a thought; and both are very different than seeing something. So, while behaviorists at the time tended to be behaviorists about all mental properties, intuitive differences between different classes of mental phenomena make this an odd commitment. For, the point about dispositions aside, there is some force to Putnam's example; it does seem very unlikely that pains should be constituted by dispositions to action. But this does not count against the corresponding point about belief.[31]

functionalism as against behaviorism: "Any ascription of functional states to an agent is logically compatible with an indefinite number of ascriptions of beliefs and desires to that agent; hence the beliefs of agents are not fixed by their functional states alone." The functionalist should reply by granting the point, and saying that the functional states of an agent *along with the true functionalist theory* is logically compatible only with the true ascriptions of beliefs and desires to that agent. But then it is not clear why the behaviorist should not make the same sort of reply. What the opponent of behaviorism should have said all along is that the behavioral dispositions of an agent along with some behaviorist theory T is compatible with an indefinite number of ascriptions of beliefs and desires to that agent; hence the beliefs of agents are not fixed by their dispositions in the manner stated by T. But now we have only an argument against a single behaviorist theory, and not a principled argument against the view that beliefs are constituted by dispositions to action. Usually, the theory opponents of behaviorism had in mind was something like the belief-desire-action principle [BDA], discussed on p. 170 below. And they were right that [BDA] is no good as a constitutive theory of belief; but, as I'll tried to show by developing a positive theory, this hardly discredits behaviorism more generally.

[31]The assumption that behaviorism should be a thesis about all mental states may explain the fact that, in his discussion, Putnam focuses on behavior rather than dispositions to behavior. For while it is natural to think of belief as a kind of dispositional state, this is far less natural in the case of pain.

Unlike the others, the Chomskyan objection [5] is correct as it stands: a science of behavior, which aims to formulate laws governing the behavior of organisms, cannot restrict itself to stimuli and responses. To see why this should pose no problem for the view that belief is constituted by behavior, it will be useful to consider briefly what sort of question we are asking when we ask, "What is it for an agent to believe p?" or "What constitutes an agent's believing p?" Intuitively, responses to questions like these are answerable to a strong modal constraint: that they apply not only to actual subjects, but also to merely possible agents capable of forming beliefs. This constraint is typical of the criteria for answering philosophical questions about the natures of things; the example of a doctor harvesting the organs of an innocent patient to save the lives of five others may be proposed as a counterexample to the view that maximization of happiness is constitutive of right action, whether or not any actual doctor has performed such an act.

The salient point as regards [5] is that this modal constraint distinguishes philosophical questions about the nature of belief from scientific questions about the laws governing behavior; in contrast with philosophical theses, the fact that a scientific claim holds only contingently is no bar to its being a law. Given this, a neo-behaviorist position in the philosophy of mind is quite consistent with an anti-behaviorist psychology. The true non-behaviorist psychological laws, if such there be, might hold only contingently, and so not qualify as constitutive accounts of mental phenomena. The true behaviorist accounts of the natures of mental states, if such there be, will hold necessarily, but need not preclude contingent scientific accounts of the causation of behavior in terms of causal interactions between the internal states of agents.

Conclusion

So far, I have argued against several facets of the individualist approach to the relationship between mind and language, as exemplified in the mentalist and private language pictures of intentionality. In this chapter, I have picked out the two assumptions which I take to underly the failure of individualism. It is worth noting that the rejection of these two assumptions — that thought is prior to language, and that mental states are constituted by internal states of individuals — are among the themes of the philosophy of mind and language of the later Wittgenstein. Regarding the priority of thought over language, Wittgenstein wrote

> When I think in language, there aren't 'meanings' going through my mind in addition to the verbal expressions: the language is itself the vehicle of thought.[32]

Here "the language" is a public language; Wittgenstein's view here is that public languages can be instruments of thought, in the sense that for an agent who understands a meaningful

It is an interesting aspect of Putnam's example that he regarded the "inner behavior" of the X-worlders as off limits to the behaviorist seeking to give an account of pain in terms of pain behavior. Part of the reason why Putnam was justified in assuming this was that behaviorism about mental facts was closely linked to verificationism; and, if one's behaviorist tendencies have verificationist motivations, inner behavior certainly is off limits. Such inner occurrences are no more verifiable to third party observers, after all, than the mental phenomena to be explained. But there is no reason for any sort of behaviorist to endorse verificationism. See p. 164 below for a discussion of the use of mental phenomena like inner speech in the explication of belief.

[32] *Philosophical Investigations*, §329.

sentence of her public language, there may be no more to say about what it is for that agent to entertain a thought with a certain content than that that agent does something with the sentence, and that that sentence has a certain meaning in her language.

I have also argued that we should not identify beliefs with inner states of agents. Regarding the identification of mental states with certain inner states of agents, Wittgenstein wrote

> No supposition seems to me more natural than that there is no process in the brain correlated with associating or with thinking; so that it would be impossible to read off thought-processes from brain processes. I mean this: if I talk or write there is, I assume, a system of impulses going out from my brain and correlated with my spoken or written thoughts. But why should the *system* continue further in the direction of the centre? Why should this order not proceed, so to speak, out of chaos? ...
>
> It is thus perfectly possible that certain psychological phenomena *cannot* be investigated physiologically, because physiologically nothing corresponds to them.[33]

Wittgenstein perhaps overstates his case here a bit; but the underlying idea is that, even if there are necessary connections between behavior ('my spoken or written thoughts') and mental states ('thought-processes'), and even if there are causal links between internal states and behavior, we need not (and should not) infer that a constitutive account of those mental states in terms of internal states is available.[34]

Beginning in the next chapter, we shall see how far these two Wittgensteinian points can take us in the direction of a more promising approach to the relationship between mind and language.

[33] *Zettel*, §§608-9.

[34] He might overstate his case in thinking that thoughts could emerge 'out of chaos'; the more cautious formulation is that because certain mental phenomena may be realized by any number of different internal mechanisms, a constitutive account of those phenomena ought not to be given in terms of such mechanisms.

Chapter 8

Belief As Constituted By Behavior

Contents

In the preceding chapter, I argued that the failure of individualism was due to the twin claims that various propositional attitudes were prior to and independent of public language, and that beliefs are inner states. I have also tried to eliminate the kinds of 'in principle' objections often raised against the idea of a constitutive account of mental states which denies these two assumptions.[1]

But this is not enough to show that such accounts will be plausible; to do that, we need to actually give an account of a mental state in these terms. The project of this chapter is to show how such an account of belief might run.

[1] See the discussion of the argument from language individuation against giving public languages an explanatory role in §6.4, and the discussion of the traditional arguments against behaviorism in §7.3.2.

8.1 The supervenience of belief on behavior

We can begin to clarify the sense in which beliefs might be constituted by dispositions to action by noting a test for any proposed constitutive account of belief. Given the modal constraints on constitutive accounts, if there is to be any hope of giving an account of belief in terms of dispositions to action, the beliefs of agents must at least supervene on their dispositions. Even if they do, it will remain a further question whether necessary conditions for belief can be given in terms of these dispositions; but if they do not, then we should scotch the view that behavior is constitutive of belief at the outset.

Do facts about the beliefs of agents supervene on their behavioral dispositions? It seems not. For consider two different agents who perform all the same bodily movements and, in particular, utter exactly the same sentences. However, one of the agents always intentionally utters these sentences, whereas the other often utters sentences spasmodically, without intending to do so. Plausibly, we could fill out the two cases in such a way that, intuitively, we would say that the two agents, despite performing the same bodily movements, had different beliefs. After all, even though the two agents utter precisely the same sentences, one is expressing beliefs while the other is not. This shows that, in giving an account of belief as constituted by behavior, the behavior relevant to the account must not be mere bodily movements — such as emitting certain sounds — but rather full-blooded intentional actions — such as accepting a sentence, or waving to someone.[2]

Do facts about the beliefs of agents supervene on their dispositions to action, as opposed to their dispositions to perform certain bodily movements? Again, it seems not. For, following Tyler Burge,[3] consider two agents in distinct linguistic communities who differ only in that the word "arthritis" means arthritis in the language of one, whereas the same word refers to a disease of the joints and thigh in the other. It seems that they have are disposed to perform just the same actions; but when the two utter the sentence "I have arthritis," they express different beliefs.

This shows that beliefs do not supervene on dispositions alone, but only on dispositions along with social facts about the meanings of words in the linguistic communities agents inhabit. The linguistic dispositions relevant to the nature of belief will not be merely dispositions to accept sentences, but dispositions to accept sentences with certain meanings in the language spoken by the agent. If we include such social facts in our description of the relevant dispositions, we can arrive at differing dispositions which promise to account for the differences in the beliefs of the two agents in the above scenario.[4]

It's worth noting that these two qualifications are not qualifications on or extensions of our ordinary practices of ascribing dispositions to agents; on the contrary, our ordinary ascriptions of dispositions to agents almost always make reference to actions rather than bodily movements of the agent in question, and sometimes make reference to facts about the

[2]Given that action-types are typically individuated partly in terms of the beliefs, desires, and intentions of agents, this will naturally raise the worry that an account of belief in terms of dispositions to action will be circular. I discuss this worry, and a few different reactions to it, later in this chapter and the next.

[3]Burge (1979).

[4]In one sense, of course, there is nothing special about sentences; performance of gestures in sign language may play the same role as acceptance of sentences, as can any action-type given a meaning in a community.

meanings of sentences. The point of these examples is only to show that, if behaviorism about belief is to be at all plausible, we must rely on ordinary descriptions of behavior rather than on descriptions restricted to a more austere vocabulary.

With these clarifications regarding the notion of a disposition at work here, do the beliefs of agents supervene on their dispositions? Very nearly; but the following example suggests that we are not quite there. Consider the case of A and B, who are two furry mammalian creatures which speak no public language, and are members of different species, but share all of their behavioral dispositions, in the above sense, in common.[5] Each is disposed, in particular, to lumber off each morning to the shore of a nearby river, stick its nose in the mud of the river bank, and inhale deeply through its snout before blowing the mud back out. They pursue this course of action, however, for different reasons. It is characteristic of members of A's species that they are able to ingest food through their snouts; when A pulls in through his snout, he quickly ingests some small insects which each morning lie embedded in the river bank he frequents. But members of B's species have chronic difficulties with bacteria in their snouts; hence they have learned to rid themselves of these bacteria by filling their snouts with mud before blowing the mud back out.

Insofar as we might be willing to ascribe mental states to A and B on the basis of this sparse description, I think that we are inclined to ascribe different mental states to them. Depending on how the rest of the story were filled out, it might be natural to say of A that he performs this morning ritual because he knows that there are insects in the river bed, and of B that he behaves this way because he knows that doing so will clear out his snout. This provides some indication that, at least in the case of some creatures, belief does not supervene on behavioral dispositions alone, but rather on behavioral dispositions together with facts about the ends or goals of the actions of those creatures.[6] As I shall argue in §8.4, the most plausible story about what it is for such a creature to have a belief will have to make use of such facts about their ends.

Such cases aside, though, it seems quite plausible that the beliefs of agents do supervene on their dispositions. I have no proof of this; but try to imagine two agents who have exactly the same dispositions, and yet differ in at least one belief. So far as I can see, this is not a conceivable situation; and this is at least a *prima facie* argument for its impossibility. The conclusion that beliefs supervene on dispositions to behavior gives some credence to the claim that what it is for an agent to have a certain belief is for the agent to be disposed to perform certain actions.

8.2 Accepting a sentence

To believe something is to take the world to be a certain way; that much is platitudinous. The thesis to be defended is then that to take the world to be a certain way is to be disposed to do certain things. There are two ways in which, by one's actions, one can take the world to be a certain way: one can be disposed to *accept a sentence* which expresses the claim that

[5] It is perhaps easier, as in the above cases, to think of them as inhabiting distinct possible worlds. One may also imagine them as superficially indistinguishable.

[6] This difference might also be described in terms of a difference in dispositions: one is disposed to go to the river bank to ingest insects through it's snout, whereas the other is disposed to go to the river bank to clear its snout.

the world is that way, or one can be disposed to *act on the basis of* the claim that the world is that way.

One way to take the world to be a certain way is to accept, or endorse, a sentence which says that the world is that way. In one sense, this is a rather obvious claim; but it has not gained much prominence in discussions of what it is for an agent to have a belief with a certain content because of reluctance to think that mental states like beliefs could be partially constituted by social facts about linguistic meaning. The key link between linguistic behavior, linguistic meaning, and belief is expressed by the following disquotational principle, discussed above:

> Necessarily, if an agent is disposed to accept a sentence which means p, then the agent believes p.[7]

This principle states one of the few non-trivial but relatively obvious necessary truths relating facts about an agent's beliefs to another class of facts. Anyone seeking to give an account of belief must either explain why this principle holds, or take the linguistic dispositions of language-using agents to be constitutive of their beliefs. By now, it should be clear that I think that the former course is indefensible; I shall defend the latter course.

The principle makes use of the dispositions of agents to *accept* sentences; but there are two other attitudes towards sentences from which acceptance should be distinguished. First, acceptance is not the same thing as *holding a sentence to be true*; if one accepts a sentence one holds it to be true, but the converse does not hold. One can, after all, hold a sentence to be true which one does not understand, and doing so is not sufficient for coming to believe what the sentence says. Second, acceptance is not the same as *uttering* a sentence. To illustrate this, suppose that one is presented with a sentence which states some deeply embarrassing truth about oneself; one may well accept the sentence without wanting to advertise this fact by uttering the sentence out loud. In addition, one may utter a string of words without understanding them, or utter a sentence which one understands without endorsing what it says, as in jokes, irony, and deception.

These distinctions bring to light a number of things which must be built into the notion of acceptance if the disquotational principle is to be true, and if sentence acceptance is to be one of the two ways of coming to have a belief. To accept a sentence (i) one must understand the sentence and (ii) one must be sincere. Finally, (iii) one can accept a sentence without performing any kind of outward behavior at all; each of us is familiar with the phenomenon of speaking to oneself in general, and, in particular, with endorsing a claim without giving voice to it. If behaviorism is to be defensible, this sort of "inward action" must be counted as a kind of action. Points (i)-(iii) might, however, give rise to the worry that even if an account of belief in terms of dispositions to accept sentences is materially adequate for some sorts of beliefs, such an account provides no real explanatory gain. After all, speaking to oneself is surely a mental act of some sort, and understanding and being sincere are surely

[7]Suitably modified for context-sensitivity, the principle says: Necessarily, if an agent is disposed accepts a sentence in a context in which that sentence expresses the proposition p, the agent believes p. For simplicity, I suppress references to context here and in what follows. I also ignore some serious difficulties which arise when we try to state the relevant dispositional property more precisely. For discussion of these, see Appendix E, pp. 223 ff.

mental states; so it seems that the terms used in explicating belief themselves need some sort of explanation.

If this worry is a worry that the sorts of facts about agents appealed to in (i)-(iii) are not verifiable, it may be easily dismissed; the view that facts about mental states are constituted by dispositions to act need not be wedded to the implausible verificationist theses which attended mid-century behaviorism. A more forceful formulation of the worry bases it on physicalism rather than verificationism. Points (i)-(iii) make it clear that explicating belief in terms of dispositions to accept sentences is explaining of one class of mental facts in terms of another; but, since all mental facts are constituted in some way or another by physical facts, one might think that this counts as no explanation at all.

In response to this, one might either claim that facts about accepting sentences — and so facts about sincerity, understanding, and inner speech — can be explained physicalist terms, or simply deny physicalism. But neither choice is forced on the defender of the view that belief is partially constituted by facts about what sentences agents accept. The most plausible versions of physicalism are stated, not in terms of the reducibility of mental phenomena to physical ones, but rather in terms of the supervenience of the mental on the physical. If some form of supervenience holds, then it does follow that there are purely physical sufficient conditions for any mental phenomenon's obtaining; but it does not follow that we can give necessary conditions as well. Since constitutive questions about the natures of things intuitively require necessary and sufficient conditions,[8] it might simply be the case that we can give an account of what it is for certain mental phenomena to obtain in terms of other mental phenomena, but that sometimes we will reach a kind of mental phenomenon which supervenes on physical facts, but which cannot be given a constitutive account in physical terms. If this is the case, then constitutive accounts of mental phenomena should not always be judged on the basis of their ability to explain these phenomena in physical terms, but rather on the basis of their ability to give a picture of the relationship between different classes of mental and linguistic phenomena which offer illuminating explanations of the interesting properties of these phenomena. So the defender of the present view of linguistic beliefs has three ways of responding to the physicalist challenge; I do not commit to one of them here.

There is, however, a third and stronger formulation of the worry that the nature of sentence acceptance empties the account of explanatory power. One might wonder whether a constitutive account of the facts appealed to in explaining sentence acceptance might involve the very facts about belief for which we are trying to give an account. If this were the case, the account of belief I've been offering would be covertly circular; and this would have important consequences for the status of the account. Perhaps we can do no better than to show the relations between the beliefs of agents and the meanings of expressions of their language, and cannot give a non-circular account of one in terms of the other. But if this is the case, then the sort of account sketched in this section does not say *what it is* for an agent to have a certain belief, since such claims typically carry with them a claim of some sort of explanatory priority. Circularity would block any such claim of priority for social facts about linguistic meaning. This is a worry worth bearing in mind; it is the topic of the next chapter.

[8]Compare a theory which says that what it is for an act to be morally right is for that act to lead directly to world peace and have no negative consequences. This probably does give a sufficient condition for an act's being morally right; but it is certainly not a full story about what it is for an act to be morally right. For more on this topic, see §1.2, "Constitutive claims," above.

But a more pressing response to the claim that dispositions to accept sentences are constitutive of beliefs is that such dispositions cannot be the whole story about belief; many agents, such as small children and animals, have beliefs without self-consciously endorsing claims in this way. There must be a way of taking the world to be a certain way which is not accepting a sentence or performing a gesture which has some social significance. Above I claimed that such creatures can have beliefs because they are capable of acting on the basis of claims; the question is then what it is to be disposed to act on the basis of a proposition.[9] Whatever our answer to this question, the result will be an account of belief which is, in a certain sense, disjunctive. Before going on to say what it is for an agent to act on the basis of a proposition in §8.4 below, in the next section I consider some reasons to think this sort of disjunctive account plausible.

8.3 Two classes of beliefs

On this sort of account, believing something is taking the world to be a certain way, and taking the world to be a certain way is being disposed to perform certain sorts of actions. There, however, the unity of the account ends; for the dispositions which may be constitutive of having a certain belief fall into two very different classes: dispositions to accept sentences, and dispositions to, in a sense to be explained below, act on the basis of a proposition. Were some account of this sort true, one would expect there to be some intuitive differences between beliefs constituted by acceptance of sentences, and beliefs constituted by various sorts of non-linguistic behavior.

In fact, there are many differences between *linguistic beliefs* — beliefs constituted by acceptance of a sentence — and *non-linguistic beliefs* — beliefs constituted not by acceptance of a sentence, but by a disposition to act on the basis of a proposition.[10] Much work on the nature of belief, judgement, and other mental states has emphasized the fact that animals which share no public language are yet capable of being in these states; and this has been thought to count against any theory which, like the present one, gives social facts about linguistic meaning a role to play in constituting the mental states of agents. Thus, for example, Fodor in *The Language of Thought*[11]:

> The only thing that's wrong with this proposal is that it isn't possible to take it seriously ... The obvious (and, I should have thought, sufficient) refutation of the claim that natural languages are the medium of thought is that there are nonverbal organisms that think.[12]

But a closer consideration of the differences between linguistic and non-linguistic beliefs shows that a theory of belief ought to account not only for the possibility of non-linguistic beliefs,

[9]Note that the question of what it is for a non-linguistic creature to have a belief is distinct from the question of what it is for a creature to have a non-linguistic belief; the latter but not the former requires an account of non-linguistic beliefs of creatures which possess a public language. For all I have said so far it is an open question whether linguistic creatures can have non-linguistic beliefs.

[10]More precisely, I count any belief p of an agent A as a linguistic belief if A is disposed to accept a sentence which means p, and count all other beliefs as non-linguistic.

[11]Fodor, Jerry

[12]Fodor (1975), 56.

but also for the differences between these and the beliefs we gain by explicitly endorsing the claims made by sentences of our language. This counts in favor, not of a theory which makes no use of facts about linguistic meaning, but rather of a theory which takes linguistic behavior to be one among several ways of having a belief, and so can explain these systematic differences.

8.3.1 The contents of perceptions and the contents of beliefs

The first such difference is an obvious one. Human beings can have beliefs about all sorts of things: mathematics, morality, fictional characters, the future, people and places they have never encountered. By contrast, the beliefs of non-linguistic animals are limited in a striking way: they are limited to objects and properties of which they have had some experience.

We are inclined to ascribe many different sorts of beliefs to animals. But consider what sorts of things a dog living in Ohio might have to do in order for us to find it reasonable to ascribe to it beliefs about, say, Japan, or the distant future. Presuming it has never travelled abroad, it is difficult to imagine a scenario in which the dog could have such a belief without imagining the dog coming to acquire a language. A constitutive account of belief should explain this difference; the theory partially developed in the preceding section is halfway to such an explanation, since the ability of people to have beliefs about things outside their immediate experience may be explained by their understanding of linguistic expressions which refer to objects outside their immediate experience.[13]

8.3.2 Determinacy of content

A further difference between linguistic and non-linguistic beliefs is that the former have a sort of determinacy of content which the latter often lack. Often, when we ascribe beliefs to non-linguistic creatures, we have the sense that the specification of the contents of these beliefs is somewhat arbitrary.

For example, consider a dog digging away in the back yard. There's more than one sort of thing which the dog is disposed to chew if he finds it as the result of his digging; among these are a bone and a piece of raw meat. So the dog is disposed to dig in the back yard, and is disposed to chew on what he finds, given that what he finds falls into some range of objects. How should we describe the dog? Should we say that he believes of a specific bone that that bone is in the back yard, and that he wishes to chew on that bone? Or that he believes that there is some bone or other in the back yard, and that he desires to chew on

[13] A related point about the relations between the contents of the occurrent thoughts and the immediate environment of the thinker has been noticed by Michael Dummett: "... the proto-thoughts of which a creature without language is capable — at least a creature of any of the kinds familiar in our experience — can occur only as integrated in his current activity, realized or frustrated. They cannot float free, as adult human thoughts can do"(Dummett (1985), 148-9). Just as we ascribe beliefs to animals, we sometimes ascribe occurrent thoughts to them. But these occurrent thoughts invariably involve activities in which the animal in question is engaged at the time when the thought is ascribed; by contrast, it is a feature of our every day life that we can think about whatever we like while walking down the street. The topic of this essay is not occurrent thought, but belief; nevertheless, Dummett's point indicates that the constraints placed by a non-linguistic creature's immediate environment on the contents of its mental states may be a pervasive distinction between language-using and languageless creatures. But why, we should ask, should this be the case?

any bone he finds? Or, alternatively, that he believes that there is either a bone or raw meat in the back yard, and that he wishes to chew on either? Or, instead, that he believes that there is something pleasantly chewable in the backyard, and that he wishes to chew on it, whatever it turns out to be? It seems to me that one does not want to come down on one of these as the uniquely correct description of the dog's mental state; the choice between them seems too arbitrary.[14] The natural thing to say is that each of these proposals gives a sort of rough characterization of the state of the dog, but that there is simply no fact of the matter which could make one of them true, and the others false.[15]

By contrast, imagine that you come home to find your slightly eccentric friend digging in your back yard. When you ask him what he's doing, he says "Your back yard contains the largest source of natural crude oil in central New Jersey, and I aim to prove it." Presuming that he is sincere in what he says, he believes that your back yard contains the largest source of natural crude oil in central New Jersey. Of course, we may presume that he believes some other things about your back yard; for example, that it contains some crude oil, and that it is in New Jersey. But there is nothing odd about saying that he believes all of these things, and there is no sense that in ascribing any of these beliefs you are giving a rough and imperfect characterization of your friend's state of mind. No doubt your information about his beliefs is only partial; but it seems perfectly determinate that he has just the beliefs that you ascribe to him. Again, this seems to be a fact which requires explanation; on the present view, it may be explained by the fact that you and your friend share a common language, and that his belief is constituted by his acceptance of the sentence used in the ascription to express the content of his belief.

8.3.3 Self-knowledge

Next, consider our knowledge of our own beliefs. It is often noted that we seem to have some sort of privileged access to the contents of our own minds; if I believe that Wednesday is

[14] It also seems wrong to say that the dog has *all* of these beliefs, and *all* of these desires; there is something absurd about viewing the dog's behavior as explained by a large number of parallel beliefs and desires working in concert.

[15] We should distinguish between the claim that there is an indeterminacy in which of our ascriptions of beliefs is true, and the stronger claim that there is an indeterminacy in the contents of the dog's belief. The former might hold without the latter if, for example, the dog possesses a concept which is not expressible in English, but has determinate thoughts involving that concept all the same; there might then be approximations of the content of its belief in English between which there would be no grounds for choosing, without the stronger form of the indeterminacy claim holding. The apparent indeterminacy would then be a result of our language being an imperfect tool for the description of the beliefs of dogs.

Which interpretation we choose does not much matter here; in either case, there is a difference to be explained. But some reason to think that there is indeterminacy "in the beliefs themselves" rather than merely in our characterizations of them is provided by the following consideration: if the thoughts of the dog were determinate, we should be willing to say of the dog that there is some property F — which may or may not be expressed by a predicate of English — such that the dog believes that there is an F in the back yard, and desires that thing. But to my ear even this sounds odd; the natural thing to say seems to be that there is no property F such that the claim that the dog believes that there is an F in the back yard is the uniquely correct characterization of the dog's mental state.

the day after Tuesday, then I need do no special research in order to determine that I do so believe. But this seems not to be the case with animals which have no public language; we might attribute to a dog the belief that there is a bone in the yard, but would not attribute to the dog the self-knowledge that he has that belief. In fact, there seems to be something absurd, or impossible, about attributing this sort of belief to a dog.

A view of the beliefs of language-using creatures as partially constituted by dispositions to accept sentences provides at least the beginnings of an explanation of the immediacy of our knowledge of the contents of our own beliefs. Consider the phenomenology of this kind of self-knowledge. If I ask myself, "Do I believe that Wednesday is the day after Tuesday?", I simply respond, "Yes," without further reflection. How is this possible? On the present picture, what it is for me to have this belief is for me to be disposed to accept the sentence, "Wednesday is the day after Tuesday." Accordingly, when I present myself with that sentence in the form of a question, my disposition manifests itself as assent. This is but a sketch of a complicated phenomenon;[16] but since self-knowledge of the obvious, unreflective, transparent sort seems invariably to be articulated in language, some of the mystery is resolved by taking our beliefs to be constituted by linguistic dispositions which are readily available for our survey.[17]

8.3.4 Productivity and systematicity

Accounts of the nature of various sorts of mental states are often taken, following the lead of Jerry Fodor, to be answerable to the dual requirement that they explain the *systematicity* and *productivity* of thought. The claim that belief is systematic is, roughly, the claim that any agent capable of believing one proposition is also capable of believing any other with the same constituents. For example, any agent capable of believing that Mary loves John is also capable of believing that John loves Mary. The claim that belief is productive is the claim that the number of propositions an agent is capable of believing is limited only by the agent's finitude. A simple example of the productivity of belief is that you can come to believe an indefinite number of propositions simply by conjoining or disjoining propositions you already believe. At some point, this process will come to an end; the propositions in question will become to complex for you to grasp them. But, intuitively, this is because of your limitations rather than because you, so to speak, ran out of propositions.[18]

Even if vague, systematicity and productivity do seem to pick out a genuine feature of human thought and belief. But are these also features of the beliefs of non-linguistic creatures?

[16]In particular, it offers no solution to the puzzles that arise from combining the thesis that we have privileged access to our own beliefs with the thesis that the contents of our beliefs are often fixed by facts external to us, such as facts about the meanings of sentences in the language we speak. For discussion of these puzzles, see the essays in Ludlow & Martin (1998) and Pryor (in preparation).

[17]The point is not that animals can never have beliefs about the mental states of others, or even themselves. (We might naturally describe a dog as knowing that his master is angry.) Rather, the point is that the kind of automatic and pervasive privileged access to one's own beliefs which seems to be a feature of human belief is missing in the case of non-linguistic creatures.

[18]So stated, neither of these claims is entirely clear. This is largely because each makes reference to facts about what an agent is *capable* of doing. The best interpretation of the claims, I think, is to take each as claiming something about what the agent can do without acquiring further conceptual capacities: without, that is, acquiring the ability to have propositional attitudes towards objects or properties with which she was previously unfamiliar.

Consider the requirement of systematicity. A dog can surely believe that there is *something* in the food bowl; that *a bone* is in the yard; that the neighbor's cat *was* just in the yard; that the sofa is not a *part of* his territory; and that the dog he is fighting is still *alive*. But can the dog also believe that *a bone was a part of something alive*? Were the dog's beliefs systematic, it could; but it is hard to imagine a dog being capable of believing something of this sort.

A similar point can be made about productivity. Human beliefs are productive in large part because we can conjoin or disjoin arbitrary beliefs to arrive at new ones; but it is not obvious that we can generate new potential beliefs for dogs by means of truth-functions. Can dogs, for example, believe conjunctions? In some cases we might be willing to ascribe conjunctive beliefs to dogs; we might say that a dog thinks that there's some food in his bowl, and that if he doesn't stop barking he's not likely to get any. But it does not seem that we can conjoin or disjoin arbitrary beliefs of a dog to get possible objects of his belief. A dog may certainly believe that there is food in his bowl and believe that his master is home; can he believe that either there is food in his bowl or that his master is home? Certainly, the number of propositions which we should be willing to ascribe to a dog on different occasions is quite large; but it does not seem to be arbitrarily large in the sense in which human beliefs are.

The systematicity and productivity of various sorts of thought were originally used by Fodor to argue that these kinds of thought must be underwritten by relations to expressions in a language which has a compositional semantics. Since many animals speak no such natural language and are yet capable of thought, this was part of an argument that the thoughts of all creatures are underwritten by an internal compositional language of thought. But the above examples indicate that this line of argument is a mistake; since productivity and systematicity seem only to hold of creatures which speak public languages, we want, not a theory which attributes productivity and systematicity to all creatures capable of belief, but an account of belief which explains why the beliefs of language-using creatures have these properties whereas the beliefs of languageless creatures lack them. And it seems that the obvious way to develop such an account is to explain these properties on the basis of properties of public languages, rather than on the basis of an internal language of thought held in common between human beings and non-linguistic creatures.

On the picture of belief I have been developing, we are put in a position to believe propositions by understanding sentences which express those propositions. On this picture, systematicity and productivity can be explained by relatively obvious facts about linguistic understanding, and not by deep facts about the nature of the brain. Beliefs are systematic because anyone who is counted as understanding a sentence $\ulcorner aRb \urcorner$ is also counted as understanding the sentence $\ulcorner bRa \urcorner$. Beliefs are productive because anyone who is counted as understanding a sentence S and a sentence S' is also counted as understanding the sentence $\ulcorner S \ \& \ S' \urcorner$, up to the limitations of their capacity to grasp very long sentences.

But, as we've seen, the beliefs of non-linguistic animals seem to exhibit neither systematicity nor productivity; so we need, not a theory which attributes these properties to them, but an account of non-linguistic beliefs which explains why they lack these features.

8.4 Acting on the basis of a proposition

These asymmetries in the natures of the beliefs had by language-using and languageless creatures provide more data for our account of what it is for non-linguistic creatures to have beliefs: we now want an account of what it is about these creatures in virtue of which they have the beliefs that they have, but also an account which makes sense of the properties of these beliefs discussed in the preceding section.

Above I claimed that having a belief is taking the world to be a certain way, and that taking the world to be a certain way is a matter of being disposed to perform certain kinds of actions. We saw above that one sort of disposition which can be constitutive of belief is the disposition to accept a sentence which expresses a claim about the world; the purpose of this section is to explain how creatures which lack language can come to have beliefs by acting on the basis of a proposition.

One traditional strategy for giving a broadly behavioral account of belief begins with the following belief-desire-action principle:

> [BDA] An agent desires q & believes that ϕing will satisfy the desire q iff the agent is disposed to ϕ.

This principle evidently has some plausibility; a dog who runs eagerly to his food bowl is naturally said to do so because it wants some food and thinks that running to the food bowl will get it some; and this seems to have something to do with the fact that the belief and desire fit together with the action in the way suggested by [BDA].

Three objections may be raised against the use of [BDA] as a constitutive account of belief or desire. First, the principle ignores the fact that an agent with these beliefs and desires might *not* be disposed to ϕ if, for example, the agent has some desire r stronger than her desire q, and satisfying r precludes acting so as to satisfy q. In addition to this problem of overriding desires, there is an analogous problem of overriding beliefs: an agent might desire q and believe that ϕing will satisfy q without being disposed to ϕ, not because of an overriding desire, but because the agent also believes that ϕ^*ing will satisfy q, and prefers this means of satisfying its desire. Finally, the principle ignores the fact that for any action ϕ which some agent is disposed to perform, there are indefinitely many different and incompatible ascriptions of beliefs and desires to the agent which fit this schema. Hence, if this sort of connection between belief, desire, and dispositions to action is to be one's story about what it is to have a given belief, one is committed to widespread indeterminacy of facts about the beliefs and desires of non-linguistic agents. And even if, as claimed in the last section, the mental states of such agents are marked by a kind of indeterminacy, not just any amount of indeterminacy is plausible; we do not want to be in the position of saying that there is nothing to choose between the hypotheses that a dog is digging in the back yard because he desires a bone to chew and believes one to be there, and that he is digging because he wants to chew a lost copy of the Gutenberg Bible, and believes one, disguised as a bone, to be in the yard.

One kind of response to the indeterminacy objection is to say that it arises from considering the dispositions of an agent singly; surely any single disposition of an agent is compatible with many different and incompatible ascriptions of mental sates to the agent, but the totality of the dispositions of an agent determine the beliefs and desire of the agent. One form this

proposal might take is the following:[19]

> B is the set of beliefs and D is the set of desires of an agent $A \equiv A$ is disposed to act in ways which would tend to satisfy every member of D in the nearest world in which every member of B is true.[20]

But this appeal to the totality of an agent's dispositions does not solve the problem of the indeterminacy of the beliefs and desires of agents; just as above we could vary the belief and desire ascribed while remaining consistent with [BDA], so here we can vary the an agent's belief and desire sets while remaining consistent with the above biconditional. The indeterminacy problem arises not from consideration of dispositions singly rather than collectively, but from the form of the belief-desire-action principle. Because the principle makes reference to both beliefs and desires, taking the principle to be constitutive of either without taking the other as given is something like taking the equation $x+y=10$ to define the value of x.

So to use [BDA] as the key to understanding the nature of belief, we need to take facts about the desires of agents as given. Now, were we using this principle to say what it is for an adult human being to have a certain belief, this would clearly be just to push the bump in the rug. Because beliefs and desires are both intentional mental states, and no facts about the desires of human beings suggest that they are in any sense more fundamental than beliefs, an account of belief in terms of desires would be unsatisfactory. But we are now concerned with the beliefs of non-linguistic creatures; and there is some reason to think that taking on board facts about the ends such creatures characteristically pursue might not be arbitrary.

Recall the discussion of the furry mammalian creatures above. The point of that example was to indicate that non-linguistic creatures with the same dispositions may yet have different beliefs if the ends pursued in their behavior sometimes differ. In that case, both creatures were disposed to stick their snouts in the mud, inhale, then blow the mud back out; but one did this with the end of ingesting small insects living in the mud, whereas the other used this technique as a way of ridding its snout of troublesome bacteria. The question then is whether it is legitimate to take ends like these for granted in giving an account of belief, and how far this move might get us.

We are on better ground in building these facts into an account of beliefs than we would be were we to make the corresponding move with respect to human beliefs and desires because in the case of non-linguistic animals, these ends seem like biological facts. It is evidence for the claim that the ends pursued by non-linguistic creatures are closely linked to biological needs that the vast majority of the desires and wants we are inclined to ascribe to animals we are inclined to ascribe to all members of their species. Note that, in the case of the furry mammalians, it was crucial that they belonged two different species with different biological needs; and this difference was the source of the difference in their beliefs and desires. This suggests that, in at least the case of the sorts of non-linguistic creatures with which we are familiar, it might not be unreasonable to regard the wants of such creatures as having a kind of explanatory priority over their beliefs.

[19]This formulation is similar to the "purely pragmatic" theory of belief and desire considered and rejected in Stalnaker (1984).

[20]Obviously this principle will run into difficulties with inconsistent beliefs and certain kinds of conflicting desires; I set these problems aside for the time being.

One might, of course, wonder what the biological needs of members of some species are, or what it is for some creature to pursue, for example, a certain kind of food as a goal. But this does not seem like the same kind of puzzlement as is registered when we ask what the facts are in virtue of which an agent can have a belief with a certain content. This is especially so since we need not take for granted any of the propositional attitudes of non-linguistic creatures. As stated, [BDA] made use of the propositional attitude of desiring that such-and-such be the case; but we can just as well restrict ourselves to object-directed attitudes like pursuing a cat or needing some food. Because such attitudes are not relations to contents, they do not seem to presuppose the facts to be explained. In what follows I shall sketch an account of the beliefs of non-linguistic creatures based on the thought that their beliefs are derived from their goals together with their dispositions.[21]

Supposing that we take facts about the ends pursued by non-linguistic creatures to be prior to and constitutive of their beliefs, the problem of indeterminacy discussed above disappears: for we can now, to speak, plug facts about their desires into the equation and solve for their beliefs. This modification is not, however, enough to turn [BDA] into a satisfactory constitutive account of belief; for one thing, the principle only mentions a sub-class of beliefs: means-end beliefs of the form "ϕing will get me x." But we often ascribe beliefs to animals which do not fit this mold. To use the previous example, it is sometimes natural to say that a dog believes that there is a bone in the yard; but this is clearly not a means-end belief of the required form.

This shows that we should recognize not only means-end beliefs of this sort, but also beliefs which are, in a certain sense, trivial preconditions for means-end beliefs. Many of the cases in which we are willing to ascribe to a dog the belief that there is a bone in the back yard are cases in which the dog is disposed to dig for the bone. In such cases, the account sketched above says that we should ascribe to the dog belief in the means-end proposition that by digging he can get the bone. But the belief that he can get the bone by digging only makes sense against the background belief that there is a bone in the yard; hence any grounds for ascribing this means-end belief are also grounds for ascription of the belief that there is a bone in the yard.

This observation leads to the following view of the beliefs of non-linguistic creatures:

> A non-linguistic creature believes $p \equiv$
>
> (i) there is some action ϕ and some end x of the creature such that the creature is disposed to ϕ in order to get x, &
>
> (ii) p either is the proposition that ϕing will get x or is a trivial precondition of the truth of this proposition

More informally, the claim is that all the beliefs of non-linguistic creatures are either beliefs

[21] One might also note, in support of this claim, that the ends pursued by non-linguistic creatures might plausibly be taken to be constituted by the dispositions characteristically had by members of their species. This is obviously in keeping with the present view of mental states as constituted by dispositions to action.

A worry for any general identification of animal needs with biological needs is that animals can clearly learn things, and acquire new desires on the basis of that learning. Think of, for example, a chimpanzee who learns to desire the approval of her master. I am not sure what to say about such cases.

that some action of theirs will satisfy a biological end, or a trivial precondition of such a belief.[22]

This is all a bit loose; I have not, for example, said exactly what the relation of being a 'trivial precondition' amounts to. But it would be a mistake to try to make matters much more precise. As noted above, our ascriptions of beliefs and wants to animals are marked by a kind of indeterminacy and arbitrariness; hence it is unlikely that much is to be gained by trying to give a precise characterization of the nature of non-linguistic beliefs when our practice of ascribing these beliefs suggests that there is no precise matter of fact about what those beliefs are. This point applies to the means-end beliefs we ascribe to animals as well as to the preconditions of those beliefs. It would surely be odd to say of the dog that it believes that digging will satisfy the end of acquiring a bone to chew on; rather, we find it more natural to say that it believes closely related propositions, such as that it can get the bone by digging. Again, there seems little point in trying to give a precise characterization of the permissible variants of means-end propositions which can be ascribed as beliefs to animals.

Even if the foregoing is enough to solve the problem of indeterminacy as it arose for [BDA], it leaves untouched the problems of overriding desires and overriding beliefs. To see how the problem of overriding desires arises in the context of the present account, consider the following plausibly possible (if overdescribed) scenario:

> Fido wants very badly to chew on a bone, and knows that one is buried in the backyard. But Fido does not go digging for the bone in the yard, since he wants to avoid a beating from his master even more than he wants to chew on a bone; and he knows that digging in the back yard will lead to a beating from his master.

The problem is to account for Fido's beliefs that digging would get him the bone, and that there is a bone in the back yard. For, contra the present theory, Fido has these beliefs even though he is not disposed to dig for the bone; the desire which would lead to his being so disposed was overridden by the stronger desire to avoid a beating.[23]

When we are inclined to ascribe beliefs to non-linguistic animals without the animal having any disposition which corresponds to the belief, this is largely because we think that the animal *would* have a certain disposition, were some interfering factor not present. Exploiting this intuition, we can arrive at an extension of the above account of non-linguistic beliefs capable of handling counterexamples like that of Fido. First, taking for granted facts about the ends of the agent and the dispositions by which the agent pursues those ends, we get from the above biconditional an account of a subclass of the agent's beliefs. In the case of the example of the preceding paragraph, this will assign the dog the belief that not digging will enable him to avoid a beating. We then consider the nearest possible world w in which the dog lacks this means-end belief, and ask what difference, if any, the lack of this belief would make

[22]The formulation here is slightly misleading, because it makes it seem as though only dispositions to pursue particular objects can yield beliefs. But sometimes it might be natural to describe an animal as, for example, hunting for some food without there being any particular food he is hunting for. In line with this, the above could be restated using corner quotes and meta-linguistic variables, and the values of 'x' restricted to singular terms and descriptions.

[23]One might say that he is disposed to dig for the bone in the absence of conflicting stronger desires; but at this stage I want to avoid building facts about desires into the conditions for the manifestation of the dispositions constitutive of the animal's belief.

in the dispositions of the dog. In this case, plausibly, the difference is that the dog would be disposed to dig in the back yard for the bone; hence the account of non-linguistic beliefs would assign to the dog at w the beliefs that digging will get him a bone, and that there is a bone in the yard. The lesson of the above case of conflicting desires shows that our ordinary practice of attributing beliefs to non-linguistic agents also licenses us to ascribe these beliefs to the dog in the actual world. That is, we often ascribe to non-linguistic animals not only the beliefs which correspond to their actual dispositions, but also the beliefs which correspond to the dispositions they would have, were they lacking one or more of their actual beliefs.[24]

The same move can accommodate the problem of overriding beliefs. Suppose that a creature wants x, believe that ϕing will satisfy x, but also believes that ϕ*ing will get it x, and so is not disposed to ϕ. On the present account, the belief that ϕing will get it x is accounted for by the fact that, plausibly, in the nearest possible world in which the creature lacks the belief that ϕ*ing will get it x, the agent would be disposed to ϕ; and this is why we say of the creature that it believes that ϕing will get it x.

In §8.3.1 above I noted that the contents of the beliefs and other attitudes of non-linguistic creatures are limited to observable features of their environment. The following counterexample to the foregoing shows that we need some such perceptual constraint on the contents of the beliefs of animals for the account to be satisfactory:

> Animals of a certain species need vitamin B for survival, and often ϕ in order to get vitamin B. But the animals certainly do not believe that ϕing will get them vitamin B; it's absurd to think that they believe anything about vitamins at all![25]

This sort of case can be resolved by imposing the requirement that the beliefs of animals be limited to things they can perceive; plausibly, vitamins are not a part of the contents of

[24]The above account may then be restated as follows:

> A non-linguistic creature believes p iff
>
> *either*
>
> (i) there is some action ϕ and some end x of the creature such that the creature is disposed to ϕ in order to get x, &
>
> (ii) p either is the proposition that ϕing will get x or is a trivial precondition of the truth of this proposition
>
> *or*
>
> p is a proposition such that there is some proposition q which satisfies (i) and (ii) and which is such that, were the agent to lack the belief q, p would satisfy (i) and (ii).

There is nothing circular about this kind of account. The biconditional we used as our account of non-linguistic beliefs assigns a series of beliefs to the animal; call these level 1 beliefs. We then consider, for each level 1 belief, the nearest possible word in which the animal lacks that belief, and assign level 2 beliefs based on the new dispositions which result in that world. Note further than it is not the case that each level 1 belief will generate a level 2 belief. For imagine that Fido is laying on the ground, and, upon hearing the rattle of food in his bowl, runs to his bowl. Fido believes that by running to his bowl he can acquire some food; but, in the nearest possible world in which Fido lacks this belief, perhaps he would simply have continued to lay on the ground. In this case there is no need to ascribe further beliefs to Fido. It is unclear how many times this process should be iterated. I have not come up with any cases in which we want to ascribe level 3 beliefs to non-linguistic animals.

A potential worry about this approach is that it will generate too many beliefs; but I have so far not been able to come up with a convincing counterexample of this kind.

[25]Thanks to Gillian Russell for the example.

the perceptions of animals.[26] This is all well and good; but what we want is not an *ad hoc* device for blocking counterexamples, but some sort of explanation of the fact that the beliefs of animals should be limited in this way.

An explanation of this sort can be provided by an account, like the present one, which takes languages to be a vehicle for the thoughts of language-using creatures. I take it that it is a necessary truth about beliefs that we are willing to ascribe a belief that x is F to an agent only if the agent is, in some sense, acquainted both with the object x and the property F. I further take it that perception is one way to become acquainted with objects and properties. One way to view the claim that public languages can be vehicles of thought is as saying that language gives us a second way to be acquainted with objects and properties: by understanding linguistic expressions which have those objects and properties as their contents. The fact that non-linguistic agents lack public languages is then the explanation for the fact that, having only perceptual means of acquaintance with objects and properties, their beliefs are limited to the things they encounter in perception. This point about the relationship between belief and acquaintance, along with the present view about the relationship between thought and language, can then provide at least a sketch of an explanation of the perception-based limitations which characterize the beliefs of animals and which block counterexamples like that of the preceding paragraph.

This provides an explanation of the first of the four differences between linguistic and non-linguistic beliefs; the account also has the resources to explain the other three. Consider first the indeterminacy in the beliefs of non-linguistic agents. This is partly the result of the fact that there is no canonical form for the expression of the means-end beliefs and their preconditions which comprise the beliefs of animals; but it also results from the fact that, on the present model, the contents of the beliefs of animals are partially derived from their needs. The key point on this score is that specifications of the needs or wants of a creature are extensional; if "x" and "y" are singular terms or descriptions, then the truth of $\ulcorner a$ needs $x\urcorner$ and and $\ulcorner x = y\urcorner$ entail the truth of $\ulcorner a$ needs $y\urcorner$.[27] But belief ascriptions are not extensional; hence, even if $\ulcorner x = y\urcorner$ is true, the ascriptions $\ulcorner a$ believes that ϕing will get it $x\urcorner$ and $\ulcorner a$ believes that ϕing will get it $y\urcorner$ may attribute different beliefs to the referent of "a". But there will be no basis for choosing between these two ascriptions, since the contents of the beliefs of a are determined by a's needs, and the claims that a needs x and that a needs y will in such cases be equivalent. Analogously, claims about the action an agent is undertaking are extensional, and this too generates some indeterminacy in the means-end belief ascribed.[28]

[26] And, if they were, we would not be reluctant to ascribe beliefs about vitamins to them.

[27] Recall that the needs and wants in question are object-directed attitudes, not the propositional attitudes of wanting that or desiring that something be the case. Note also that the needs claims, insofar as they can generate means-end beliefs, will be restricted by the perceptual constraint above.

[28] There is a worry about this: the fact that needs claims are extensional yields far too much indeterminacy in the beliefs of non-linguistic creatures, even if we apply the perceptual constraint on the constituents of the object of the need. As noted above, we must make room for the fact that some needs are not directed toward specific objects, but rather are more generally directed at, for example, *finding some food*. But if we allow descriptions into the specifications of the needs of animals, it will be possible, for virtually any case in which an animal is pursuing some object o, to come up with some very complicated description D, every constituent of which is perceivable by the creature in question, which refers to o. The worry is that the extensionality of needs claims, along with the truth of $\ulcorner a$ needs $o\urcorner$, will entail an implausible belief of the form $\ulcorner a$ believes that ϕing will

The fact that the non-linguistic agents with which we are familiar lack beliefs about their own beliefs — and so lack self-knowledge — falls out of the present account. The beliefs of such agents are always means-end beliefs which relate an action to some end, or preconditions for such beliefs; but the non-linguistic agents with which we are familiar never pursue facts about their own mental states, and the action employed in pursuit of an end is never the formation of a mental state.

The failure of productivity and systematicity may be similarly derived from the source of non-linguistic beliefs in means-end beliefs of this sort. Given that a creature may have a limited number of ends and a limited number of action-types at its disposal for pursuing those ends, it is not surprising that its beliefs will not be productive; the beliefs it is capable of forming will be limited to the means-end actions it can pursue and preconditions for the success of those actions. Similarly, in the case of systematicity, this an animal may be such as to pursue an end x by ϕing — and so believe that it can get x by ϕing — but not be the sort of creature which could pursue ϕing as an end by getting x — and so not have the capacity to believe that it can ϕ by getting x.

This completes the sketch of the manner in which the beliefs of non-linguistic creatures are constituted by their dispositions to action.[29] But the account still leaves a fundamental question unanswered: how are linguistic and non-linguistic beliefs related?

8.5 Integrating linguistic & non-linguistic behavior

In the present context, this is a pressing question because the account sketched above may well encourage a picture of the class of possible believers as falling into two categories: those, like us, whose beliefs have everything to do with their linguistic behavior, and nothing to do with their non-linguistic actions; and those, like our pets, whose beliefs are rooted in biological needs and end-directed actions of a specific kind. But this picture of belief is absurd; there must be *some* kind of connection between linguistic and non-linguistic beliefs, and between linguistic and non-linguistic behavior.

This may be brought out by considering three points:

1. Language using creatures can apparently have beliefs which they are not disposed to express with any sentence they understand; so the realm of non-linguistic beliefs is not restricted to languageless creatures.

get it $D\urcorner$. I am not sure what to say about this.

[29] There is another challenge to the material adequacy of the account of the beliefs of non-linguistic creatures, in addition to the worry about learned behavior mentioned in fn. 21 above, which may require some modification. Suppose that your dog is laying on the lawn, and a cat runs by in front of it; the dog lifts its head, and follows the path of the cat across the lawn. Clearly the dog sees the cat, and sees that the cat is running across the lawn. Does the dog also believe that the cat is running across the lawn? I am inclined to say not; but, if the dog does believe this, the case is not covered by the above account. The modification required would not be very significant, though, because since I am taking belief to be explained partly in terms of the contents of perceptions, we can use facts about the contents of the dogs perceptions to explain the belief acquired. Such cases seem always to be cases in which the belief in question is tied closely to the content of a perception; and there is always some doubt whether we should say that the creature merely sees or hears that something is the case, or that it believes something to be the case.

2. We often explain our non-linguistic actions by citing beliefs; but if those beliefs really are fixed by linguistic behavior alone, it is hard to see how beliefs could figure in any such explanation.

3. The very idea that there is no common thread at all between our beliefs and those of non-linguistic creatures is counterintuitive. The word 'believes' is evidently not ambiguous in the same sense that 'bank' is; so we should expect there to be some reason why we use this term to describe both the states of adult language-using humans and those of dogs.

So the picture sketched above really is absurd; there are widespread connections between the two realms which it simply ignores. If the present account of belief is on the right track, it must be able to make sense of these three observations.

8.5.1 Non-linguistic beliefs of language-using creatures?

It may seem that the first of these observations — that language-using human beings have beliefs which are not expressed by any sentence they are disposed to accept — poses no serious problem. After all, we've just finished giving an account of the beliefs of creatures which speak no public language; why not just carry this account over to the case of language-using creatures, and use it to explain their non-linguistic beliefs?

But this move seems not to be very promising, for a reason mentioned in the preceding section. The account of the beliefs of non-linguistic creatures given there took as given the desires of those creatures; and this was justified on the grounds that those desires were closely related to the biological needs of those creatures. But we are now considering the beliefs of adult human beings; and it is clear that the desires of human beings need not be so closely tied to biological needs common to all human beings. So it seems objectionable in this case to take desire for granted for the purposes of explaining belief.

There are, I think, three kinds of cases in which a person can plausibly said to believe p without being disposed to accept a sentence which means p.[30]

Perceptual beliefs

The first place to look is to perception; for it is obvious that the contents of perceptions provide many of the contents of our beliefs, and equally obvious that we are not, in general, capable of expressing everything presented to us in experience in expressions of a natural language. So it seems as though perception should be a rich source of non-linguistic beliefs for language-using agents.

[30]It's important to note that we're only concerned here with the case of *belief*. So the claim that one can *think* about something without having words which express the object of your thought "running through your head," while true, is not to the point. Believing a proposition and thinking about one are different propositional attitudes. Further, being disposed to accept a sentence need not require that you be aware of this disposition, or that the acquisition of such dispositions have any special phenomenal quality. So the relevant cases are not ones in which a belief is present without attention to language being a part of the relevant agent's phenomenology, but ones in which a belief is present without the corresponding disposition.

But, on reflection, this is not at all clear. Imagine scanning your eyes across a bookcase full of books of many different colors; the scene which presents itself to you when you do this is a fairly complex one, and it is most implausible to think that you should able to articulate in language every aspect of it. But it is equally implausible to think that you formed beliefs about every aspect of the book case which was presented to you. If asked about your experience afterwards, you will be able to describe some aspects of it; and it is plausible that these recollections do reflect beliefs you have formed. But, were we to take every aspect of the scene visually presented to you to be the object of belief, we would have to claim, implausibly, that in perception you acquire a vast number of beliefs, only to forget them with surprising speed. Better to sharply separate belief from visual experience: each are ways of being related to claims about the world, but they are distinct ways of being so related.

So we need some way of distinguishing between cases in which a proposition p is merely part of the content of the experience of an agent, and those cases in which the agent also goes on to believe p. It is not entirely implausible, I think, to make this criterion the acquisition of a disposition to accept a sentence which means p.[31] Consider again the case of scanning your eyes across a bookcase; suppose that it was part of your visual experience that there was a grey book near the middle of one of the lower shelves. It seems reasonable to think that you formed the *belief* that there was a grey book on one of the middle shelves just in case you acquired the disposition to accept a sentence which expresses this claim. Plausibly, if you did not gain such a disposition, the placement of the grey book was simply a part of the content of your visual experience about which you did not form a belief.

In any case, though, this is a class of cases about which we can afford to be agnostic. The present account of belief takes perceptual content as given for the purposes of explaining belief. The priority of perceptual content over the contents of beliefs seems plausible in any case; but it is worth noting that it is entailed by any account of beliefs in terms of dispositions to action. A full specification of dispositions requires a statement of the conditions under which the disposition is manifested; and the manifestation conditions for dispositions to action will typically make reference to the contents of the perceptions of the agent in question. For example, a full specification of the disposition to accept a sentence will perhaps say that the agent is disposed to accept the sentence *when presented with it*; and being presented with it will typically involve hearing an utterance of it, or seeing an inscription of it. Since in the case of perceptual belief the belief acquired is derived from the content of a perception, we can leave open the question of how to distinguish between cases of merely having p as part of the content of an experience and believing p, while being sure that such cases will not provide an obstacle to the project of explicating belief in terms of perception, action, and linguistic meaning.

Beliefs manifest in action

There is some temptation to recognize another class of language-independent beliefs in beliefs which are, in a certain sense, manifest in action. Imagine that you are walking down some

[31] This will obviously only work for language-using creatures. I am inclined to think that in the case of animals, the distinction between perceptual contents which are and are not believed is very vague, and has something to do with the animal's dispositions to act on the basis of the perception.

familiar stairs with a friend, and the friend trips on the last step. You might say, "Sorry; I knew that there was another step there. I should have told you." You might well speak truly; but it is not clear that there is any sentence which expresses the proposition knowledge of which you self-ascribe which, prior to the experience, you would have been disposed to accept.

It is not totally implausible to deny that there are any such cases; it seems likely that, were you stopped before the last step and asked, "Is there another stair there?", that you would have been willing to say that there was. But it seems theoretically driven to insist that this must always be the case; and in many cases we are willing to ascribe such beliefs to non-language using animals on the basis of what looks to be very similar behavior.

Another response is to think of these cases as exhibiting a kind of practical knowledge of how to do something. On this view, knowledge how to do something is not always analyzable in terms of knowledge that something or other is the case. If this were right, then it would be natural to think that we sometimes attribute theoretical knowledge to agents on the basis of their knowing how to do something. In the example above, the proposition known is clearly closely related to the action — walking down the stairs — that you knew how to perform. This response still requires an independent account of "knowledge how;" but it is not implausible to think that such an account could be given in roughly the same dispositional terms presently being used to analyze belief.[32]

Unconscious belief & self-deception

A final class of problematic cases are alleged cases of suppressed beliefs. Now, only a subclass of the states often referred to as suppressed beliefs are problematic; sometimes a suppressed belief is simply a belief one has had for a long time, but did not know that one held. These

[32]A related case concerns intentional actions which one pursues rather automatically, without a great deal of conscious thought. For example, when you type, you might well intentionally hit the 'y' key with your right index finger, without thinking at all about how to accomplish this task. In such cases, it seems that you know how to do something; but you cannot express any corresponding means-end belief of the form "By moving my right index finger in such and such ways, I can hit the 'y' key." Now, if we adopt the view that all intentional action is produced by a belief and desire, such cases pose a serious problem. For then there must be a means-end belief of the sort I described, and yet it seems clear that, even though I know how to hit this key with my right index finger, I could not express any such belief with a sentence of English. (It may be that some sentence of English which I understand does express the required proposition; but I am not disposed to accept it, since, given several sentences which purport to describe the motion I perform with my right index finger, I would have no idea which is true.)

In my view, the right response is to reject the view that every intentional action must be explained on the basis of an appropriate desire and means-end belief. It is implausible that I am disposed to accept a sentence which expresses an appropriate means-end proposition; but, fortunately, it is just as implausible that the sort of practical know-how exemplified by typing and riding a bike is to be explained on the basis of a complex means-end belief. Rather, we should accept that knowledge how to do something is not always reducible to knowledge that a particular means-end proposition is true. (For a defense of the contrary view that knowledge how is a species of knowledge that, see Stanley & Williamson (2001).)

One might also see this case as pointing towards a stronger conclusion: that action-types are not individuated by their associated beliefs and desires. Were this true, this would do much to allay the worry that the present account's appeal to actions in explaining belief is viciously circular. More on this below.

cases are no problem; the present account of belief only requires that an agent be disposed to accept a sentence which expresses the belief when presented with it, and does not require that the agent actually have accepted the sentence. The suppressed belief cases which are problematic are those in which one wants to say that an agent believes something, even though she will deny the relevant claim if asked about it, and will not "inwardly assent" to it either.

A hopeful thought is that it seems that in all cases of suppressed belief, there is some behavior on the basis of which the belief is attributed; and this makes it seem possible that there is some broadly behaviorist treatment of suppressed belief. But I am not sure exactly what should be said about these cases. A tempting move is to appeal to counterfactuals which abstract from the features of particular cases which make agents unable to recognize the suppressed belief. But it is not obvious that this sort of move is non-circular, and far from obvious that there will be any unified treatment of these 'blocking' features across diverse cases.

I think that the most promising way of thinking about these cases is to treat them as a kind of extension of our ordinary notion of belief. To see that there is something odd about cases of suppressed belief, note that ordinarily we think that it cannot be true that I just realized p and that I have believed p for some time. But in cases of suppressed or unconscious belief, it will often be very natural to say both of these things. But this is a challenging class of cases which requires a more thorough discussion.[33]

8.5.2 Linguistic beliefs and the explanation of non-linguistic behavior

Consider the claim that Chris went to the store because he knew that there were no beans in the cupboard. On the face of it, this explains a certain event — Chris's going to the store — on the basis of a certain state of Chris — his believing that there are no beans in the cupboard. But on the present view this state is constituted by Chris's disposition to accept the sentence, "There are no beans in the cupboard"; and how could a disposition to accept a sentence explain his getting in the car to drive to the store?

This kind of case is a more serious problem than those discussed in the preceding section, for it concerns not just the fact that people can have non-linguistic beliefs, but rather challenges directly the claim that many of the beliefs of adult human beings are constituted solely by linguistic behavior. And the challenge is quite a general one, for we often explain the behavior of agents by citing beliefs which, on the present view, are constituted by linguistic dispositions.

There is a general problem about how beliefs can provide non-trivial explanations of behavior at all, if beliefs are dispositional states. This does indeed require some explanation; but it is not the problem we are confronting here. The present problem is to explain how, if a belief is constituted by a disposition to accept a sentence, the belief could be so much as *relevant* to activities like going to the store.

Though I do not have a fully satisfactory answer to this problem, I think that I do see the direction in which a solution can be found. The present objection derives much of its force

[33] A different but related sort of case is provided by putative cases of 'tacit' or 'sub-personal' belief, as in, according to some theories, our knowledge of the grammar of our native tongue. I am somewhat skeptical of the claim that these are cases of genuine belief, but am not sure exactly what should be said about them.

from the thought that a disposition to emit a string of sounds could not possibly explain going to the store. This is a plausible thought; but here it is important to recall that dispositions to accept sentences are not just dispositions to emit strings of sounds.[34] Rather, it is a precondition of accepting a sentence in the relevant sense that the agent understand the sentence, and be sincere in assenting to it.[35] Further, it may turn out that it is a condition on *sincerely* accepting a sentence which means p that one have certain non-linguistic dispositions. If this is right, then this may be the bridge between linguistically constituted beliefs and non-linguistic behavior needed to make sense of ordinary explanations of behavior like the one described above.

To see how this might work in a specific case, return to the example of Chris and the beans. Suppose that Chris wants some beans, and says "There are no beans in the cupboard" and then, after a moment's pause, "I can get some beans by going to the store." But then Chris simply sits there, and shows no inclination to get up and go to the store.

There are several different interpretations of Chris's state which would make sense of his behavior. It may be that Chris is feeling lazy, and that his desire for beans is simply outweighed by his desire to remain on the couch a bit longer. Or, it may be that he's reluctant to go to the store because he thinks that there's some preferable way of getting beans; perhaps he thinks that you'll go instead.[36] A third interpretation, though, is also possible: one or both of his utterances were not sincere; perhaps he's really not sure whether there are beans in the cupboard, or has some doubts about whether the store is open right now. If the last interpretation were correct, then this would be a case in which failure to have a certain non-linguistic disposition — the disposition to go to the store — ruled out Chris's accepting the two sentence he uttered.

But, even given that there could be cases in which the last is correct, how does this help with the problem of explaining action by means of beliefs? What this story shows is that, even in cases where an agent's belief p is constituted by his disposition to accept a sentence which means p, the claim that the agent believes p entails some facts about his non-linguistic dispositions. In particular, it entails that, absent interfering desires or beliefs, if the agent desires x and believes that if p, ϕing is the best way to get x, the agent will be disposed to ϕ. Crucially, the value of "ϕ" here might be a non-linguistic disposition like the disposition to go to the store. And this makes it intelligible how, even given the above story about the relationship between language and belief, attributing beliefs to an agent could be a way of explaining her behavior.

To be sure, this is only a sketch of an answer to the challenge about explaining behavior. It leaves two fundamental questions unanswered. This first asks what makes it the case that, in the situation described above, one rather than another of the three competing interpretations of Chris's state is the correct one. That is, we need, not just the claim that non-linguistic dispositions are sometimes preconditions for sincerely accepting a claim, but an account of the conditions under which assent to a sentence is sincere. This is particularly important since,

[34] See §8.2 above.

[35] Of course, this does raise the worry, mentioned above, that the account will end up being circular; more on this in the next section.

[36] These cases should be familiar from the discussion in §8.4; we've basically described a case where [BDA] fails, and noted that this may be because the agent has either what I called above an overriding desire or overriding belief.

given the above discussion, there is some reason to think that such an account will involve the facts about belief to be explicated. Second, we have only begun to give an account of how explanations of actions in terms of mental states can be understood on the present picture; it is not at all obvious that all explanations of actions in terms of beliefs can be made to fit the mold of Chris and the beans. The second task is beyond the scope of this essay; I return to the question of the nature of sincere acceptance in §9.2 below.

8.5.3 The disunity of the account

We are now in a better position to discuss the third of the worries raises at the outset of this section. This was that it is implausible to think that there is *nothing* in common between the facts in virtue of which human beings have the beliefs that they do and the facts in virtue of which non-linguistic animals have the beliefs that they do. Were this the case, it would be a mystery why we apply the same word to each.

Above, we noted that non-linguistic dispositions are relevant even to the foundations of linguistic beliefs, for certain non-linguistic dispositions may be required for an agent to sincerely accept a sentence. We should also note that the non-linguistic dispositions which were relevant exactly mirrored the dispositions used in giving our account of non-linguistic creatures. In each case, the key principle connecting belief to action was the belief-desire-action principle [BDA]. In the case of non-linguistic creatures, we took their desires for granted, and used [BDA] to derive their means-end beliefs, which were in turn used to derive their other beliefs; in the case of linguistic creatures, we took conformity with [BDA] to be a (admittedly, defeasible) criterion for sincere acceptance of a sentence.

Roughly, then, the principle which was constitutive of the beliefs of non-linguistic creatures is a constraint on the beliefs of language-using agents. As above, this is so far quite vague, and needs to be supplemented with an account of sincerity which states the conditions under which violation of [BDA] is sufficient to count an utterance of an agent as insincere. But the picture we get is not the unreasonable picture of two realms of believers, whose beliefs are constituted by entirely different facts about them, and which are grouped together only by linguistic accident. Rather, the view is that a principle connecting belief, desire, and action plays a fundamental role in our story about what it is for any agent to have a certain belief. It's just that it isn't the whole story, or even close to the whole story, about the beliefs of human beings. Membership in a linguistic community gives human beings a new medium for the formation of beliefs, and makes the character of human beliefs different in a number of fundamental ways from those of non-linguistic agents.

Chapter 9

Action and Communitarianism

Contents

9.1 Circularity and interdependence

It is now time to turn to a worry about this view of belief which was in the background throughout the previous chapter. The worry is that the account of belief there is obviously very far from being an account of belief in non-mental and non-intentional terms. The account was given in terms of dispositions to perform certain intentional *actions*; among the conditions of manifestation for these dispositions were the having of certain *perceptions*; among the relevant actions were instances of accepting sentences, which seem to presuppose *sincerity* and *understanding* of the sentence; and the sentences were taken as having certain *meanings*. It is not unreasonable to think that, in giving an account of one or more of these notions, we will have to bring back in the very facts about belief and related mental states we sought to give an account of.

There are really two questions here. First, is the account circular — i.e., do we need to employ facts about beliefs in giving constitutive accounts of any of the notions employed by the theory of belief sketched in the last chapter? Second, if it is circular, what does this show about the present account of belief?

This chapter is intended to be a kind of answer to the first of these questions. Though I do not think that the arguments which follow are as conclusive as the negative arguments

given in previous chapters, I do think that there are strong considerations in favor of the view that the threat of circularity is not so strong as it appears.

But suppose for the moment that the account *were* circular — that is, that a constitutive account of one or another of meaning, understanding, sincerity, and intentional action made use of facts about belief. A sanguine response to this possibility is that it does not show that the account of belief is false; it only shows that belief and one or another of meaning, understanding, sincerity, and intentional action are *interdependent*.[1] This response is a bit overly sanguine. It is hard to see this kind of use of "interdependent" as anything other than a euphemism for "circular." A verdict of circularity would be a genuine loss. After all, we have been trying to state and evaluate constitutive accounts of various kinds of phenomena to do with intentionality; and, as discussed above, it seems clear that giving a satisfactory constitutive account — saying what it is for a certain kind of fact to obtain — requires that the facts invoked in the account have a kind of conceptual or explanatory priority over the facts under investigation.[2] If the present account were circular, no such claim of explanatory priority would be tenable.

But not all would be lost. The task of Parts I and II of this essay was to argue against two versions of the individualist view of the relationship between thought and language. Those arguments would still stand, and the positive account of belief given in the last chapter would still be a non-individualist alternative to individualist accounts of the natures of mental states.

So much for the question of what circularity would mean. The more interesting question is whether the present view of belief really is circular; what follows is an argument that the account of belief given in the last chapter is not just an 'interdependence' account of belief, but a genuinely priority-revealing constitutive account.

There are two ways to argue for this conclusion. The more ambitious is simply to provide constitutive accounts of all the circularity-threatening notions — meaning, sincerity, understanding, intentional action, etc. — in terms which do not presuppose facts about belief. This would be an impressive feat; but it is not one which I know how to perform. Accordingly, my argumentative strategy will be more modest.

The argument will have two parts. First, I shall argue that the circularity-threatening notions can all be accounted for in terms of one of them: facts about the intentional actions, and dispositions to action, of agents. This will not take the form of giving full constitutive accounts of meaning, understanding, and the like in terms of action; rather, I simply suggest some *prima facie* reasons for thinking that such accounts might be possible. So this first stage of the argument aims to reduce the problem of circularity to the issue of the nature of intentional action.

Here the problem of circularity is at its most difficult; so far as I know, every going account of what it is for an agent to perform an intentional action of a given type is given in terms of the mental states of the agent: that agent's beliefs, desires, and intentions. To give a constitutive account of intentional action which did not have this feature would amount to a revolution in the philosophy of action; I needn't add that I am not prepared to give such an account. What I am prepared to do is to present several plausible arguments, which employ no specifically communitarian premises, for the conclusion that beliefs are not constitutive of

[1] One might support this with the McDowellian denial that any "sideways-on" account of intentional notions are available. See McDowell (1994), Lecture II.

[2] For discussion of this point, see §1.2.

intentional action.

One might think that this is not enough: that the threat of circularity remains until it is shown how a constitutive account of intentional action which makes no use of the beliefs of agents might be constructed. I am not sure whether this is right. In order to raise some doubts about this thought, it will be useful to pursue an analogy with perception.

The picture of intentionality I am inclined to endorse takes there to be two fundamental sources of intentionality: the contents of the perceptions of agents, and the actions agents are disposed to perform. (The account of belief of the last chapter presupposes facts about the contents of perceptions just as surely as it presupposes facts about intentional action.) So far we have been focusing on circularity worries as they arise in the case of intentional action rather than in the case of perceptual experience; this is natural, since it is very implausible that belief is prior to perceptual experience in the order of explanation, and hence very implausible to think that perception poses any special problems of circularity.

The relevant point at present is that the thought that belief is not prior to and constitutive of perception does not rest on our having in hand a constitutive account of perception which makes no use of facts about belief. Indeed, it does not even rest on any confidence that such a constitutive account of perception is even possible. For suppose (as seems plausible to me) that there is no way of filling out formulae like

a has a visual experience with content $p \equiv \ldots$

to yield a constitutive account of one kind of perceptual experience. We would still think that an account of belief which made use of facts about perceptual content was not in danger of being circular.

There is probably more than one reason for our confidence in this; but, I suggest, one fact lying behind this confidence is that it seems that we can explain a number of facts about belief in terms of facts about perception. We might, for example, explain our ability to have beliefs about certain objects in terms of perceptual acquaintance with those objects; we might explain the justification of some beliefs in terms of perceptual experiences. What this suggests is that we can take one class of facts to be constitutive of another on the basis of the former explaining features of the latter, even if we can have no independent account of the former.

This is a roundabout way of suggesting that we should take the analogous thought seriously in the case of intentional action as well. Perhaps, as we supposed in the case of perception, we can give no constitutive account of intentional action; but perhaps we would be justified in taking action to be prior to belief anyway on the grounds that we can explain a number of facts about belief in terms of features of the actions of agents. I claimed above in Chapter 8 that we can explain what it is to have a belief with a certain content in terms of facts about the dispositions to action of agents; below I shall argue that we can give an explanation of certain features of de re belief by taking action as prior to belief. This is all by way of adding credence to the claim that action is prior to belief.

But, to some extent, this is speculation. A full defense of this point would require more discussion of the philosophy of action than is contained in the remainder of this essay. The point is only that the argumentative strategy outlined above, if successful, should be grounds enough taking seriously the idea that we should reverse the usual explanatory priorities of mental states like belief and intentional action.

9.2 Four sources of circularity

It is hardly an overstatement to say that the account of belief given in the last chapter is threatened with circularity from every direction. Above I listed four of these: the account employed dispositions to perform certain kinds of *action*, at least one of which presupposes facts about linguistic *meaning*, *understanding*, and *sincerity*.[3] The purpose of this section is to indicate some reasons for thinking that the only pressing problem of circularity lies in the account of action. If action can be regarded as prior to and constitutive of belief, no special problems arise with meaning, understanding, or sincerity.

Meaning

There are at least two kinds of worry about an account of belief that makes use of facts about linguistic meaning. The first stems from allegiance to certain kinds of foundational accounts of linguistic meaning. On several influential accounts of the nature of linguistic meaning, after all, facts about meaning are constituted either by correlations between the utterances of agents and their beliefs,[4] or by conventions governing such correlations.[5] The second stems from a more general worry about how any account of linguistic meaning which did *not* make use of facts about the beliefs of agents (or their intentions, desires, and other mental states) could ever succeed; how could we possibly give a constitutive account of linguistic meaning in terms of something so sparse as the sounds emitted by speakers of a language on certain occasions?

I have already responded to the first of these worries. The charge of circularity on the basis of accounts of linguistic meaning in terms of the beliefs and intentions of agents is worrying only insofar as those accounts are plausible. And I have already argued that neither the intentions nor the beliefs of speakers can serve as the foundation for a constitutive account of linguistic meaning.[6]

As for the second of these worries, some of the despair at giving an account of linguistic meaning which does not make use of the beliefs or intentions of the relevant language-users stems from the idea that our only alternative is to make use of facts, described in an austere physical language, about patterns of sounds emitted by speakers of a language. But, simply as a strategic point, there is no reason to go so far in the context of the present theory. After all, the account of belief given above already presupposes facts about a kind of *action* an agent may perform with a sentence: the act of accepting a sentence. But if we are already making use of facts about acceptance of sentences, there is nothing to be lost by also making use of facts about what sentences agents assertively utter, or perform other sorts of linguistic

[3] The account also presupposes facts about the contents of the perceptions of agents; as noted above, I ignore this for the time being, since it is natural to think that perceptual content should be regarded as explanatorily prior to the contents of mental states like beliefs.

A less obvious but more troublesome source of circularity comes from the notion of a linguistic community, which was also tacitly presupposed by the account of the preceding chapter. As noted in Chapter 6, though, we need some account of the individuation of public languages, and there is some reason to think that it will be given partly in terms of the beliefs or expectations of language-users.

[4] For a possible example, see Ramsey (1927).

[5] See Lewis (1969).

[6] See, respectively, Chapters 2 and 3.

acts with. These facts seem no more (and no less) problematic than facts about acceptance of sentences. And, while defense of the claim is far beyond this essay, it may not be unreasonable to think that we could give an account of what it is for an expression of a language to have a certain meaning in terms of situations in which agents assertively utter, or accept, or perform other actions with sentences involving that expression.

The plausibility of this claim may be illustrated by noting that the meanings of expressions in linguistic communities seem to supervene on the dispositions to action of members of those communities.[7] It seems to be a necessary truth that any two communities alike with respect to their dispositions to perform certain acts with sentences of their language will speak languages with identical semantics. Indeed, we can put this point about supervenience in terms of individual expressions: necessarily, if the dispositions of speakers of a linguistic community are (and always have been) alike with respect to two expressions x and y, then x and y have the same meaning in that community.

This is hardly enough to deliver a constitutive account of linguistic meaning; this supervenience claim gives sufficient conditions for two expressions meaning the same, but doesn't give either necessary or sufficient conditions for an expression to have a given meaning. But it is enough to encourage confidence that an account of public language meaning in terms of use — that is, in terms of the dispositions of language users to perform certain intentional actions with a given expression in particular contexts — is at least as plausible as the individualist alternatives.[8]

Understanding

What is it for an agent to understand an expression? The requirements we set for understanding seem to vary more from situation to situation than is the case even with most mental states; one might well say of an undergraduate student of philosophy that he really does not understand the meaning of "a priori knowledge," but still be willing to credit him with beliefs as a result of his asserting sentences involving this phrase. The sense of understanding with which we are interested here is a fairly minimal one; to understand a sentence in this sense is to be in a position to acquire a belief by assenting to the sentence.

[7] Strictly, the supervenience base would have to include not only these dispositions to action, but also facts about the environment of the speakers of the language. But such facts — which will include, most obviously, facts such as the chemical constitution of the liquid which falls from the sky and is in lakes and rivers — raise no worries about circularity. I ignore this complication in what follows.

[8] For discussion of these alternatives, see Chapters 2, 3, and 6. For discussion of an account of public language meaning in terms of use, see Brandom (1994).

There are of course a number of problems with this thought; a particularly incisive one has been noticed by Robert Stalnaker. On the present account of belief, the semantic contents of sentences in contexts are taken for granted. On the standard picture of semantic content, the contents of sentences in contexts are determined by the sentences character along with the context in question; so it looks as though the kind of use theory of meaning envisaged above will not be an account of content, but of character. The problem is that in many cases it is tempting to count the mental states of speakers as parts of the context. But, if one does this, then the contents of mental states are among the determinants of content; hence an explanation of those mental states in terms of the contents of sentences will be circular. See Stalnaker (1984), 40. I am inclined to hope that this problem can be solved by 'building up' to the relevant intentions from mastery of other linguistic items.

Further problems of the same general sort are raised by utterances of ambiguous sentences.

A good place to begin an analysis of understanding in this minimal sense is with Kripke's examples of agents in various states of ignorance about the referent of a proper name who nonetheless count as competent users of the name in the sense that they can acquire beliefs about the referent by accepting sentences involving the name. As regards their beliefs about the referent, these agents need have very little in common. Rather, it seems, we count them all as understanding the name in question because (i) they are members of a linguistic community in which the name has a meaning, (ii) they were introduced to the name by another member of the community competent with the name, and (iii) they satisfy certain minimal standards regarding the use of the name; for example, they use the name in sentences in ways appropriate to its grammatical category. Perhaps (iv) they must also be amenable to correction in the use of the name by other members of the community, at least in the early stages of their acquisition of the name. More needs to be said about what these constraints amount to, and about whether other constraints might be needed. But the key point at present is that, so long as we help ourselves to full-blooded rather than austere descriptions of the actions undertaken by the agent with respect to the name, none of (i)-(iv) seem to presuppose facts about the beliefs of agents.[9]

If this point applies not just to proper names, but to linguistic expressions more generally — as seems plausible — then this model might provide the beginnings of an account of understanding as at root a practical ability, rather than as a set of beliefs about linguistic expressions.

Sincerity

Sincerity may seem to be a more difficult case. What could be more natural than to say that sincere acceptance is constituted simply by believing what one is saying?

Some indication that this account of what it is to be sincere is on the wrong track is provided by Wittgenstein's question,

> Why can't a dog simulate pain? Is he too honest? Could one teach a dog to simulate pain? Perhaps it is possible to teach him to howl on particular occasions

[9]One might object that to be competent with a name "N" which refers to an individual N, one must at the least *believe* that "N" refers to N. This seems right; but it need not make the account circular. For it may be that this kind of meta-linguistic knowledge of the meanings of words is explained by understanding of the words. On this view, anyone who satisfied (i)-(iv) would thereby also be counted as knowing that "N" refers to N.

It is also worth noting that, since (i)-(iv) give an account of understanding an expression partly in terms of the understanding of others, we would need a separate account of competence with a name one has introduced oneself. But this is more a problem for a theory of meaning than a theory of understanding; presumably, if one has succeeded in introducing a meaningful term, one is thereby a competent user of that term.

The importance of a more difficult objection has been pointed out to me by Gideon Rosen: an account of understanding in the context of a communitarian account of thought and language will have to make room not only for standard cases of understanding and misunderstanding, but also cases in which — without intentionally doing so — a speaker misunderstands an expression in such a way as to invest it with a meaning for her which differs from its meaning in the linguistic community of which she is a member. (Think of a child who calls everything purple "yellow.") I do not know what should be said about such cases.

> as if he were in pain, even when he is not. But the surroundings which are necessary for this behaviour to be real simulation are missing.[10]

We can reformulate Wittgenstein's question to suit our purposes as, "Why can't a dog be insincere?" Wittgenstein's answer is that the conditions necessary for insincerity are missing in the case of a dog. This is hardly a precise account of sincerity; but the basic idea — that insincerity requires certain background conditions — is an important one. It is important for our purposes because it suggests that sincerity is a kind of default condition, and that the natural order of explanation is to begin with an account of what it is for an agent to be insincere.

This might not look like much of an advance. But if we think of insincerity as the more fundamental notion, the problem of circularity posed by sincerity no longer looks insoluble. For it seems as though cases of insincerity fall into two classes, neither of which need pose problems of principle for the kind of communitarianism being developed.

The first kind of case of insincerity is what we might call *self-conscious insincerity*. This includes typical cases of lying, and is defined as the utterance of a sentence which the speaker believes that he does not believe. Of course, this too looks circular, since it defines a kind of insincerity in terms of a certain belief of an agent about his own utterance. But this can be restated in a noncircular way as follows: an agent is self-consciously insincere with respect to an utterance x just in case she is disposed to accept some sentence which means that she does not believe what x says. As Wittgenstein points out, the conditions necessary for insincerity are lacking in the case of non-linguistic creatures; so here we can just identify the beliefs definitive of self-conscious sincerity with linguistic dispositions.[11]

But this doesn't exhaust cases of insincerity; there are also cases of *unconscious insincerity*. Take someone who professes to believe certain religious views, but never acts in accord with any of them. Does he really believe the views, or is his profession insincere? Sometimes we want to say that the agent has inconsistent beliefs: that he believes both the religious views and propositions which entail their falsity. But other times we want to say that he simply does not believe the religious views, and that his acceptance of the claims was insincere. There seems to be a kind of continuum here; and the axis of the continuum seems to be how well the actions of the agent fit with his desires and his professed beliefs. If the evidence for the desires is strong enough, and the conflict between the desires, beliefs, and actions great enough, we will be inclined to count the agent as insincere in her utterances. This is roughly the account I sketched above:[12] the belief-desire-action principle states a general but defeasible rule connecting the beliefs, desires, and actions of agents. To be sure, the details

[10] *Philosophical Investigations*, §250.

[11] Doesn't this lead to a vicious regress, since the agent might be insincere with respect to the higher-order sentence as well? I don't think so. Agents will only have so many higher-order sentences about sentences in mind, so the regress will come to a natural end.

[12] See 8.5.2, "Linguistic beliefs and the explanation of non-linguistic behavior" pp. 180 ff.

Objection: Consider an agent who speaks a fairly impoverished language in which there are no words for 'believes', 'says', 'true' or 'lying.' Then such a creature cannot be insincere in the first way. But the second kind of insincerity only seems to apply to utterances of sentences which have a meaning which bears some fairly direct connection to actions. So consider some sentence S which is not of this sort; if the agent utters S, she cannot be insincere in either of the two ways listed above. So now suppose that she utters a sentence S out loud, and five seconds later says the sentence $\neg S$ to herself. Surely, one of these must have been insincere!

of this account are unclear at this stage; but there is no 'in principle' problem of circularity here.

Action

In two senses, the most fundamental level of the challenge of circularity is at the level of intentional action. First, as we have seen, there is some reason to think that various notions which seem to pose problems of circularity for the communitarian may be explicated in terms of the actions and dispositions to actions of agents. I suggested that we might be able to give an account of public language meaning in terms of the dispositions of agents to use (assert, ask, exclaim) sentences in various contexts; that we might be able to give an account of understanding in terms of possessing dispositions to use expressions in ways similar to others in your community; and that we might be able to explain sincerity in terms of the absence of dispositions to (among others) accept sentences which say that one does not believe a certain utterance.

But the problem of circularity in the case of intentional action is also the most fundamental in that it is the most difficult level of the challenge. The next section is devoted to addressing this issue, and providing some *prima facie* arguments for the conclusion that action is prior to belief.

9.3 Action as prior to belief

9.3.1 Functionalism and the priority of action

The first argument for the priority of action is directed at someone who objects to 'interdependence' accounts of mental notions, and uses this as part of an argument against the account of belief sketched in the preceding chapter. Such a philosopher might argue as follows:

> (i) Actions are constituted by the beliefs, desires, and intentions of agents. (ii) Circular constitutive accounts are inadmissible; it cannot be the case both that A's are constituted by B's, and that B's are constituted by A's. (iii) Any plausible behaviorist communitarianism account will take beliefs to be constituted by intentional actions (rather than mere bodily movements). (iv) So behaviorism is false and, contra the arguments of Chapters 4 and 5, some version of functionalism must be true.

As noted in the preceding chapter, I accept (iii). But I have argued that (iv) is false; so I must reject either (i) or (ii).

Premise (ii) is, in a way, fairly easy to reject. The point I want to make here is that, if (ii) is correct, we can use it in an argument against (i). This is because, I claim, any functionalist

Reply: Why think that one of these utterances must have been insincere? I suggest that what one is imagining in this case is the agent uttering S out loud while thinking of her utterance as false, and then saying its negation to herself with a feeling of certainty. In other words, one is imagining the agent to have all of the complex thoughts about the truth and falsity of utterances without having words for these thoughts. But this kind of separability of the linguistic capacities of creatures from their capacities to have certain thoughts is just what the communitarian denies; agents without the capacity to express certain kinds of thoughts would also lack the capacity for certain kinds of insincerity.

account of belief will take beliefs to be constituted in part by the intentional actions of agents. If I am correct that functionalism and the kind of neo-behaviorism I have been defending are the only two plausible ways of giving a constitutive account of belief, this will be enough to show that the conjunction of (ii) with *any* constitutive account of belief will be enough to entail that intentional action is prior to belief in the order of explanation.

Recall that a functionalist account of belief must say which properties of internal states make such states belief states, and which properties determine the contents of those states. The point to be shown is that every attempt to state such properties ends up using properties of causing intentional actions with certain characteristics. Suppose (for *reductio* that one tries to construct an account of functionalism without this characteristic. A natural first thought is that we should try to give an account of belief in terms of causal relations between the external world and the states in question; such an account might then have no need of facts about intentional actions. The proposed functionalist account might then be of the form

> x is a belief state with content $p \equiv x$ bears the right causal relation to the fact p

But recall Stalnaker's objection to this sort of account of belief. Of such a purely causal account, he wrote

> ... if a bald head is shiny enough to reflect some features of its environment, then the states of that head might be described in terms of a kind of indication — in terms of a relation between the person owning the head and a proposition. But no one would be tempted to call such states belief states. ...
>
> Beliefs have determinate content because of their presumed causal connections with the world. Beliefs are *beliefs* rather than some other representational state, because of their connection, through desire, with action.[13]

Stalnaker's point here seems to me decisive. Even a purely causal account of content will need some account of which states which bear the requisite causal relation to the world are belief states; and it is hard to see how such an account could be given, if not in terms of the behavioral output of the relevant agents.

This is not, however, quite enough to show that any plausible version of functionalism will have to take beliefs to be constituted by intentional actions. So far I have argued that any plausible version of functionalism will have to make some use of facts about the behavior of agents; but so far it is unclear whether this 'behavior' will be the intentional actions of agents, or whether we might give the account in terms of the dispositions to perform certain bodily movements of agents.

But it seems fairly clear that the relevant sense of 'behavior' will be the intentional actions of agents. The point of the appeal to behavior was to rule out some things as states of rocks from counting as beliefs; but to rule these out, we need to appeal to facts about intentional action. Again, the point can be illustrated using Stalnaker's causal-pragmatic account of belief. According to Stalnaker, a state which indicates p will be the belief p if the agent in question is disposed to act so as to satisfy her desires in a world in which p and the agents other beliefs are true. But consider a rock, the surface of which indicates the temperature of the surrounding air. If a wind comes along and blows the rock a bit to the left into some

[13]Stalnaker (1984), 15-19.

shade, there is no obvious way to block the unwanted result that the rock believes that the air is at such-and-such temperature, and desires to be in cooler air. Less fanciful examples can be generated using examples of bodily movements of human agents. The functionalist should not want the behavioral constraints on having certain beliefs to be satisfied by, e.g., facial tics, spasmodic movements, or unconscious generation of sounds.

The modest conclusion to be drawn from this argument is that one cannot use a premise like (ii) to argue against behaviorism, since it's also an argument against functionalism and, on the present broad construal of functionalism, there is no third option. The stronger result is that if you grant (ii), and if you think that there is some true constitutive account of belief (whether functionalist or behaviorist), you must also think that intentional action is prior to and constitutive of belief. So belief-desire-intention accounts of action must be wrong.

9.3.2 De re belief and de re action

A second sort of argument for the priority of action over belief can be generated from a few plausible principles concerning the conditions agents must satisfy in order to have de re beliefs about objects. The principles I have in mind, and discuss below, are all proposals due to Jim Pryor.[14]

Intuitively, it seems that certain kinds of evidence concerning an object o can put an agent in a position to have de re beliefs concerning o.[15] Call these kinds of evidence *introductory evidence*. Certain perceptual experiences will surely count as introductory evidence, as will, on the present view, coming to be competent with a singular term which licenses exportation from the complements of propositional attitude ascriptions. Concerning this introductory evidence, we can then state the following plausible *non-emergence principle*: only introductory evidence with respect to an object o can put an agent in position to have de re beliefs about o; no amount of non-introductory evidence, such as descriptive information about the object, can do so.[16]

But this non-emergence principle seems to conflict with another plausible thesis about de re belief. Intuitively, just as we can distinguish between beliefs which are and are not de re, we can distinguish between actions which are and are not de re. Consider the action of touching something. We can distinguish between, so to speak, touching an object under a description — as might happen if I try to touch *whatever is to my left* — and touching an object de re — as happens when there is some object such that I am trying to touch *it*. With this in mind, consider the following example:

> I am standing in a classroom facing away from the chalkboard; a reliable friend has told me that there is a square on the chalkboard just behind my left shoulder.

[14] He should not be taken to endorse these principles, and still less to endorse the use to which I put them here.

[15] Roughly, I take de re beliefs to be those beliefs which are truly attributed by de re belief ascriptions: those ascriptions which are such that the inference from $\ulcorner a$ believes that t is $F\urcorner$ to $\ulcorner a$ believes of t that it is $F\urcorner$ is valid.

[16] This is a step to which latitudinarians, who think that (under some conditions) descriptive information about an object is sufficient for having de re thoughts about the object. I am not sure if the argument which follows could be reformulated in a way which would be convincing to a latitudinarian.

> I might try to touch this square by pointing my finger over my shoulder, and very slowly inching back towards the chalkboard.

I claim that, before touching the chalkboard, I am able to have de re beliefs about the square on the chalkboard toward which my finger is inching. This is partly because I am trying to perform a de re action with respect to that shape: there is a shape on the board such that I am trying to touch *it*. But I know that I am trying to perform this de re action; there is a square on the board such that I know that I am trying to touch it. This conflicts with the non-emergence principle stated above. I have heard about the square on the board only under a description; so, because I have had no perceptual contact with the square prior to touching it, and have not come to be competent with a singular term of my language of the relevant sort, I have no introductory evidence with respect to the square. No amount of inching toward the board should be enough to remedy this lack.

One interpretation of what's going on here is that my trying to perform a de re action with respect to the square puts me in a position to have de re thoughts about it; on this interpretation, until I tried to touch the square, I would have been unable to have de re thoughts about it.[17] On this interpretation, the intuition in the chalkboard case can be made consistent with the non-emergence principle by supplementing the list of kinds of evidence which can be introductory with respect to an object. In particular, we might recognize a species of *agential justification*, to the effect that my trying to perform some action gives me (defeasible) evidence for believing that I am performing that action. If this sort of agential justification exists, then it seems that we should admit acts of trying to perform certain kinds of actions as introductory evidence. In particular, trying to perform a de re action will provide agential justification for — and hence, presumably, put one in a position to entertain — the belief that one is trying to perform that de re action. But, as we saw above, this will be a de re belief.

This view of de re belief, which involves the non-emergence principle along with three recognized kinds of introductory evidence — perception, competence with a singular term of the relevant sort, and trying to perform a de re action — fits well with the picture of intentionality I have been developing, according to which public languages (including singular terms of the relevant sort), can be vehicles of thought, and perception and action are the two fundamental sources of intentionality. And, more to the present purpose, this view of de re belief also provides support for the thesis that intentional action is prior to belief in the order of explanation. For consider the de re action I performed in trying to touch the square on the chalkboard; this case poses a challenge to the view that actions are constituted by, among other things, the beliefs of agents. What belief could play the relevant role here? Presumably no purely descriptive belief can be constitutive of performing a de re action; and we cannot appeal to the belief that I am trying to perform the action, since I have that belief in virtue of pursuing the intentional action of trying to touch the square. And, by hypothesis, I have no other de re beliefs about the square. Hence the intentional actions of agents are not constituted by their beliefs. This counts against the view that the present account is circular, and in favor of the view that the order of explanation I have been pursuing is the correct one.

But there is another interpretation of the chalkboard case available. Above I suggested that my trying to touch the square was what put me in a position to have de re thoughts

[17] As we'll see below, this isn't the only plausible interpretation of the case.

about the square. But it is not implausible to think that my actually trying to touch the square was not necessary. On this view, what matters is that the square was, in Pryor's phrase, in my *volitional field*: that is, what matters is that I am in a position to perform de re actions with respect to it, whether or not I try to perform any of them. In support of this interpretation, one might note that it is intuitively right to say that, before trying to touch the square, I was in a position to *wonder whether I could touch it*; and this seems to be a de re attitude toward the square.

There is no straightforward argument from this interpretation of the case to the priority of action; but perhaps there is still something to be taken from it. For how could the case under this interpretation be made consistent with the framework of giving the conditions for de re thought in terms of introductory evidence and the non-emergence principle? A natural thought is that we should include among the kinds of introductory evidence which can put an agent in position to have de re thoughts about an object a certain kind of descriptive information about the object: evidence that the object is in one's volitional field. But then we need some explanation of why this sort of descriptive information, but not all kinds, can play this role.[18] And again it seems natural to say that this kind of descriptive information can play this special role because actions, and abilities to perform actions, have a special foundational role to play relative to beliefs. What else could explain the special character of descriptive evidence having to do with the actions one is able to perform?

Like the argument of the preceding section, the case of de re action and de re belief does not provide an argument from premises that everyone will accept to the priority of action. But, also like that argument, it shows that matters with respect to action and belief are not so simple as the standard view of propositional attitudes as prior to and constitutive of intentional action would have it. Given certain theoretical commitments (other than the kind of communitarianism I have been defending), the thesis of the priority of intentional action over belief is a natural one.[19]

[18] Again, this argument has no bite against the latitudinarian, since she thinks that, given the right stage-setting, any descriptive information can play this role.

[19] Is there anything positive to be said about what it is perform an intentional action of a certain type? An example from Wittgenstein which we discussed above might point in the direction of a positive account of intentional action: "Let us imagine a god creating a country instantaneously in the middle of the wilderness, which exists for two minutes and is an exact reproduction of a part of England, with everything that is going on there for two minutes. Just like those in England, the people are pursuing a variety of occupations. Children are in school. Some people are doing mathematics. Now let us contemplate the activity of some human being during these two minutes. One of these people is doing exactly what a mathematician in England is doing, who is just doing a calculation.—Ought we to say that this two-minute man is calculating? Could we for example not imagine a past and a continuation of these two minutes, which would make us call the processes something quite different?" (*Remarks on the Foundations of Mathematics*, Part VI, §34). Above in §6.3 I adapted this quote to serve as an argument against individualist views of public language meaning. But Wittgenstein's concern here is with intentional action rather than with meaning; and perhaps his thought here points in the direction of an interesting analogy between the two.

The analogy would be this: bodily movements stand to intentional actions as words stand to their meanings. Just as the social history of a community fixes the meanings of the linguistic expressions it uses in communication, the history of that community determines which bodily movements (in which contexts) count as which intentional action-types. Just as we can come to be competent with a system of linguistic expressions, and use that system as vehicle for thought, so we can come to be

9.4 Consequences of communitarianism

So far in this essay I have argued at length against two pictures of the relationship between mind and language, and sketched and more briefly defended a third. If this third picture is correct, there are obvious consequences for our views of the relationship between mind, language, and action. But there are also a number of less obvious consequences of the transition from an individualist to a communitarian view of intentionality.

9.4.1 Language in thought and in communication

There are two main positions on the relative roles of language in thought and in communication. First, one may regard the role of language in communication as its primary, and perhaps only, function. This is, as Dummett puts it, a view of language as a kind of *code*: facts about the contents of mental states are constituted quite independently of language, and language is simply the medium in which we encode these thoughts in order to transmit them to each other in communication.[20] A second view of the relationship between the roles of language in thought and in communication takes its role in thought to be primary; according to this view, the meanings of linguistic items in the language of an agent acquire meaning through their role in the thought of that agent. On this view, languages are primarily *private languages*: languages of individual agents. Expressions in these private languages are then assigned contents on the basis of their role in the thoughts of the relevant agents; insofar as public languages are recognized at all, they are to be understood as derivative from these private languages.

We've now seen some reason to distrust both of these views. The code conception of language is threatened by the idea that the contents of the beliefs and other mental states of agents are often constituted by the meanings of linguistic items; the private language view is threatened by the fact that the linguistic items which play this role are expressions of public, rather than of individual, languages. On the view of language and thought encouraged by the view of belief sketched above, the code conception is correct in taking the primary sorts of languages to be public languages, and is correct in taking the meanings of expressions in these languages to be constituted by their role in communication; the private language view is correct in taking language to play a role in constituting the contents of the mental states of individuals. The meanings of linguistic expressions are fixed by their communicative uses in a community, and these meanings — rather than the meanings of expressions in private languages — play a role in determining the contents of the thoughts of agents.

9.4.2 Skepticism about belief

The main alternative to the view that beliefs are constituted by dispositions to perform certain actions is that they are constituted by the second-order properties of internal states

competent with a system of correlations between bodily movements and actions employed by one's community, and use that system as a vehicle for thought. There are a number of obvious problems here, not the least of which is that it is hard to describe the analogy without using the language of intentional action to describe the process of coming to be competent with a community's correlations of bodily movements with actions; but the analogy strikes me as an interesting one.

[20]See Dummett (1985), 149-151.

of agents. This functionalist view of belief opens the door to a new kind of skepticism about the attribution of beliefs even to adult human subjects. For suppose that we give a theory of belief along broadly functionalist lines; such a theory will say that what it is for an agent to believe p is for that agent to be in some internal state x such that, for some relation R defined by the theory, $R(x, p)$. Such a theory then makes a specific claim about the structure the internal states of an organism must have in order to qualify that organism as one capable of having beliefs; in particular, it claims that the internal states of human beings must have such a structure in order for any of our attributions of beliefs to human beings to be true. If psychological research then shows that the internal states of human beings lack this sort of structure, it will have shown that, contrary to what we thought all along, none of us believes anything.[21]

It is, I think, difficult to believe that psychological research should convince any of us that we have no beliefs. The present view of belief provides one way of avoiding the conclusion that it could;[22] on this view, what it is for an adult human to have a belief is for that agent to be disposed to accept a sentence, rather than for that agent to have internal states with certain second-order properties. Hence a skeptical view of the existence of beliefs would have to be grounded in an argument that no human being is disposed to accept any sentence, or perform any action; and it is hard to see how psychological research could provide such an argument.

9.4.3 The relationship between attitude and content

Many ascriptions of mental states to agents are ascriptions of propositional attitudes. Such ascriptions say of an agent that she bears some attitude — e.g., *believing*, *forgetting*, *imagining*, *thinking*, *supposing* — to a proposition. One standard way to give an account of the nature of a propositional attitude R is to say that what it is for an agent a to bear R to a proposition p is for that agent to bear some relation R' to an internal representation x, which *has the content* p.[23]

On this strategy for giving an account of the nature of propositional attitudes, the tasks of giving a theory of content and a theory of a given attitude may be pursued more or less independently. The former task is a matter of giving an account of what determines the contents of internal states; the latter is a matter of giving an account of what sort of relation one must bear to an internal state in order to be counted as bearing the relevant attitude to the proposition which is the content of that state.

If, one the other hand, one takes the propositional attitudes of agents to be constituted by their dispositions to perform certain sorts of actions, this division of labor breaks down. For, on this picture, there is no such thing as a theory of mental content simpliciter. There are instead a series of questions, each particular to a given mental state type: What is it for an agent to believe p? What is it for an agent to entertain p? What is it for an agent to remember

[21] For an example of this sort of view of what belief requires, and an argument that psychological research may well show that none of us have any beliefs, see Ramsey et al. (1996).

[22] It is not, however, the only way to avoid this conclusion; one could also claim that, because we are more sure that we have beliefs than we are of any particular theory of the nature of belief, evidence that we fail to satisfy the conditions for belief set by a theory should be taken as grounds for rejecting the theory rather than as grounds for accepting the skeptical conclusion.

[23] See, e.g., Field (1978).

p? There may well be connections between these questions; but there is no reason to think that our answers to them should be given in terms of a series of different relations to a single sort of content-bearing internal state. Rather, each sort of attitude may be constituted by different sorts of dispositions, and there may be no one thing held in common between each. If so, then the project of giving a theory of content *simpliciter* is radically misconceived.

9.4.4 Philosophical anthropology

The task of philosophical anthropology is to say what sorts of things human beings are. One may approach this task either by asking a single overarching question — What is a human being? — or by asking about the nature of various activities which seem central to being human.

If the kind of communitarianism I have been defending is correct, its most important consequence is the contribution it makes to the second approach to philosophical anthropology. For one fundamental question in philosophical anthropology concerns the relation, with respect to some sphere of activity, of the relationship between human beings and the social groups they comprise. Communitarianism suggests that, at least in the sphere of the mental, we should reject the traditional view that societies are best understood as aggregates of more of less independently constituted units. In its place, communitarianism suggests that social facts have a role to play in the constitution of thought.

APPENDICES

Appendix A

Deception and self-referential intentions

In the discussion of Grice's views of speaker-meaning in §2.2 above, I used a formulation of the account which involves self-referential intentions, in the sense explained in the text. This is not, however, the version of Grice's account most often discussed. Instead, theorists focus on the following account, which seems to have been the view that Grice held in the early 1960's, after the publication of meaning but before Strawson's "Intention and Communication in Speech Acts" of 1964:

[G*] a means p by uttering $x \equiv a$ intends that
(1) his audience come to believe p,
(2) his audience recognize that a intends (1), &
(3) (1) occur on the basis of (2)

So I owe some explanation of why I used version [G] instead of this more prominent account. The explanation comes in an short and a long version. The short version is that the formulation in terms of self-referential intentions does what no other version of the Gricean account does: avoid a class of counterexamples involving deceptive utterances first raised by Strawson,[1] and repeated in many different forms since. The long version will occupy the remainder of this appendix. I shall, by examination of suggested revisions to [G*], argue that there is no way to avoid the problem posed by Strawson's cases of deception other than via self-referential intentions; and, in the closing pages, defend Gilbert Harman's view that there is nothing paradoxical or otherwise troublesome about such intentions.

The simplest cases of deception are those in which an agent a has intentions (1), (2), and (3), but in addition has a deceptive intention with the following content:

[D] that his audience not recognize that a intends that his audience recognize that a intends (1) — i.e., that his audience not recognize (2)

[1] Strawson (1964), 446-7.

The following case[2] is an illustration of such a deceptive intention: a intends that his audience b should believe that the house b is about to buy is infested with rats. So a lets loose a rat in the house when he knows that b is watching him, and knows further that b believes that a does not know this. a then intends b to reason as follows: "My friend is setting loose a rat in the house. He surely expects me to see the rat, though he does not expect me to know that he set it loose. Why would he do this? He knows that if I see a rat running around the house, I will come to believe that the house is rat-infested. So he must intend that I come to this belief. I trust my friend; hence his having this intention is enough for me to believe that the house is rat-infested."

So a has each of the intentions (1)-(3): he intends (1) that b believe that the house is rat-infested; he intends (2) that b recognize intention (1), since he knows that b is watching him set loose the rat and expects b to infer from this that he intends that he should believe that the house is rat-infested; and he intends (3) that the first of these intentions be satisfied in the basis of the satisfaction of the second. Despite the facts that a has each of these intentions, though, it seems wrong to say that a meant, by setting loose the rat, that the house was infested with rats. And the reason why it seems wrong to say this is that, in addition to intentions (1)-(3), a also has deceptive intention [D]: he intends that b not recognize intention (2).

The easiest way to rule out this sort of counterexample would be to add another clause to [G*], requiring that

(4) a intends that his audience recognize intention (2).[3]

But this leaves open the possibility of more recondite cases of deception, in which an agent has intentions (1)-(4), but also has the following deceptive intention:

[D′] that his audience not recognize that a intends that his audience recognize intention (2) — i.e., that his audience not recognize (4).[4]

Furthermore, were we to add a clause (5) requiring that a intend that his audience recognize the intention described by (4), we could generate a new deceptive intention, [D″], to the effect that a intends that his audience not recognize intention (5). The fact that cases of deception can be made successively more complex to avoid revisions like (4) or (5) seems to lead to the need to attribute to agents an infinite number of increasingly complex intentions.[5]

A variety of expansions of [G*] have been suggested to avoid this consequence, none of which are altogether satisfactory. The strategy originally suggested by Grice, and modified many times by subsequent writers on the issue, is that we solve the problem "by requiring U (the speaker) *not* to have a certain sort of intention or complex of intentions."[6] Incorporating

[2]The case is from Schiffer (1972), 17-18, and adapted from a version given in Strawson (1964).

[3] Strawson (1964) suggests a revision along these lines, but expresses caution — justifiably, as it turns out — about whether it will be enough to block all cases of deception.

[4]This and more complicated cases of deception are described in Schiffer (1972), 18 ff.

[5]Grice (1969) notes correctly that only a few clauses like (4) and (5) would be enough to rule out all deceptive intentions simple enough to be humanly possible (99). But as long as the aim is to give metaphysically necessary and sufficient conditions (or, as Grice has it, "logically necessary and sufficient conditions") for speaker-meaning, this contingent fact about human beings is no help here.

[6]Grice (1969), 99. Grice proposes a slightly different solution in his (1980), though I am not

this suggestion involves adding to our list of the conditions a speaker must satisfy to mean something by an utterance a negative requirement such as the following:

> [N] there is no proposition q such that a uttered x intending both (i) that his audience's coming to believe p should rely on their believing q and (ii) that his audience should think that a intends (i) to be false[7]

One advantage of this addition is that it avoids attributing yet another intention to speakers; the more complex the Gricean analysis of speaker-meaning becomes, the less plausible it is to say that speakers must have all of the intentions it specifies in order to mean something by an utterance.

But, as Schiffer notes, the addition of [N] is not sufficient to block cases of deception.[8] This can be shown by slightly modifying Strawson's example of the rat-infested house, discussed above. It was part of this example that a intended that b believe that a intends b not to rely on the belief that a set the rat loose in coming to believe that the house is rat-infested; and clause (ii) of [N] rules this case out. But the example is no less effective if, instead of having a intend that b believe this, we have a be *indifferent* as to whether b believes that a intends b to rely on the belief that a set the rat loose in coming to believe that the house is rat-infested. And this sort of case is not ruled out by [N]: [N] rules out cases of deceptive higher-order intentions, but some cases of deception, such as the one just described, turn on the speaker simply failing to have the requisite non-deceptive higher-order intentions.[9]

Schiffer's own suggestion introduces the notion of mutual knowledge*. In Schiffer's terms, a and b mutually know* p if and only if a knows p, b knows that a knows p, a knows that b knows that a knows p, and so on and vice versa. Schiffer's solution to cases of deception

---

entirely clear what it is.

[7] Grice (1969) puts the condition as follows: there is no inference-element E such that a uttered x intending both (i) that his audience's coming to believe p should rely on E and (ii) that his audience should think that a intends (i) to be false. The switch in the text from talk of inference-elements to talk of beliefs seems to me to lose none of the advantages of Grice's suggestion, while avoiding the introduction of a new technical term.

[8] Schiffer (1972), 26. A different counterexample of the same sort is suggested in Christensen (1997), 504-5; Christensen's counterexample, however, seems too plausibly regarded as a genuine case of speaker-meaning to be an effective counterexample.

[9] Variants of Grice's strategy for blocking cases of deception have been proposed by Recanati (1986), Neale (1992), and Bennett (1976). Recanati's version requires the speaker to have no intentions inconsistent with any of the infinitely many intentions that would be required to block cases of deception; this requirement fails to block the above counterexample for the same reason as does [N]. Neale attempts to block cases of deception by requiring that the speaker not intend that his audience be deceived about intentions (1) or (2); and this too fails to solve the variant of Strawson's example discussed above. Bennett requires that the speaker expect the audience to have no *beliefs* which would entail that any of the infinitely many intentions in the series are not realized. This too fails to rule out the above counterexample and, in addition, faces the following problem, pointed out by Harman (1977). Harman notes that Bennett himself believes that the 100th member of the expansion of this series of intentions "does not correspond to any humanly possible state of mind" (Bennett (1976), 126); and this belief entails that some of the intentions in the series are not realized by any human being. Hence, by Bennett's requirement, "no one who is aware of his opinions in this matter could ever mean anything in talking to him" (Harman (1977), 422).

is considerably more complicated than that of Grice; a paraphrase of his solution using the notation from [G*] above is as follows:

> a means p by uttering $x \equiv a$ utters x intending thereby to realize a certain state of affairs E which is (intended by a to be) such that the obtainment of E is sufficient for a and his audience mutually knowing* that E obtains and that E is conclusive evidence that a uttered x intending
>
> (0) that E obtain,
> (1) that his audience believe p,
> (2) that his audience recognize that a intends (1), &
> (3) that (1) occur on the basis of (2).[10]

As Harman has noted, there is an ambiguity in this formulation according to whether we take seriously Schiffer's parenthetical remark "intended by a to be."[11] If we do, and place what follows within the scope of the intention attributed to a, then we arrive at the following interpretation of Schiffer's account, which gives the speaker's intention wide-scope:

> [WS] a means p by uttering $x \equiv a$ uttered x intending
> (a) that a certain state of affairs E be realized, &
> (b) that the obtainment of E be sufficient for a and his audience mutually knowing* that E obtains and that E is conclusive evidence that a uttered x intending
>
> (0) that E obtain,
> (1) that his audience believe p,
> (2) that his audience recognize that a intends (1), &
> (3) that (1) occur on the basis of (2).

Nothing seems to be lost if we remove the reference to the state of affairs E and the mention of evidence to simplify [WS] as follows:

> [WS*] a means p by uttering $x \equiv a$ intends in uttering x that he and his audience mutually know* that he intends that
>
> (1) his audience come to believe p,
> (2) his audience recognize that a intends (1), &
> (3) (1) occur on the basis of (2)

With this simplification of the wide-scope interpretation of Schiffer's proposal in hand, it is

[10] See Schiffer (1972), 39. The exact quote from Schiffer is as follows: "S meant something by (or in) uttering x iff S uttered x intending thereby to realize a certain state of affairs E which is (intended by S to be) such that the obtainment of E is sufficient for S and a certain audience a mutually knowing* (or believing*) that E obtains and that E is conclusive (very good or good) evidence that S uttered x intending (1) to produce a certain response r in a; (2) a's recognition of S's intention (1) to function as at least part of a's reason for a's response r; (3) to realize E." I've changed this quote to make it (slightly) more readable by omitting several of Schiffer's qualifications. In addition, in keeping with the topic under discussion, I have changed Schiffer's analysis from an analysis of a speaker meaning something by an utterance to a speaker meaning p by an utterance, for any proposition p; I have also made cosmetic changes in the specification of the speaker's intention, for consistency with [G*].

[11] Harman (1974), 226.

clear that it succeeds in blocking both of the kinds of cases of deception mentioned above. It blocks the original scenario imagined by Strawson, since in that case a intends that his audience *not* know that he intends (2); and [WS*] stipulates that a intend that he and his audience mutually know* this. Similarly, in the revised version of the case used against Grice's proposed solution, a is indifferent as to whether his audience recognizes that he intends (2), which again conflicts with the requirement of [WS*] that a intend that he and his audience mutually know* this.

So far, then, the wide-scope version of Schiffer's proposal seems an improvement over that of Grice. But it succeeds at the cost of considerably complicating the analysis. A general problem with the Gricean analysis concerns the complexity of the intentions attributed to speakers. But even if it is somewhat plausible that meaning something by an utterance requires the intentions specified in [G*], it seems most implausible to attribute to them intentions whose content involves the technical notion of mutual knowledge* discussed by Schiffer. It is hard to believe that most people even possess the concept of mutual knowledge*, let alone have intentions involving it every time they mean something by an utterance.

One thought is that the narrow-scope interpretation of Schiffer's biconditional, obtained by ignoring his parenthetical remark, might avoid this problem, by taking the reference to mutual knowledge* outside the scope of the speaker's intention.[12] The narrow-scope version of Schiffer's account of speaker-meaning is as follows:

[NS] a means p by uttering $x \equiv$
- (a) there is a state of affairs E such that the obtainment of E is sufficient for a and his audience mutually knowing*
  - (i) that E obtains,
  - (i) that a uttered x intending
    - (1) that his audience believe p,
    - (2) that his audience recognize that a intends (1),
    - (3) that (1) occur on the basis of (2); &
- (b) a uttered x intending that E obtain

The problem with this, as Harman notes, is that it is difficult to see what the state of affairs E could be such that speakers are, according to [NS], ever able to mean anything by utterances. One option would be to identify E with the utterance of x by a in the presence of his audience in certain conditions. This may be sufficient for a and his audience mutually knowing* (i); but it is clearly not sufficient for a and his audience mutually knowing (ii), since a's audience might well know that a has uttered x in his presence without knowing whether a intends that he believe p. Harman suggests that "if something like self-referential states of affairs are possible, E might be the conjunctive state of affairs, *S's uttering x intending (1)-(3) and S and A's mutually knowing that E obtains.*"[13] But even if this does yield the conclusion that speakers are able, by the standards of [NS], to mean things by utterances, it negates any advantage [NS] might have had over the wide-scope interpretation of Schiffer's account of speaker-meaning. To attribute to speakers the intention that such self-referential states of affairs obtain, after all, is to attribute intentions no less complex than those attributed by

[12] Harman (1974) suggests that this is the best interpretation of Schiffer's remarks.

[13] Harman (1974), 227. Harman does not endorse this account of speaker-meaning, however.

[WS*].

Of the main solutions offered to the problem posed by cases of deception, then, the only ones to meet with any success attribute to speakers intentions of such complexity that the account is robbed of its plausibility. As Harman suggests, a better solution to this class of counterexamples is to be found by returning to the analysis of speaker-meaning which Grice originally gave in his 1957 article "Meaning," and abandoned soon thereafter. Before moving to [G*], Grice advocated something very like the account [G] discussed in the main text above. Recall that this account ran as follows:

[G] a means p by uttering $x \equiv a$ intends in uttering x that
(1) his audience come to believe p,
(2) his audience recognize this intention, &
(3) (1) occur on the basis of (2)[14]

where the expression "this intention" refers to the intention whose content is given by the conjunction of (1), (2), and (3). The salient difference between [G*] and [G] is that [G], unlike [G*], attributes to speakers self-referential intentions: speakers are said to intend that their audience come to have a certain belief on the basis of their recognition of that very intention. This can be seen by the fact that whereas it is essential to [G] that a have a single conjunctive intention whose content is given by (1), (2), and (3) which is the referent of the expression "this intention" in (2), [G*] can be seen just as well as attributing to the speaker three separate intentions.

Some time between "Meaning" and his 1969 article "Utterer's Meaning and Intentions," Grice modified his analysis of speaker-meaning to avoid using these sorts of self-referential intentions in the analysans, and moved to a definition along the lines of [G*]. What is important for our purposes is that [G] avoids the cases of deception which have been raised against [G*], without the complications introduced by Schiffer's reliance on speakers having intentions involving the concept of mutual knowledge*. Clause (2) of [G*] requires that a intend that his audience recognize that a intends that his audience believe p; the cases of deception mentioned above get off the ground by supposing that an agent could satisfy this clause and yet intend that his audience not recognize that he satisfies it. Clause (2) of [G], however, requires that a intend that his audience recognize that he intends that (1) & (2) & (3); and this rules out, without need of further intentions attributed to the speaker, the possibility of an agent satisfying the conditions set by [G] while intending that his audience not recognize that he intends any of (1), (2), or (3).

If Grice's original analysis avoided one of the principle counterexamples to the version of the view he held in the 1960's, then, why did he abandon the original view? So far as I know, Grice nowhere explicitly answers this question. It seems plausible that he feared some sort of paradox of self-reference; he mentions "reflexive paradox" in "Meaning" without explaining

[14]Grice (1957), 219. In his original analysis in "Meaning", Grice gave no analysis of what it is for a speaker to mean p by uttering x, but only of what it is for a speaker to mean *something* by uttering x. He did say, though, that what the speaker meant could be discerned from the intended effect on his audience; identifying this effect with a belief whose content is identical with what the speaker meant was the first attempt to give an analysis of a speaker meaning p by uttering x Grice considered in his (1969). He came up with other versions later in that paper, and in other papers, some of which I discuss in §2.2.

exactly what he has in mind.[15] Given the problems caused by the cases of deception for versions of the Gricean account of meaning which do not make use of self-referential intentions, it is worth considering whether there is anything incoherent in the attribution of these sorts of intentions to speakers.

It seems to me that there is no incoherence here and that, for this reason, [G] should be preferred over [G*]. One way to lend credence to this claim is to argue that there are self-referential beliefs; if there is nothing incoherent about the idea of a self-referential belief, then it is hard to see why intentions should be any different.

This is the strategy pursued by Gilbert Harman in his defense of the coherence of self-referential attitudes.[16] Harman's main example concerns beliefs whose content is the proposition expressed by an instance of the Liar Paradox. These beliefs are self-referential; and in many cases it is hard to deny that people have such beliefs. Harman mentions, using examples from Kripke's "Outline of a Theory of Truth,"[17] that one can believe that another person's belief is not true; but this is enough to yield a self-referential, and a Liar-paradoxical, belief if the other person's belief is that one's own belief is true. It seems clear that Harman is correct that there can be self-referential beliefs in this sense.

What is not so clear is that this claim lends credence to the claim that intentions which are self-referential in the manner of those used in [G] are coherent.[18] In general, we can distinguish two ways in which the fact that an agent bears a certain attitude toward a proposition can be self-referential. Say that a proposition p is referential with respect to x just in case one of the constituents of p either is x or has x as its referent.[19] Then an example of a propositional attitude relation R to a proposition p of an agent a may be self-referential in either of the following two ways:

(i) aRp and p is referential with respect to the proposition p

(ii) aRp and p is referential with respect to the fact that aRp

The examples Harman gives are self-referential in sense (i): they concern an agent standing in the belief relation to a proposition which refers to itself.[20] Intentions of the sort mentioned in [G], however, are self-referential in way (ii). The phrase "this intention" in (2), after all, does not refer to the proposition expressed by the conjunction of (1), (2), and (3), but rather to the fact that a has an intention with this proposition as its content. That is, (2) requires not that a intend that the audience recognize *that the audience believe* p, but rather requires that a intend that the audience recognize *that* a *intends that the audience believe* p. Harman's examples of self-referential attitudes, because self-referential in way (i), can provide no direct

[15] Grice (1957), 219.

[16] See Harman (1986), the section entitled "Self-Referential Attitudes" on pp. 87-8. See also Harman (1974).

[17] Kripke (1975).

[18] Harman goes on to discuss Gricean self-referential intentions in the passage immediately following the discussion of belief mentioned above; but he does not explicitly claim that we have the same sort of self-reference in the two cases.

[19] This presumes a conception of propositions on which they are structured. It could be rephrased so as to apply to other conceptions of propositions.

[20] Here and below I am thinking of descriptions as singular terms which have reference.

support for the coherence of Gricean self-referential intentions, which are self-referential in way (ii).

However, a variant of Harman's strategy of focusing on the attitude of belief fares better. Consider the following disquotational schema:

> If a understands S and reflectively and seriously accepts S in C, and S means p in C, then a believes p

It is plausible that all instances of this schema are true.[21] If they are, then the following case is an example of an agent having a belief which is self-referential in way (ii). Let "SR" be a name of the sentence "John believes that the proposition expressed by SR is true." A close friend of John's now approaches him, informs him that he has just named a sentence "SR", and tells John that the proposition expressed by SR is true. John, trusting his friend, says (and accepts) the following sentence: "The proposition expressed by SR is true."

By the relevant instance of the disquotational schema, it follows that John believes that the proposition expressed by SR is true. The content of John's belief may be represented as follows:

<BELIEF<John, (the x:x is the proposition expressed by SR) (x is true)>>

The referent of the description is, of course, the proposition expressed by SR. Since SR is the sentence "John believes that the proposition expressed by SR is true," the proposition expressed by SR may be represented as follows:

<BELIEF<John, (the x:x is the proposition expressed by SR) (x is true)>>

So the fact about belief noted above — that John believes that the proposition expressed by SR is true — is self-referential in way (ii), since the referent of one of the constituents of the proposition believed is the proposition that John believes that the proposition expressed by SR is true.[22] This result lends some credence to the claim that agents might also have intentions which are self-referential in this way.

It is hard to find any principles connecting uses of sentences with intentions in so direct a way as the disquotational schema connects acceptance of sentences with beliefs; for this reason, it is hard to find any direct arguments for the existence of intentions which are self-referential in way (ii) analogous to the above argument for this conclusion in the case of belief. The following example from Blackburn, however, seems to be a plausible case of an intention which is self-referential in way (ii) and seems to be clearly coherent:

[21] For some discussion of this schema, see above pp. 143 ff.

[22] Here I am relying on an identification of facts with true propositions; but I think that this is a dispensable part of the argument.

One might be inclined to deny that John understands the sentence he accepts, and so to deny that the disquotational principle applies here. But the only grounds for claiming this are that John does not understand (is not a competent user of) the name "SR," and this is implausible. "SR" is a proper name; and usually we require far less information than John possesses to count someone as a competent user of, and as understanding, a proper name. John, after all, is possessed of a uniquely identifying description of the referent of the name; he knows that it is the one and only sentence that his friend has just named. Moreover, the example could be reconstructed so that the sentence accepted is something like "The proposition expressed by the sentence written on the blackboard in Room 201 is true," which includes no such unfamiliar names.

> Imagine a certain kind of love affair. I want you to know *everything* about me. And everything includes, especially, the fact that I have this want. If you didn't know that about me, you might suspect me of concealment, and I wouldn't want that.[23]

It seems clear that we can grasp this intention, even though it is self-referential in way (ii). But, if this is right, it is difficult to see why one should think that the self-referential intentions employed in [G] should be thought incoherent or impossible to grasp. This, finally, is why I use [G] in the discussion of Gricean accounts in the main text.[24]

[23] Blackburn(1984), 117.

[24] Another prominent counterexample to the sufficiency of Grice's conditions is Searle's "Kennst du land" example, first presented in his (1969), 44. See Schiffer (1972), 26-30 for an argument that, on the interpretation of the example on which it does pose a challenge to the sufficiency claim, it is really only a version of the cases of deception discussed above.

Appendix B

Conditions of satisfaction and the expression of belief

In §2.2 above, I argued against the Gricean account of speaker-meaning. I concluded, partly on the basis of this argument, that there is no true account of speaker-meaning — or any other linguistic propositional attitude — in terms of facts about mental content. In this appendix, I consider the options of the mentalist who seeks to reply to this argument by giving a non-Gricean, but mentalist, account of speaker-meaning in terms of mental content. So far as I can see, these options are two: the accounts of speaker-meaning presented by John Searle and Christopher Peacocke.

Searle's account of speaker-meaning

The main early alternative to the Gricean account of speaker-meaning was given by John Searle. Initially, Searle attempted to give an account of speaker-meaning in terms of the intentions of speakers to bring about understanding in an audience via the Gricean mechanism. That is,

> a means p by uttering $x \equiv a$ intends in uttering x that
> (i) his audience understand that a means p,
> (ii) his audience recognize this intention, &
> (iii) (i) occur on the basis of (ii)[1]

As Searle later came to see, however, this account of speaker-meaning is open to precisely the same objection as Searle brought against Grice: a speaker may mean something by an utterance without trying to bring about any effects whatsoever in his audience. Hence one of the principle objections against the Gricean account of meaning applies also against Searle's early account.[2]

[1] Searle (1969), 47 ff. Searle goes on to try to unpack the notion of understanding; the details of that attempt won't matter here.

[2] See Searle (1983), (1986). This objection was first raised against Searle in Chomsky (1975). A revision of Searle's early approach given in Vlach (1981) avoids this problem. Vlach suggests

Searle later revised his account to avoid this objection. In *Intentionality*, he outlined his new approach as follows: "The mind imposes Intentionality on entities that are not intrinsically Intentional by intentionally conferring the conditions of satisfaction of the expressed psychological state upon the external physical entity."[3] The key to Searle's account is that the basis for an account of speaker-meaning is to be found in the fact that speakers, when meaning something by utterances, intend that their utterances have certain conditions of satisfaction. If we limit ourselves to the case of indicative sentences, these conditions of satisfaction may be identified with truth conditions. Thus Searle's theory of speaker-meaning for indicative sentences can be stated as follows:

a means p by uttering $x \equiv a$ intends that (his utterance of) x be true iff p

This account avoids the fundamental problem, noted above, with Searle's early account.

One might be inclined to object to this version of Searle's account that meanings are more finely grained than truth conditions; a speaker may say, "It is raining" and mean thereby that it is raining, *without* meaning that it is raining and arithmetic is incomplete, even though the proposition that it is raining is true if and only the proposition that it is raining and arithmetic is incomplete is true. But note that, by itself, this observation is no objection to Searle's account. The reference to truth conditions in Searle's account of speaker-meaning, after all, occurs within the scope of "intends"; the fact that a speaker can mean that it is raining by an utterance without meaning that it is raining and arithmetic is incomplete by that utterance may be explained, on Searle's account, by the speaker's intending that her utterance be true if and only if it is raining, and not intending that her utterance be true if and only if it is raining and arithmetic is incomplete.

But this response to the original objection makes it clear that the objection can be reformulated so as to apply to Searle's account. While it is true that my intentions may be as described, they need not be; I *may* intend that my utterance be true if and only if it is raining and arithmetic is incomplete, and yet not mean by my utterance that it is raining and arithmetic is incomplete. This is enough to show that the conditions on the intentions of speakers given in Searle's account are not sufficient for speaker-meaning.

Perhaps, though, a revision of Searle's account which incorporates the fine-grained character of meanings will fare better. We might try the following:

a means p by uttering $x \equiv a$ intends that (his utterance of) x mean p

It seems clear that Searle would not accept this revision. In discussing an attempt to account for facts about speaker-meaning in terms of facts about speakers' intentions that their utterance represent a certain state of affairs, Searle writes that it is "obviously no more than

---

the following account of speaker-meaning:

> a means p by uttering $x \equiv a$ utters x in the belief that he is thereby committing himself to p.

A serious problem with this account is that an utterance can commit you to a proposition even if that proposition was not among the things you meant by the utterance. Suppose that I mean p by an utterance, that q is a necessary consequence of p to which I was not previously committed, and that I know this. Then I made my utterance in the belief that I was thereby committing myself to q, even though I did not mean q by my utterance.

[3] Searle (1983), 27.

a preliminary to the analysis of the concept because it employs the notion of representation itself. One cannot analyse representation in terms of the intention to represent without circularity or infinite regress."[4] If this objection applies to an account of speaker-meaning in terms of intentions to represent, then it applies *a fortiori* to an account of speaker-meaning in terms of intentions about meaning.

It seems to me that Searle is a bit hasty here. It is not as though the above account analyzes speaker-meaning in terms of speaker-meaning; the reference to speaker-meaning on the right-hand side of the biconditional occurs within the scope of "intends." There seems to be nothing circular, or viciously regressive, in analyzing speaker-meaning in terms of the intention to mean something. But although I think that Searle misidentifies the problem with the above account, there is a problem in the same neighborhood. According to our revised version of Searle's account of speaker-meaning, it is a necessary condition on a speaker-meaning anything that the speaker possess the concept of speaker-meaning, and be capable of having intentions in which the concept figures. But this seems to get the order of explanation backwards.

There is, however, a more fundamental problem which affects both Searle's original account in *Intentionality* and our revised version of that account. In effect, both accounts attempt to explicate what *speakers mean* by utterances in terms of the intentions of speakers regarding what their *utterances* mean. But this seems to link too closely the notions of speaker-meaning and utterance-meaning. Suppose that you are standing on my foot, and I say to you, "You are standing on my foot", meaning thereby that you should get off of my foot. I do intend to mean that you should get off of my foot; but I do not intend that my *utterance* should mean that you get off of my foot. I know that my utterance means only that you are standing on my foot and do not intend, *per impossible*, that it mean something else. This is a fundamental problem with the account Searle defends in *Intentionality* and later papers, and with the improvements on that account suggested above. There seems to be no obvious way of solving this problem within Searle's framework.[5]

Peacocke on the expression of belief

Inasmuch as Searle's account of speaker-meaning tries to reduce the notion to the intentions of speakers, it is a variant of Grice's strategy rather than a completely new approach. A different strategy altogether is suggested by Christopher Peacocke in his defense of a version of mentalism in *Thoughts: An Essay on Content*. Peacocke sets out to give an account of what it is for a speaker to express a given belief by uttering a sentence; it is not unreasonable to think that the notion of expression of belief by an utterance could, if explicated in terms of mental content, serve as the basis for a mentalist account of linguistic meaning.

Peacocke gives the following analysis of the expression of belief:

> a expresses his believing p by uttering $x \equiv$
> (a) a meets condition (I),
> (b) a utters x intentionally, &
> (c) a utters x because a thinks that x is true

[4] Searle (1986), 214.

[5] See Talmage (1996) for further criticisms of Searle's account.

Peacocke describes condition (I) as follows:

> When a sentence s has the sense that p on someone's lips, the following is true of him: the conditions whose obtaining would give him reason to judge that p are precisely those which give him reason to think that an utterance of s would in fact be true. ... This identity of conditions holds in actual circumstances, and in any counterfactual circumstances in which s retains for him the sense that p. When such an identity holds, let us say that the subject meets condition (I).[6]

This can be paraphrased as follows:

> $\forall a \forall x \forall p \forall c$ (a meets condition (I) $\equiv \Box$(x has the sense that p on a's lips $\rightarrow$ (c obtaining gives a reason to judge $p \equiv c$ obtaining gives a reason to believe that x is true)))

Will this do as an account of what it is to express a belief by an utterance?

Whether or not Peacocke's analysis is true, it is of no use to the mentalist. The key problem comes with the definition of condition (I), and the phrase "when a sentence s has the sense that p on someone's lips." It is not entirely clear here whether Peacocke is referring to utterance-meaning or sentence-meaning; but in either case, his account does not advance the mentalist case. If Peacocke is referring to utterance-meaning, then his analysis of the expression of belief is given partly in terms of facts about linguistic meaning, and so in this case an account of linguistic meaning in terms of facts about the expression of beliefs by utterances will not be an analysis of linguistic meaning in non-semantic terms. If Peacocke is referring to speaker-meaning here, then any account of linguistic meaning in terms of the expression of belief will make use of facts about speaker-meaning; the discussion of Gricean attempts to reduce speaker-meaning to mental content is enough to show that this is not a notion which has been made unproblematic for the purpose of a mentalist account of linguistic meaning.[7]

So far as I know, this exhausts the attempts which have been made to give an account of speaker-meaning and related notions in terms of mental content.[8] The prospects for success for modifications of any of the three main strategies I've discussed appear dim: there are fundamental problems with the proposals of Searle and Peacocke, and the Gricean program

[6] Peacocke (1986), 115.

[7] One might think, though, that a variant of Peacocke's condition (I) itself could provide the key to an account of speaker-meaning. The natural choice is the following:

> $\Box \forall a \forall x \forall p \forall c$ (a means p by uttering $x \rightarrow$ (c obtaining gives a reason to judge $p \equiv c$ obtaining gives a reason to believe that x is true))

But this gives, at best, only a necessary condition for speaker-meaning; and it turns out that it does not give even that. I can utter a sentence x which I know to be false, and mean p by this utterance; and in such a case I might have reason to judge that p while having no reason to believe that x is true. This is related to the problems concerning the relation between speaker-meaning and utterance-meaning which proved problematic for Searle's later account of speaker-meaning.

[8] There are accounts of what it is for a speaker to mean something by an utterance which I haven't discussed which make no attempt to give an account in terms of mental content alone. For a recent example, see Parikh (2000), in which the analysis is partly in terms of facts about linguistic meaning.

faces counterexamples from so many different sides that it is difficult to see how any modification of the clauses of any version of the analysis could solve them all without giving rise to new problem cases. While there may be a true account of the kind sought which no one has yet imagined, the failure of each of these proposals provides good evidence that there is no true account of speaker-meaning, or related propositional attitudes such as the expression of belief, in terms of mental content.

Appendix C

A communitarian account of speaker-meaning

Grice's approach to speaker-meaning has us think about speaker-meaning in terms of the relationship between what a speaker means by uttering a sentence x and what effects the speaker intends his utterance of x to bring about. The failure of this account — and the sorts of counterexamples on which it fails — lead to a new way of thinking about speaker-meaning, on which the focus shifts from the effects a speaker intends her utterance to bring about to the meaning of the sentence uttered. Because Gricean accounts aim to make speaker-meaning safe for use in a constitutive account of linguistic meaning, they cannot make use of the meanings of expressions of languages as part of the story about what it is for a speaker to mean something by uttering an expression of such a language. The aim of this appendix is to show that this is a flaw in the Gricean account, by sketching a communitarian picture of speaker-meaning on which what speakers mean by utterances is partly constituted by the meanings of the expressions they utter.

It is obvious that speaker-meaning cannot be identified with sentence meaning; but perhaps we can give an account of speaker-meaning by thinking about different cases of a speaker meaning p by an utterance of x in terms of the relationship between p and the proposition expressed by x in the context. On this sort of account, what it is for a speaker a to mean p by uttering x is either for a to utter x seriously in a context in which x expresses p, or for a to .... The question is then how to fill in the ellipsis.

This is more than a formal departure from Grice's picture of speaker-meaning. According to that view, speaker-meaning was a notion well-suited to play an explanatory theoretical role; it was thought of as a sort of natural kind, unified by the sorts of reflexive audience-directed intentions discussed above. But on the present communitarian picture, speaker-meaning is not a unified kind in this way, and does not play the same kind of theoretical role; rather than explaining linguistic meaning, it is explained by linguistic meaning and the mental states of speakers. On this picture, the meanings of sentences are tools at the disposal of speakers of a language, which may be used in a number of ways to mean something by an utterance. From a communitarian perspective, there is thus no reason to think that we should be able to give a unified account of speaker-meaning.

Suppose that a speaker a means p by uttering x in a context C, even though x means q in C, and $p \neq q$. For a start, we can distinguish between three ways in which this might come about: (i) p is an obvious consequence of q; (ii) a uses x ironically, metaphorically, or otherwise non-literally to mean p; and (iii) a conveys p by means of some pragmatic device, such as a conversational implicature.[1] Can we give an account of these three sorts of cases adequate to fill out the account of speaker-meaning suggested above?

One idea is that the Gricean notion of S-meaning might be capable of covering these cases. I noted above that each of the counterexamples to Grice's analysis of meaning discussed above was a case in which a speaker meant p by an utterance of a sentence x which, in the context, literally meant p. It is thus not implausible to think that the Gricean analysis might work as an account of precisely those cases of speaker-meaning in which speaker-meaning and utterance meaning come apart. That is, abstracting again from context-sensitivity,

> a means p by uttering $x \equiv$((a utters x seriously & x means p) $\vee$ a S-means p by uttering x)[2]

On further reflection, though, it is clear that the Gricean account is ill-suited for even this limited role. Variants of all the counterexamples discussed above can be raised again here. A speaker might mean p by uttering a sentence x as the conclusion of an argument even if x does not mean p in the context by using x metaphorically, or by relying on the fact that p is an obvious consequence of x, as used in the context; but neither of these cases are cases of S-meaning, for the reasons given in the discussion of persuasive discourse above. Audienceless cases arise again as well; a speaker might mean p by an audienceless utterance of x, even though x does not mean p in the context, if the speaker is using x metaphorically or is aware of the fact that p is an immediate consequence of x, as used in the context.[3]

The problem here is another illustration of the general problem with the Gricean account noted above: the Gricean approach to speaker-meaning underestimates the relevance of the meaning of the sentence uttered to what the speaker means by uttering it. The account

[1] I leave it open whether there is substantial overlap between these categories; in particular, I leave it open whether metaphor could be treated as a species of conversational implicature. I do take this list to be jointly exhaustive, though I'm not very sure that it is.

It is not clear to me whether Grice regarded conversationally implicating p as a way of meaning p. He certainly distinguishes what a speaker implicates from what the speaker says (see, e.g., Grice (1978), 41); but he also distinguishes what a speaker says from what she means by an utterance (see Grice (1969)). While the fact that Grice consistently uses "implicates" rather than "means" when discussing implicature explicitly (see Grice (1975)) indicates that he may not have regarded conversational implicature as a kind of speaker-meaning, the accounts of speaker-meaning he favored to seem to imply that cases of conversational implicature are cases of speaker-meaning. In any case, it seems clear that ordinary usage is on the side of the claim that some — though, as I shall argue below, not all — cases of conversational implicature are cases of speaker-meaning; to use Grice's famous example, a professor might mean by writing "Tom has excellent penmanship" in a certain context that Tom is a mediocre student of philosophy.

[2] For a definition of 'S-means', see above p. 36.

[3] Even the problematic cases of reminding are problems here as well; if the name of your friend is on the tip of your tongue, I might say to you "He's no friend of mine", knowing that the friend that you are thinking of is a hated enemy of mine and that we've just been discussing this fact. But this will not meet Grice's conditions either, for the reasons given in §2.2.2.

discussed above tried to avoid this problem by assigning the Gricean account the limited task of accounting for cases of speaker-meaning in which what the speaker means diverges from what the speaker's utterance means; but, even in these cases, what the speaker means is closely related to the meaning of the speaker's utterance. Note that, in each of cases (i)-(iii), the meaning of the sentence uttered by the speaker seems essential to the speaker's meaning p. Although an agent a need not, in order to mean p by an utterance, use a sentence x which means p, a cannot use just any sentence; the meaning of the sentence uttered must be related, in some way or other, to what the speaker means. This indicates that, in giving an account of these sorts of cases of speaker-meaning, we will need to tie the account to the meaning of the sentence uttered in a way that Grice's approach does not. This is, I think, best done by considering the three sorts of cases of speaker-meaning listed above individually.

Obvious consequences of utterances

In the first sort of case, we have a speaker meaning p by an utterance of x which literally means q, by exploiting the fact that p is an obvious consequence of q. It is important to note that this class of cases includes only serious utterances of sentences: those in which a speaker utters a sentence x which means q without sarcasm, irony, or conversational implicature of the sort to make it the case that the speaker does not mean q by uttering x. So what must the relationship between p and q be for the speaker to also mean p by uttering x?

It is not sufficient that p simply *be* an obvious consequence of q; the speaker must recognize this in order for him to mean p in this sort of case. Nor, it seems, is it necessary for p to really be an immediate consequence of q; if a speaker takes p to be an immediate consequence of q, even if he is mistaken in this belief, this might be enough. Suppose that a speaker mistakenly believes that it is not a leap year. A friend asks him, "Is this the last day of the month?" He replies, "Well, it's the 28^{th} of February, isn't it?" He means by this utterance, not only that it is the 28^{th} of February, but also that it is the last day of the month; and he means this because he believes that the latter proposition is an immediate consequence of the one expressed by his utterance. The fact that he is mistaken in this belief seems not to be relevant to the question of whether he means by his utterance that it is the last day of the month.

In cases in which a speaker has an audience, must the speaker also believe that his audience will take p to be an immediate consequence of q? I don't think so. Suppose that a is a professor administering an oral examination to a particularly incompetent student b. a realizes that b has contradicted himself by uttering sentences x and x', though he knows that b's grasp of elementary logic is such that b is utterly incapable of working out the fact that he has asserted contradictory sentences. b again utters x; in reply, a utters x'. By uttering x', a means that x is false, because the negation of x is, for a, an obvious consequence of x'. But a is painfully aware of the fact that b doesn't recognize this fact, and indeed may not do so even if given a great deal of time to work it out. But this doesn't seem to stop a from meaning that x is false by his utterance. In short, to build into the account facts about a speaker's expectations regarding an audience is to recapitulate Grice's mistake of linking speaker-meaning too closely to intended effects in an audience.

Nevertheless, it seems clear that a's seriously uttering a sentence x which means q and a's believing p to be an immediate consequence of q, though necessary in cases of this type, are not jointly sufficient for x meaning p by his utterance of x. a might be an accomplished logician, capable of deriving at a moment's notice countless immediate consequences of q; it

doesn't follow that she means countless things by her utterance of x. What seems to be needed is an extra constraint to do with relevance — or, more precisely, with what the speaker takes to be relevant. Sometimes the important consideration may be relevance to a conversation among different individuals; other times, as in the absence of an audience, it may be relevance to what the speaker is thinking about at the time. The addition of this requirement seems to give us necessary and sufficient conditions for cases of speaker-meaning of type (i).[4]

Metaphor

The case (ii) of metaphor is considerably more complicated. First, note that the concern here is with *speakers* meaning things by their metaphorical and ironic utterances; thus we need not enter into the debate over whether these utterances themselves should be said to have a meaning distinct from their literal meaning, or whether cases of metaphoric meaning are only cases of speaker-meaning.[5]

Second, we can usefully distinguish between 'live' and 'dead' metaphorical expressions. The latter are expressions which have, through repeated metaphorical use, acquired a new (literal) meaning. We can imagine, for example, that the phrase "the legs of a table" was once used as a metaphor which derived its force from the similarity between the function of the legs of animals and the function of the structures which support tables; now, however, uses of this phrase to refer to the structures which support a table is a literal use. Dead metaphors are not really metaphors at all; they are expressions which used to be metaphors. As such, they don't pose any special problem for accounts of speaker-meaning. Live metaphors, on the other hand, are new or little used enough that they have not acquired a new literal meaning; paradigmatic examples are to be found in poetry. When Eliot wrote, "I have seen the eternal Footman hold my coat, and snicker,/And in short, I was afraid"[6] he was not relying on a pattern of previous metaphorical uses of sentences involving eternal footmen holding coats and snickering to give his sentence a new, immediately obvious, literal meaning.[7] Our problem

[4]One worry here is that these conditions may conflict with my earlier claim that the propositional attitude of speaker-meaning distributes over conjunction. (See p. 33.) Cases where we want to invoke the distribution of speaker-meaning over conjunction certainly seem to fit into category (i) above; so my claim about relevance above commits me to the claim that a speaker who utters a conjunctive sentence x seriously — and so takes the proposition expressed by x to be relevant — must also take the propositions expressed by each of the conjuncts to be relevant. I am inclined to endorse this consequence.

[5]See Moran (1997) for a summary of this debate.

[6]Eliot, "The Love Song of J. Alfred Prufrock."

[7]It is characteristic of live metaphors that we can clearly distinguish between what the speaker means by his utterance of the metaphorical sentence, and what the sentence literally means; and that the proposition literally expressed by the sentence in the context is not among the things the speaker meant by uttering the sentence.

It is worth noting that expressions can be widely used as metaphors, even hackneyed, and yet still be, in the relevant sense, live metaphors. Consider "President George Bush did not fall on his face or reach for his gun; instead he has shown considerable amounts of skill, subtlety, leadership and, above all, intelligence" (from "Closing In," *The Economist* 360:8241 (29 September 2001), 11). The metaphors of falling on one's face or reaching for a gun are hardly original; but they clearly qualify as live metaphors. Suppose it turned out that President Bush did (literally) fall on his face or reach for his gun; it would not follow that any of the propositions which the author of this article meant

is to give an account of speakers meaning things by utterances of live metaphors.

I don't know how to give such an account. Two points, though, indicate that the sort of account of speaker-meaning being sketched in this section is better positioned to handle metaphorical language than the mentalist accounts discussed above. First, as noted above, uses of sentences as live metaphors to mean things rely on the literal meanings of the sentences; even though speaker-meaning diverges from literal linguistic meaning here, it seems plausible that there is some dependence of speaker-meaning on that literal meaning. Eliot could not have meant by writing "I have seen the transitory Footman hold my shoes and chortle,/And in short, I was famished" what he meant by writing the line from "Prufrock" quoted above. Second, metaphorical uses of language are not restricted to uses of language in Gricean communication. One could use a metaphor to convey the conclusion of an argument, to express something while writing in one's diary, or when saying a prayer. Yet, as we've seen, none of these cases of speaker-meaning fit into the Gricean story.

Pragmatic Implicatures

This still leaves cases of type (iii), in which a speaker means p by exploiting various pragmatic devices to do with communication. In many such cases, the speaker's utterance will not be a metaphorical one, and what the speaker means by the utterance cannot be explained as an obvious consequence of the proposition semantically expressed by the utterance.

The most well-known sorts of pragmatic devices of this kind are Gricean conversational implicatures. Grice gives the following necessary and sufficient conditions for a speaker to conversationally implicate p by an utterance x:

(1) the speaker is presumed to be observing the Cooperative Principle, and/or the conversational maxims,
(2) the supposition that the speaker is aware that, or thinks that p is required to make his utterance of x consistent with (1), &
(3) the speaker thinks, and expects his audience to think that the speaker thinks, that it is within the competence of his audience to grasp (2).[8]

by writing this sentence were false. This sort of metaphorical use of language, though, is different from the poetic case in that it does exploit previous uses of the relevant expressions to mean exactly what the speaker means by them.

[8]See Grice (1975), 30-1. I intentionally leave it open whether, if the speaker's utterance x literally means q in the context, the speaker means q by his utterance as well as the conversationally implicated proposition p. I think that in some cases a will, and in some cases he won't, depending on the maxim violated. If, for example, a violates one of the Maxims of Quality, it seems likely that a will not mean q; if he merely violates one of the Maxims of Quantity, he probably will.

I do have some doubts about this treatment of implicature, which are connected to the objections given above to Grice's other work on meaning. Specifically, I wonder whether there are cases in which a phenomenon very similar to conversational implicature occurs in non-communicative uses of language — for example, without there being an audience (and hence without there being a conversation). Consider the following variant of one of Grice's cases. b approaches a, and asks him, "Do you know where I could get some gas for my car?" a, who despises b, replies, "No; wish I could help." Walking away, a then says under his breath, "Of course, there's a gas station around the corner." It seems clear that by uttering this a means that b could have gotten some gas for his car

Grice was interested in characterizing the phenomenon of conversational implicature; what we're interested in are pragmatic devices for non-literal speaker-meaning. An initial thought is that the two coincide: that all and only cases of non-literal speaker-meaning which exploit pragmatic devices are instances of Gricean conversational implicature.

Unfortunately, neither direction of this claim is true; but seeing how it fails will help us arrive at a better characterization of class (iii) of cases of non-literal speaker-meaning. First, meeting conditions (1)-(3) above is not sufficient for speaker-meaning: not all cases in which a speaker conversationally implicates p are cases of speaker-meaning. Suppose that a man is asked where his son Bob is, and answers, "Bob is either in Los Angeles or in Chicago."[9] If the man knows that Bob is in Los Angeles (or knows that he is in Chicago), then his utterance would violate the maxim of Quantity: he would be giving less information than the conversational setting requires. Hence the supposition that the man does not know whether his son is in Los Angeles or in Chicago is required for his utterance to be in conformity to the conversational maxims. We can presume that the man is taken to be obeying the conversational maxims, and that he takes it to be within the powers of his audience to arrive at the above supposition. So in this sort of case, Grice's conditions on conversational implicature entail that the man conversationally implicates that he does not know whether his son is in Los Angeles or in Chicago. And, while this seems right, it seems wrong to say that the man *meant* by his utterance that he does not know whether his son is in Los Angeles or in Chicago.[10]

Neither are Grice's conditions on conversational implicature necessary for a speaker to mean p by an utterance via a pragmatic device, as is shown by the following example, due to Jennifer Saul.[11] The case is a variant of one of Grice's classic examples of conversational implicature. Grice imagined a professor faced with the task of writing a letter of recommendation for a philosophy student a whom she considers to be a very poor student. Reluctant to make any negative comment explicitly, the professor writes in the letter, "a is an excellent typist, and is very punctual," intending thereby to convey the information to his readers that she does not think much of a as a student. But, Saul imagines, unbeknownst to the professor, the student was applying for a job as a typist rather than as a philosopher. Hence, as it turns out, the supposition that the professor thinks a to be a poor student is *not* required for his letter to be in conformity with the relevant conversational maxims; since condition (2) above is not satisfied, in this case the professor is not counted as conversationally implicating p. Yet it seems clear that the professor did mean by writing the letter that a is a poor student.

around the corner. How can this case be accounted for? Had a said this aloud, we could give a Gricean account, using the Maxim of Relation ("Be relevant"). But this sentence was not uttered in conversation; should we say that the conversational maxims are still at work? But why, since there is no cooperative endeavor going on here at all? Maybe this could be treated under case (i) of speakers meaning by an utterance obvious consequences of the proposition expressed by the utterance; I'm not sure.

[9]This case was pointed out to me by Scott Soames.

[10]Note that this is not directly an argument against Grice himself, since, so far as I know, he nowhere claims that all cases of conversational implicature are cases of speaker-meaning. It may be that this sort of case could be turned into a further argument against Grice's analysis of speaker-meaning, however; it may be that in such cases the Gricean analysans entails wrongly that the man did mean that he did not know which city his son was in. I'm not sure about this.

[11]Saul (2002).

In response to this case, Saul defines a notion of utterer's implicature. A speakers utterance of x is an utterer's implicature of p just in case the following conditions are satisfied:

(1*) the speaker takes himself to be observing the Cooperative Principle, and/or the conversational maxims,
(2*) the speaker thinks that the supposition that the speaker is aware that, or thinks that p is required to make his utterance of x consistent with (1), &
(3) the speaker thinks, and expects his audience to think that the speaker thinks, that it is within the competence of his audience to grasp (2).

Plausibly, these are necessary conditions for a speaker meaning p via a pragmatic device of some kind. But we still have the example of the man's utterance regarding his son's location which shows that (1*), (2*), and (3) are not jointly *sufficient* for speaker-meaning.

Why does this case not seem to be a genuine case of speaker-meaning? It seems to me that it is a necessary — though not sufficient — condition for a speaker meaning p by an utterance that the speaker take himself to be, in some strong sense, *committed* to p. This may be what is lacking in the case discussed above: though the speaker does convey the information that he does not know which of the cities his son is in, he is not committed to defending this proposition in that way that, for example, the professor in the case described above is. It would take some work to make this more precise; but it seems that there must be some such way of distinguishing between the two cases.[12]

This is, as yet, only a sketch of a theory of speaker-meaning; but it does seem that this way of thinking about speaker-meaning faces far fewer obstacles than the sorts of mentalist approaches discussed above. When trying to give an account of speaker-meaning, we should not begin by considering the intended effects of utterances in isolation from the meanings of the sentences uttered. Rather, we should take meaning as prior to speaker-meaning, and explain cases in which speaker-meaning outstrips linguistic meaning in terms of the literal meanings of utterances, along with facts about what those speakers recognize, believe, and intend. This sketch of an account, along with the failure of Gricean accounts is, it seems to me, enough to substantiate the claim suggested above: facts about the meanings of sentences are prior to and constitutive of facts about what speakers mean by uttering those sentences, rather than the other way around.

This result, moreover, seems to generalize to the other action-based propositional attitudes associated with utterances of sentences: saying, asserting, communicating, informing, and so on. Given the failure of the Gricean account to capture these notions, the scope of the argument given in §2.2.4 against mentalist accounts of these notions, and the — I think, hopeful — prospects for giving a workable account of these notions along the lines of that suggested for speaker-meaning in this section, this conclusion seems plausible.

[12] I'm less and less sure about the claim made in this paragraph the more I think about it.

A further worry is about the completeness of this characterization of speaker-meaning via pragmatic devices. In particular, this may run into trouble with conventional implicatures and with the generalized or "short-circuited" conversational implicatures discussed in Michael Nelson's *Using Words: Pragmatic Implicatures & Semantic Contents* (PhD dissertation, Princeton University, unpublished).

Appendix D

Convention, serious circumstances, and word-meaning

In Chapter 2 above, I discussed one route for the mentalist to go in giving an account of public language meaning in terms of the beliefs and intentions of speakers. This indirect mentalist strategy reduced meaning to some class of the action-based propositional attitudes (assertion, speaker-meaning, etc.) and then, in turn, gave an account of this class of propositional attitudes in terms of the communicative intentions of speakers. One problem with this strategy was that there is no account of any of the action-based propositional attitudes in terms of the intentions of speakers; for this reason, I suggested that the proponent of the indirect mentalist strategy should stipulatively define a propositional attitude of S-meaning in terms of the intentions of speakers, and then try to give an account of linguistic meaning in terms of S-meaning. Above I discussed one way in which one might try to reduce linguistic meaning to S-meaning, which ran as follows:

x means p in a population $G \equiv$
(1) almost all members of G utter x only when they S-mean p by uttering x,
(2) almost all members of G mutually know (1), &
(3) (1) obtains because of (2)

Above I discussed a number of what I take to be fundamental objections to this view of linguistic meaning. The purpose of this appendix is to discuss a few slightly more technical problems which also attend this version of the mentalist program.

The first of these is an objection to clause (1) of each of the versions of the convention-based approach discussed above. As advocates of this approach are aware, the simplified versions of this clause given above will not do, since there are many circumstances in which, although a sentence x means p, a speaker might utter x without thereby meaning p. Brian Loar gives the following list: "creative metaphors, hyperbole, telling jokes, testing microphones and typewriters, and, in general, utterances such that the speaker knows that his hearer will not

take his utterance literally."[1] It seems clear that, as Loar suggests, clause (1) of convention based analyses will have to be restricted to 'serious circumstances,' to read:

> almost all members of G utter x in serious circumstances only when they S-mean p by uttering x

The challenge for the mentalist is then to say exactly what these serious circumstances are.

This is difficult enough if the goal is to restrict ourselves to those circumstances which are such that, for some sentence x which means p, if speakers utter x in those circumstances, they mean p by their utterance; it is not obvious how to explicate serious or literal utterances without appealing to the meaning of the sentence uttered. But it seems virtually impossible if the goal is, as it must be for the mentalist, to restrict ourselves to those circumstances which are such that, if speakers utter x in those circumstances, they will *S-mean p* by their utterance. For the definition of 'serious circumstances' will have to rule out not only the sorts of cases of non-literal uses of language mentioned by Loar, but also the long list of cases of speaker-meaning which are not cases of S-meaning: the broad classes of cases grouped under the headings of persuasive discourse and speaker-meaning without intended effects, as well as the cases of reminding and answering discussed above. There seems no easy way to group these cases together; they are certainly not uniformly 'un-serious' or non-literal.

If there is no way to group these cases together without using the concept of S-meaning, the first clause of the convention-theorist's analysis will seem *ad hoc* and uninformative; it will be something like the following:

> Almost all members of G utter x in circumstances which are (i) serious and (ii) not circumstances in which an utterance of x by the speaker would be a case of speaker-meaning but not S-meaning only if they S-mean p by uttering x

This is far less appealing as an analysis of what it is for a sentence to have a given meaning — even if there is an account of what it is for a circumstance to be serious which does not appeal to the meaning of the sentence uttered — than was our original clause (1). Further, unless a solution to the problem raised against the appeal to knowledge of speaker-meaning is found, whatever is built into a revised version of clause (1) will be attributed to the content of the mutual knowledge possessed by speakers by clause (2); and, the more complex clause (1) becomes, the less plausible clause (2) will be.[2]

The second further problem for the convention-theorist concerns the need to give an account of the meanings of sub-sentential expressions. The mentalist needs an account of facts about the meanings of such expressions for two reasons. First and most obviously, sub-sentential expressions have meanings; hence facts about their meaning are among those for

[1] Loar (1976), 155.

[2] Clause (2), for example, might read as follows:

> $\forall x \forall p \forall G$ (x means p in a population $G \rightarrow$ almost all members of G mutually know that almost all members of G utter x in circumstances which are (i) serious and (ii) not circumstances in which an utterance of x by the speaker would be a case of speaker-meaning but not S-meaning only when they S-mean p by uttering x)

Here I use the *in sensu diviso* formulation of (2), which attributes less to the mutual knowledge of speakers and so is presumably more appealing to the convention-theorist. See note 49 on p. 39 above on Lewis's distinction between knowledge *in sensu diviso* and knowledge *in sensu composito*.

which the mentalist owes an account. Second, facts about the meanings of such expressions are needed to give an account of the meanings of some sentences; even if some version of [C] turns out to be true, it will not in general deliver an account of the meanings of all the sentences of a language. Some sentences of a language may be so long or complex that no speakers of the language are able to understand them, or have any dispositions to utter them at all. No convention-based theory of meaning can account for the meanings of such sentences directly, for the obvious reason that, if speakers of the language can't understand such sentences, they have no mutual knowledge about what speakers would mean by utterances of those sentences in the sorts of circumstances mentioned above.[3]

Without getting into particular convention-based accounts of the meanings of sub-sentential expressions, I'd just like to note that here too the failure of the Gricean analysis of speaker-meaning poses a problem for the convention-theorist. Presumably, the convention-theorist will account for the meanings of sub-sentential expressions in terms, somehow, of the meanings of some of the sentences in which they occur — those sentences whose meaning can be explicated by some version of [C]. The problem is that it seems perfectly possible that some expression of a language might only be used in sentences which are never used to *S*-mean anything; if this were the case, then there would be no way for the convention-theorist to work back from the meanings of sentences to the meaning of that expression. Suppose, for example, that the word "therefore" had a more specialized meaning than it does, and was only used to state the conclusions of arguments. Because such uses of sentences in persuasive discourse, as argued in §2.2.2 above, do not qualify as instances of *S*-meaning, there would be no sentences involving the expression for which the convention-theorist could directly give an account; and, in this case, it is hard to see how any account of the expression could be given.

[3]One response to this last problem might be to deny that such sentences are sentences of the language of the population in question. (Loar (1976) considers this.) In a way, this is a minor point — as noted above, the mentalist owes an account of the meanings of sub-sentential expressions independently of this claim. Though the thought that such sentences are not really part of the language of the population does not sound wholly implausible to me, the following sort of argument does seem to indicate that some sentences too long for almost everyone in the community to understand should be considered part of the language; and these sentences raise much the same problems for the convention-theorist as sentences too long or complex for *anyone* in the community to understand. Suppose that a very small group of English speakers have an extraordinarily high capacity for understanding complex sentences of their language; let S be some sentence which means p and which only these speakers have any dispositions to utter. As stated, the conditions on conventions will not account for the meaning of S, since it is not the case that almost all speakers of English are disposed to assert S in such-and-such circumstances only when they mean p by their utterance. Hence the meaning of S can only be accounted for on the basis of the meanings of the expressions which occur in it.

The convention-theorist might reply by saying that S is only a sentence in the language spoken by this minority of English speakers, and is not a sentence of English *simpliciter*. But suppose that there are only five such speakers; isn't it possible that all of these speakers misunderstand and habitually use S in the wrong way? But this possibility is ruled out by this move on the part of the convention-theorist. Suppose that each of these speakers are disposed to utter S only when they mean q. Even if $q \neq p$, it seems that it might still be the case that S means p; and the only way to account for this result is by means of the meanings of the expressions in S.

Appendix E

Complications with dispositions

In giving a positive account of belief, I appealed to the dispositions of language-using agents to accept meaningful sentences of public languages, using the disquotational principle connecting these dispositions with beliefs. But, though this principle sounds simple enough, not much reflection is required to show that making it precise is not an easy task.

One problem emerges from noticing that the disquotational principle gives an incomplete description of the relevant linguistic dispositions. I wrote of dispositions to accept sentences; but, typically, the specification of a dispositional property requires not only specification of the manifestation of the disposition, but also of the conditions under which the disposition is manifested. A dispositional property like fragility is not just a disposition to break; presumably almost anything, including many un-fragile things, would break under certain conditions. Rather, fragility is the property of being disposed to break when struck, or break when dropped, or .... It is controversial how the ellipsis should be filled in; but it is not controversial that it needs to be filled in somehow or other.

Similar remarks are in order for dispositions to accept sentences. There are many propositions which I do not believe, but which are such that there is some sentence which expresses them, which I understand, and which I would accept in *some circumstances*. Noting the possibility of brainwashing and the like is enough to show that we need some limitations on the conditions for the manifestation of these linguistic dispositions.

One possible source of complications here is that the factors relevant to my accepting a sentence in different circumstances of being presented with it might vary with my psychological state. But if the relevant parts of my psychological state include facts about my beliefs, this gives rise to worries about circularity. So it is not legitimate for a proponent of the kind of account I defend in Chapter 8 to simply rest easy with the thought that there will be some way or other of filling out the conditions for manifestation of the dispositions; not all ways of filling out these conditions will serve the purposes of a communitarian account of belief.

A second kind of complication raised by the above use of dispositional properties has to do with the relationship between dispositions and counterfactuals. For any specification of a disposition to do perform some action under certain conditions of manifestation, there is an associated counterfactual to the effect that, were those conditions satisfied, one would perform the action. But, at least under the standard possible worlds interpretation of counterfactuals, the truth of the associated counterfactual must, if the account I have been defending is to

be correct, not be sufficient for the instantiation of the dispositional property. To see this, consider the following case:

> *A* does not, at present, believe that he has conflicting lunch dates for Wednesday. But he does have conflicting lunch dates; and, were he presented with a sentence expressing the proposition that he has conflicting lunch dates for Wednesday, he would immediately realize this, and endorse the sentence.[1]

In this situation, prior to being presented with the sentence, *A* satisfies the counterfactual claim that were *A* presented with a sentence which means that *A* has conflicting lunch dates for Wednesday, *A* would accept that sentence. But this had better not mean that, at that time, *A* was disposed to accept such a sentence; for, if he was, then the disquotational principle would entail that, contrary to the description of the case, *A already* believed that he had conflicting lunch dates. But this is surely wrong.[2]

A different kind of case arises with non-linguistic animals. It is not strange to attribute beliefs about water to a horse, if the horse is disposed to, say, go to water and drink when water is presented to it in perception. But it is also the case that, were the horse presented with some XYZ in perception — a different but superficially indistinguishable substance — the horse would go to the XYZ and take a drink. If counterfactuals were sufficient for their associated dispositional properties, the account of non-linguistic beliefs in Chapter 8 might force us into saying that the horse has beliefs about XYZ. But again, this seems wrong.

So we need an account of dispositions not in terms of their associated counterfactuals. One might worry that this dissociation of dispositions from comparatively well-understood counterfactuals make the dispositional properties at the heart of the communitarian account of belief defended in Chapter 8 rather mysterious. This is a worry; some solace comes from the fact that it is plausible to think that the problems here stem from the counterfactual analysis of dispositions rather than from the present dispositional analysis of belief. After all, it does seem strange to say that *A* was disposed to accept the above sentence before he was presented with it; intuitively, he acquired this disposition upon being presented with the sentence. It would not be unnatural to say the same about the horse and XYZ. But this intuition is no substitute for a worked-out account of these dispositions.

To sum up: I do not regard these as decisive problems; for one, the same problems arise for every other constitutive account of belief, since all such accounts make some use of dispositions to action. But they do show that matters are more complicated than the discussion in Chapter 8 makes them seem, and so that the present account of belief must be regarded as, in an important way, incomplete without a satisfactory account of the relevant dispositional properties.

[1] This example is due to Jonathan Beere.

[2] It should be possible to come up with analogous cases which show that the truth of the associated counterfactual also cannot be necessary for the holding of the dispositional property, but I have not been able to come up with a case of this kind.

Bibliography

Armstrong, David [1971]. Meaning and communication. In *Philosophical Review* 80:427–447.

Avramides, Anita [1989]. *Meaning and Mind*. Cambridge, MA: MIT Press.

Bar-On, Dorit [1995]. 'Meaning' reconstructed: Grice and the naturalizing of semantics. In *Pacific Philosophical Quarterly* 76:2:83–116.

Bennett, Jonathan [1973]. The meaning-nominalist strategy. In *Foundations of Language* 10:141–168.

— [1976]. *Linguistic Behavior*. Cambridge: Cambridge U.P.

Biro, J. [1979]. Intentionalism in the theory of meaning. In *The Monist* 62:238–258.

Blackburn, Simon [1984]. *Spreading the Word*. Oxford: Clarendon Press.

Block, Ned [1986]. Advertisement for a semantics for psychology. In Stich & Warfield (1994), 81–141.

Brandom, Robert [1994]. *Making It Explicit*. Cambridge, MA: Harvard University Press.

Burge, Tyler [1975]. On knowledge and convention. In *Philosophical Review* 84:2:xx–xx.

— [1979]. Individualism and the mental. In Ludlow & Martin (1998), 21–83.

— [1986]. Intellectual norms and the foundations of mind. In *Journal of Philosophy* 83:12:697–720.

Chisholm, Roderick & Sellars, Wilfrid [1958]. Correspondence on intentionality. In *Intentionality, Mind, & Language*, edited by Marres, Ausonio. Urbana, Illinois: University of Illinois Press, 214–248.

Chomsky, Noam [1959]. Review of Skinner's "Verbal Behavior". In *Language* 35:26–58.

— [1975]. *Reflections on Language*. London: Temple-Smith.

— [1980]. *Rules and Representations*. Oxford: Basil Blackwell.

— [1986]. *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger Publishers.

Christensen, Carleton [1997]. Meaning things and meaning others. In *Philosophy and Phenomenological Research* 57:3:495–522.

Clark, Michael [1975]. Utterer's meaning and implications about belief. In *Analysis* 35:105–108.

Davidson, Donald [1967]. Truth and meaning. In his *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press, 17–36.

— [1986]. A nice derangement of epitaphs. In Hale & Wright (1997), 443–446.

Davis, Wayne [1999]. Communicating, telling, and informing. In *Philosophical Inquiry* 21:1:21–43.

Dretske, Fred [1981]. *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

— [1988]. *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: MIT Press.

— [1990]. Reply to reviewers. In *Philosophy and Phenomenological Research* 50:4:819–839.

Dummett, Michael [1985]. Truth and meaning. In Dummett (1993), 147–165.

— [1986]. 'A Nice Derangement of Epitaphs': Some comments on Davidson and Hacking. In Hale & Wright (1997), 459–476.

— [1989]. Language and communication. In Dummett (1993), 166–187.

— [1993]. *The Seas of Language*. Oxford: Oxford University Press.

Evans, Gareth [1977]. Pronouns, quantifiers, and relative clauses (I). In his *Collected Papers*. Oxford: Oxford University Press, 76–152.

Field, Hartry [1977]. Logic, meaning, and conceptual role. In *Journal of Philosophy* 74:7:379–409.

— [1978]. Mental representation. In Stich & Warfield (1994), 34–77.

— [1986]. Stalnaker on intentionality. In *Pacific Philosophical Quarterly* 67:2:98–112.

— [1994]. Deflationist views of meaning and content. In *Mind* 103:249–285.

Fodor, Jerry [1975]. *The Language of Thought*. Hassocks: Harvester Press.

— [1980]. Psychosemantics or Where do truth conditions come from? In *Mind and Cognition*, edited by Lycan, William. Oxford: Basil Blackwell, 312–338.

— [1985]. Fodor's guide to mental representation. In Fodor (1990*a*), 3–30.

— [1987]. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

— [1990*a*]. *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

— [1990*b*]. A theory of content, II: The theory. In Fodor (1990*a*), 89–136.

— [1994]. *The Elm and the Expert*. Cambridge, MA: MIT Press.

— [2001]. Language, thought, and compositionality. In *Mind & Language* 16:1:1–15.

Grandy, Richard E. & Warner, Richard, eds. [1986]. *Philosophical Grounds of Rationality: Intentions, Categories, Ends*. Oxford: Clarendon Press.

Greenberg, Mark [in preparation]. Incomplete understanding, deference, and the content of thought. In .

Grice, Paul [1957]. Meaning. In Grice (1989), 213–223.

— [1968]. Utterer's meaning, sentence meaning, and word-meaning. In Grice (1989), 117–137.

— [1969]. Utterer's meaning and intentions. In Grice (1989), 86–116.

— [1975]. Logic and conversation. In Grice (1989), 22–40.

— [1978]. Further notes on logic and conversation. In Grice (1989), 41–57.

— [1982]. Meaning revisited. In Grice (1989), 283–303.

— [1987]. Retrospective epilogue. In Grice (1989), 339–386.

— [1989]. *Studies in the Way of Words*. Cambridge, MA: Harvard U.P.

Hale, Bob & Wright, Crispin, eds. [1997]. *A Companion to the Philosophy of Language*. Oxford: Basil Blackwell.

Harman, Gilbert [1974]. Stephen Schiffer: *Meaning*. In *Journal of Philosophy* 71:7:224–229.

— [1977]. Review of Jonathan Bennett's *Linguistic Behavior*. In *Language* 53:417–424.

— [1986]. *Change in View: Principles of Reasoning*. Cambridge, MA: MIT Press.

— [1987]. (Nonsolipsistic) conceptual role semantics. In Harman (1999), 206–232.

— [1988]. Wide functionalism. In Harman (1999), 235–243.

— [1999]. *Reasoning, Meaning, and Mind*. Oxford: Clarendon Press.

Hyslop, Alec [1977]. Grice without an audience. In *Analysis* 37:67–69.

Johnston, Mark [1988]. The end of the theory of meaning. In *Mind & Language* 3:1:28–42.

— [1993]. Objectivity refigured: Pragmatism without verificationism. In *Reality, Representation, & Projection*, edited by Haldane, John & Wright, Crispin. New York: Oxford University Press, 85–130.

Kripke, Saul [1972]. *Naming and Necessity*. Cambridge, MA: Harvard University Press.

— [1975]. Outline of a theory of truth. In *Journal of Philosophy* 72:609–716.

— [1982]. *Wittgenstein on Rules and Private Language: An Elementary Exposition*. Cambridge, MA: Harvard U.P.

Laurence, Stephen [1996]. A Chomskian alternative to convention-based semantics. In *Mind* 105:269–301.

LePore, Ernest, ed. [1986]. *Truth & Interpretation: Perspectives on the Philosophy of Donald Davidson*. Oxford: Basil Blackwell.

Lewis, David [1969]. *Convention*. Cambridge, MA: Harvard University Press.

— [1970]. How to define theoretical terms. In *Journal of Philosophy* 67:427–446.

— [1975]. Languages and language. 163–188.

— [1980]. Mad pain and Martian pain. In Rosenthal (1991), 229–233.

Loar, Brian [1976]. Two theories of meaning. In *Truth and Meaning: Essays in Semantics*, edited by Evans, Gareth & McDowell, John. Oxford: Oxford University Press, 138–161.

— [1981]. *Mind and Meaning*. Cambridge: Cambridge University Press.

— [1982]. Conceptual role and truth-conditions. In *Notre Dame Journal of Formal Logic* 23:3:272–283.

— [1991]. Can we explain intentionality? In Loewer & Rey (1991), 119–136.

Loewer, Barry [1997]. A guide to naturalizing semantics. In LePore (1986), 108–126.

Loewer, Barry & Rey, Georges, eds. [1991]. *Meaning in Mind: Fodor and His Critics*. Cambridge, MA: Blackwell.

Ludlow, Peter [1995]. Externalism, self-knowledge, and the prevalence of slow switching. In Ludlow & Martin (1998), 225–230.

— [1999]. *Semantics, Tense, and Time: An Essay in the Metaphyics of Natural Language*. Cambridge, MA: MIT Press.

Ludlow, Peter & Martin, Norah, eds. [1998]. *Externalism and Self-Knowledge*. Stanford, CA: CSLI Publications.

Lycan, William [1999]. *Philosophy of Language: A Contemporary Introduction*. New York: Routledge.

McDowell, John [1980]. Meaning, communication, and knowledge. In *Philosophical Subjects: Essays Presented to P.F. Strawson*, edited by van Straaten, Zak. Oxford: Clarendon Press, 117–139.

— [1994]. *Mind & World*. Cambridge, MA: Harvard University Press.

McKeown-Green, Jonathan [2002]. *The Primacy of Public Language*. PhD Dissertation, Princeton University: unpublished.

Millikan, Ruth [1989]. Biosemantics. In *Journal of Philosophy* 86:6:281–297.

Moran, Richard [1997]. Metaphor. In LePore (1986), 248–270.

Neale, Stephen [1992]. Paul Grice and the philosophy of language. In *Linguistics & Philosophy* 15:5:509–559.

Nelson, Michael [2001]. *Using Words: Pragmatic Implicatures & Semantic Contents*. PhD Dissertation, Princeton University: unpublished.

Papineau, David [1987]. *Reality and Representation*. Oxford: Blackwell.

Parikh, Prashant [2000]. Communication, meaning, and interpretation. In *Linguistics & Philosophy* 23:2:185–212.

Peacocke, Christopher [1986]. *Thoughts: An Essay on Content*. New York: Blackwell.

— [1992]. *A Study of Concepts*. Cambridge, MA: MIT Press.

Pietroski, Paul [1992]. Intentionality and teleological error. In *Pacific Philosophical Quarterly* 73:267–282.

Platts, Mark [1979]. *Ways of Meaning: An Introduction to a Philosophy of Language*. Cambridge, MA: MIT Press.

Pryor, James [2000]. The skeptic and the dogmatist. In *Nous* 34:4:517–549.

— [in preparation]. Externalism and McKinsey-style reasoning. In .

Putnam, Hilary [1968]. Brains and behaviour. In Rosenthal (1991), 151–159.

Ramsey, Frank P. [1927]. Facts and propositions. In his *Philosophical Papers*. Oxford: Basil Blackwell, 34–51.

Ramsey, William, Garon, Joseph, & Stich, Stephen [1996]. Connectionism, eliminativism, and the future of folk psychology. In *Deconstructing the Mind*, edited by Stich, Stephen. New York: Oxford University Press, 91–114.

Recanati, Francois [1986]. On defining communicative intentions. In *Mind & Language* 1:213–242.

Rosenthal, David, ed. [1991]. *The Nature of Mind*. New York: Oxford University Press.

Rumfitt, Ian [1995]. Truth conditions and communication. In *Mind* 104:827–862.

Ryle, Gilbert [1949]. *The Concept of Mind*. London: Hutchinson.

Saul, Jennifer [2002]. Speaker meaning, what is said, and what is implicated. In *Nous* 36:2:228–248.

Schiffer, Stephen [1972]. *Meaning*. Oxford: Oxford University Press.

— [1982]. Intention-based semantics. In *Notre Dame Journal of Formal Logic* 23:119–156.

— [1986]. Stalnaker's problem of intentionality. In *Pacific Philosophical Quarterly* 67:2:87–97.

— [1987]. *Remnants of Meaning*. Cambridge, MA: MIT Press.

— [1993]. Actual-language relations. In *Philosophical Perspectives* 7:231–258.

SEARLE, John [1969]. *Speech Acts*. London: Cambridge University Press.

— [1983]. *Intentionality*. New York: Cambridge U.P.

— [1986]. Meaning, communication, and representation. In Grandy & Warner (1986), 209–226.

SKYRMS, Brian [1996]. *The Evolution of the Social Contract*. Cambridge: Cambridge University Press.

SOAMES, Scott [1984]. Linguistics and psychology. In *Linguistics & Philosophy* 7:155–180.

— [1985]. Lost innocence. In *Linguistics and Philosophy* 8:1:59–72.

— [1988]. Direct reference, propositional attitudes, and semantic content. In *Propositions and Attitudes*, edited by SALMON, Nathan & SOAMES, Scott. Oxford: Oxford University Press, 197–239.

— [2002]. *Beyond Rigidity: The Unfinished Semantic Agenda of* Naming and Necessity. Oxford: Oxford University Press.

STALNAKER, Robert [1984]. *Inquiry*. Cambridge, MA: MIT Press.

— [1986]. Replies to Schiffer and Field. In *Pacific Philosophical Quarterly* 67:2:113–123.

— [1990]. Mental content and linguistic form. In his *Context and Content*. New York: Oxford University Press, 225–240.

STAMPE, Dennis [1979]. Toward a causal theory of linguistic representation. In *Contemporary Perspectives in the Philosophy of Language*, edited by FRENCH, Peter, UEHLING, Theodore, & WETTSTEIN, Howard. Minneapolis: University of Minnesota Press, 81–102.

STANLEY, Jason & WILLIAMSON, Timothy [2001]. Knowing how. In *Journal of Philosophy* 98:8:411–444.

STERELNY, Kim [1990]. *The Representational Theory of Mind*. Cambridge, MA: Blackwell.

STICH, Stephen [1990]. Building belief: Some queries about representation, indication, and function. In *Philosophy and Phenomenological Research* 50:4:801–806.

— [1994]. What is a theory of mental representation? In Stich & Warfield (1994), 347–363.

STICH, Stephen & WARFIELD, Ted, eds. [1994]. *Mental Representation: A Reader*. Cambridge, MA: Basil Blackwell.

STRAWSON, P.F. [1964]. Intention and convention in speech acts. In *Philosophical Review* 73:439–460.

SUPPES, Patrick [1986]. The primacy of utterer's meaning. In Grandy & Warner (1986), 109–129.

TALMAGE, Catherine [1996]. Meaning intentions. In *Australasian Journal of Philosophy* 74:2:341–346.

VLACH, Frank [1981]. Speaker's meaning. In *Linguistics & Philosophy* 4:359–392.

WHYTE, J. T. [1990]. Success semantics. In *Analysis* 50:149–157.

WITTGENSTEIN, Ludwig [1937-1944]. *Remarks on the Foundations of Mathematics.* Edited by G.E.M. Anscombe, G.H. von Wright, and Rush Rhees. Translated by G.E.M. Anscombe.. Revised edn. New York: MacMillan, 1994.

— [1945-1948]. *Zettel.* Edited by G.E.M. Anscombe and G.H. von Wright. Translated by G.E.M. Anscombe.. Berkeley, CA: University of California Press, 1967.

— [1953]. *Philosophical Investigations.* Translated by G. E. M. Anscombe.. 3rd edn. New York: MacMillan.

YU, Paul [1979]. On the Gricean program about meaning. In *Linguistics & Philosophy* 3:273–288.