NOÛS 40:3 (2006) 428–467

Is Mental Content Prior to Linguistic Meaning?

Jeff Speaks
University of Notre Dame

Most contemporary work on the nature of intentionality proceeds from the thesis that the fundamental sort of representation is mental representation. The purpose of this essay is to argue that, to a large extent, this starting point is mistaken. A clear view of some of the phenomena with which the philosophies of language and mind are centrally concerned—including the nature of mental content and linguistic meaning—requires taking seriously the idea that public languages can and often do serve as a vehicle for the thoughts of agents.

The picture of intentionality which informs most contemporary work on mental and linguistic representation may be brought out by considering three questions. First, there is a question about the relation between thought and language, namely

> Are facts about the beliefs, desires, and thoughts of agents prior to and constitutive of facts about the meanings of expressions in public languages, or does the order of explanation run in the opposite direction?

Inasmuch as mental states are ascribed to individuals whereas public languages are typically shared among members of a wider social group, this question about the relation between thought and language is closely related to the following question about the relationship between individuals and the societies of which they are members:

> Are social facts about communities constitutive of the capacities of their members to be in certain kinds of mental states, or are the latter largely independent of and constitutive of the former?

A third question concerns the foundation of mental content and linguistic meaning:

> Should the representational capacities of individuals be explained in terms of properties of their internal states, or in terms of the actions they are disposed to perform?

Since the 1960's, work in the analytic tradition on the nature of mental and linguistic content has converged on answers to each of these questions, which together comprise what might be called the *mentalist* picture of intentionality: the view that social facts about public language meaning are derived from facts about the thoughts of individuals, and that these thoughts—and hence, on this picture, also indirectly facts about public languages—are constituted by properties of the internal states of agents.[1]

In what follows, I shall argue that the mentalist picture goes wrong in its answers to each of these three questions and, in the final section, suggest an opposed picture of the relationship between language and the mind which avoids the problems which face mentalism.

This aim, however, leads to an immediate problem. Partly due to widespread acceptance of the mentalist picture of intentionality, there are many different and competing accounts of the nature of mental representation consistent with mentalism. Accordingly, it seems, any convincing argument against the mentalist picture should either consider all of these, or mount an argument directly against the general theses definitive of mentalism. The former is a topic fit for a book rather than an article; a convincing argument of sufficient abstraction to accomplish the latter is difficult to imagine.

Here I'll attempt a middle course. Following a discussion of the constraints on answers to questions about the nature of intentionality, I'll discuss one fundamental issue which divides mentalist theories of content: the relative priorities of belief states and 'sub-sentential' mental representations. I shall argue that mentalist theories which take the contents of belief states to be inherited from the contents of mental representations, thought of as constituents of those states, face a number of fundamental problems.

If I am right that these problems discredit the views in question, then, the field of mentalist theories thus narrowed, we will be in a position to consider an exemplar of the class of mentalist theories which do *not* take the contents of belief states to be underwritten by the contents of components of those states: the compelling view of belief presented and defended by Robert Stalnaker in his *Inquiry*.[2] The heart of the paper will be concerned with a number of what I take to be decisive objections to that theory.

First, though, I turn to the constraints on theories of intentionality, and some of the motivations for mentalist views of intentionality of the form we will be considering.

1. The Problem of Intentionality

Our target, to borrow a phrase from Stalnaker, is 'the problem of intentionality': the problem of saying what it is for something—a mental state, an expression of English, a gesture—to represent the world as being some way. This question calls for an answer which does not merely tell us contingent facts about the way that representation happens to work in our linguistic community, among humans, or even in the actual world; rather, what is sought is an account of the conditions under which, in any possible world, something represents the world as being a certain way. This is not an arbitrary constraint, but rather is a general feature of philosophical questions about the natures of things. If, for example, in moral philosophy one is trying to answer the question, "What is it for an action to be morally right?" one cannot restrict oneself to actual actions; it is clearly permissible and useful to test moral theories against situations which could have arisen, but actually have not. Since we are interested in the nature of representation, the same sort of criterion applies here. Call this the *modal constraint* on solutions to the problem of intentionality.

The thesis about intentionality to be evaluated is the priority of mental content over public language meaning:

> *The priority of mental content over public language meaning*: facts about the contents of the mental states of agents are prior to and independent of facts about the meanings of expressions in public languages spoken by those agents.

This is an intuitively appealing thesis, which can be supported by an intuitively compelling argument. It is very natural to think that there should be some connection between the meaningfulness of sentences of a language and the contentfulness of mental states of users of the language, and so very natural to think that we should either give an account of linguistic meaning in terms of mental content, or the reverse. But the datum that human infants and many animals (as well as possible creatures) intuitively have the capacity to form, for example, beliefs without sharing a language with any of their fellow creatures seems to be strong *prima facie* evidence that the second of these directions of explanation is a nonstarter. Hence the priority of mental content over linguistic meaning.

If we accept this argument, this helps to narrow down the range of possible solutions to the problem of intentionality; we can rule out theories which try to analyze the contentfulness of mental states in terms of facts about the meanings of expressions of public languages. But if this is right, then we need some independent account of the nature of intentional mental states: some account of what it is for an agent to have a belief, desire, or other mental state with a given content.

Fortunately, the modal constraint on solutions to the problem of intentionality also seems to give us some guidance here, and points to the following

thesis about the natures of these sorts of mental states (here I focus on the mental state of belief):

> *Functionalism, broadly construed*: facts about the beliefs of agents are constituted by second-order relational properties of their internal states.[3]

We can argue for this thesis as follows: if we are trying to give the nature of belief, then our account must apply to possible as well as actual believers. But then, given the multiple realizability of mental states, our account cannot be given in terms of the intrinsic properties of internal states of agents. This seems to leave only two possible positions: mental states are either constituted by dispositions to behavior, or by relational properties of internal states. But facts about belief cannot be constituted by dispositions to behavior, since, among other reasons, it is difficult to see in the case of many beliefs—such as, for example, very abstract beliefs about mathematics—what sort of behavior could be constitutive of an agent's having that belief.[4] Hence beliefs must be constituted by relational properties of internal states, and some version of functionalism, broadly construed, must be true.

This leads us to ask: what are the relational properties of internal states that constitute the mental states of agents? Here again, attention to the fact that our question is about the nature of intentionality points us to an answer:

> *Externalism*: facts about the contents of the beliefs of agents are partly determined by relations between those agents and facts external to them.

Once again, if we want to give an account of the nature of intentionality, then our account of what it is to have a certain belief must apply to possible as well as actual believers, and so must account for the difference in beliefs standardly supposed to obtain between us and our intrinsic duplicates on Twin-Earth (or the various other counterfactual scenarios imagined by externalists). The obvious way to do this is to include among the relational properties of internal states relevant to the determination of their content relations between those internal states and objects, properties, and events external to the agent in question.

On the basis of the modal constraint and the supposition that mental content is prior to linguistic meaning, we have so far given arguments for the conclusion that the right account of intentionality will have to have a fairly specific form: it will have to treat the contents of mental states rather than the meanings of expressions of public languages as basic, and will give an account of the natures of various mental states in terms of the relational properties of internal states of agents, among which will be relations between those internal states and facts external to the agents in question.

2. Belief States and Mental Representations

Here, though, the mentalist faces a choice. The sort of theory we have been developing takes having a belief to be a matter of being in an internal state with certain relational properties (to be specified by the theory). Let a *belief state* be an internal state with the relational properties required to make it a belief with a certain content. Presumably, belief states will be complex physical states. Call the parts of these belief states *mental representations*. The question which the mentalist must answer is: are the relational properties which constitute the contents of internal states properties, in the first instance, of belief states, or of mental representations?

This question can be clarified a bit by considering an analogous question with respect to linguistic meaning. Just as belief states have propositions as their contents and mental representations as their parts, so sentences of natural languages have propositions as their contents and words as their parts. And were our focus the meanings of such sentences (rather than the contents of the mental states of agents) we could ask an analogous question: are the meanings of sentences determined primarily by properties of those sentences as a whole, or by properties of the words which comprise those sentences? (For a simple and crude example of the former kind of theory, imagine a theory according to which the meaning of any sentence is the proposition belief in which would be expressed by utterances of the sentence; for a simple and crude example of the latter, imagine a theory according to which the meaning of any expression is the object or property in the world most likely to cause an utterance of that expression.)

This may not seem a particularly pressing question about mental content. The burden of this section is to argue that this appearance is misleading: mentalist theories which give primacy to mental representations (MR-theories, for short) rather than belief states are nonstarters. This will be an important step in establishing our conditional conclusion that *if* the thesis of the priority of mental content over public language meaning is correct, *then* a theory much like Stalnaker's causal-pragmatic theory of belief and desire must be as well.

2.1. Stalnaker's Objection to Mental Representation-Based Theories

Fittingly then, the main argument against the primacy of mental representations is due to Stalnaker. Recall that if we are after a solution to the problem of intentionality, then our account of what it is for an agent to have a given belief must meet the modal constraint and so make no use of contingent psychological claims particular to some proper subset of those agents. The problem, Stalnaker claims, is that 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:

> 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 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, an MR-theorist is not likely to be moved by Stalnaker's implication that her 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 MR-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 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 theory about the constituents of the states underlying our beliefs.

In response, the MR-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.[6]

But the MR-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. Rather, the MR-theorist is committed to the much stronger claim that any agent capable of having beliefs must have mental representations which are related in a certain way to the environment of the agent. Suppose for illustration that an MR-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 modal constraint, 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 MR-theorists have no promising way of giving a constitutive account of belief (or of any other sort of intentional fact, for that matter).[7]

There is, then, some reason to be skeptical about whether complex properties of mental representations should have any role to play in a constitutive account of belief. But this worry derives from modal intuitions which, trustworthy though they seem to me, would presumably be denied by MR-theorists, and are difficult to argue for. Fortunately, I think that we can strengthen Stalnaker's argument that MR-theories fail the modal constraint by being a bit clearer on the shape an MR-theory will have to take.

2.2. 'Tokening' Mental Representations[8]

The MR-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 such a 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 what is it for a mental representation to bear a relation of the right kind 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 MR-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*.

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. Using the terminology of MR-theorists, we can express this by saying that the relevant relations are between objects and properties in the world and *tokenings* of mental representations.[9] Given its centrality to MR-theories, one would like to know a bit more about what this mental state of tokening a mental representation is.

We got some grip on the notion of a belief state by saying that a belief state with content *p* is an internal state possession of which qualifies an agent as believing *p*. We can give similar glosses on the internal states underlying other propositional attitudes; e.g., a *thought state* with content *p* can be an internal state possession of which qualifies an agent as having the occurrent thought *p*. These glosses do not tell us everything we might want to know about the internal states in question, but they do give us some idea of what we are talking about when we are talking about belief states, thought states, or other propositional attitude states.

One way to sharpen our question about the nature of tokening a mental representation is to ask: is tokening a mental representation a *sui generis* state, or is it a matter of being in a belief state, thought state, or some other propositional attitude state, one constituent of which is that mental representation? Either response, I shall argue, conflicts with the modal constraint.

Suppose first that tokening a mental representation is *sui generis*, in the sense of not being a matter of being in some propositional attitude state including the representation in question. In this case, it is difficult to see how the notion can play any role in a constitutive theory of intentionality. We know that such a theory must be accountable to facts about possible as well as actual thinkers; so we know that if our account of belief is stated in terms of facts about tokenings of mental representations, it had better be a necessary rather than a merely contingent truth that all agents with beliefs also token mental representations. But if tokening a mental representation is a *sui generis* mental state distinct from being in a belief state or any other propositional attitude state, what justification can there be for believing this to be a necessary truth?

We can make this more concrete by exploring a kind of picture of tokenings which is suggested by the writings of MR-theorists. Consider the phenomenon of simply emitting 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 an utterance 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 noticeable 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 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 MR-theorist should try to define tokening in terms of occurrences of mental representations 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:

> 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 MR-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. 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 the MR-theorist faces a problem here even if we abstract away from the possibility of error. We need to restrict the thought-states occurrences of which are relevant to fixing the contents of the mental representations they include 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 far from 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 MR-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 MR-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 thoughts, 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.

Let an *I-type* proposition be a proposition whose truth requires that every property which occurs in the proposition be instantiated. Then the natural response on the part of the MR-theorist is to define some condition C on thought states which is met only by states which have I-type contents. Then she might modify the schematic account given above as follows:

> 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 problem with this idea is that the MR-theorist cannot specify condition C—which is a property of thought-*states*, not of their contents—in terms of the contents of those states. The point of MR theories of content is to explain the contents of belief states, thought states, and propositional attitude states generally in terms of the contents of mental representations. Given this, the account of the contents of mental representations had better not take for granted the contents of the propositional attitudes it was introduced to explain. For this reason, the problem the MR-theorist faces is not to define the class of I-type propositions, which is easy enough; the problem is to define the class of thought states which have I-type propositions as their contents without building into this definition facts about the contents of the states in question. This means, in effect, that the MR-theorist must find some purely syntactic property of thought-states which is a sufficient condition for such a state to have as its content an I-type proposition. But 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 *must* be any syntactic difference of the sort the MR-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 MR-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. 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 MR-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. So here too the MR-theorist fails to meet the modal constraint.

To sum up: the argument of this section presents a dilemma. On the one hand, the MR-theorist may take tokening a mental representation to be a *sui generis* mental state distinct from belief states and other propositional attitude states; but then it is implausible to think that tokening a mental representation should be a necessary condition on having beliefs. On the other hand, the MR-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 states between those which have I-type propositions as their contents and those which do not is a metaphysically necessary condition on having beliefs at all.

In the previous section, we argued from the modal constraint along with the supposition that mental content is prior to linguistic meaning to two further conclusions about the nature of intentionality: functionalism (broadly construed) and externalism. The argument of the present section allows us to add a further thesis to our mentalist theory of intentionality:

> *Priority of belief states*: the relational properties of belief states determinative of their content are relations between those states and the world, rather than between constituents of those states and the world.

With these theses on the table, we are now in a position to see why Stalnaker's account of the nature of intentionality provides the best hope for the mentalist.

3. Stalnaker's Causal-Pragmatic Account of Belief

Given that my aim is not so much to criticize the details of Stalnaker's account of belief as to use it as a way of bringing into critical focus the mentalist picture which lies in the background of that account, it will be useful to present Stalnaker's account of belief and desire as emerging naturally from these theses about the nature of mental states.

3.1. Causal Theories and Optimal Conditions

Consider a simple belief, like the belief that grass is green. The externalist thesis along with functionalism tells us that what it is for an agent to believe that grass is green is for that agent to be in some state that is related in a certain way to something external to him. The priority of belief states tells us that this relation cannot be analyzed away in favor of relations between parts of the state and objects and properties in the world. The priority of mental content over linguistic meaning tells us that the external thing to which the belief state bears the relevant relation is not a sentence of a public language which means that grass is green. Once this option is ruled out, a natural alternative is to take the agent to be in some state which is related in some way to the fact that grass is green itself. With this on the table, it is a further step—but, again, a natural one—to regard this relation as a causal one.

Our four theses about intentionality have led us, then, to a simple causal theory of the following sort:

> 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.

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,"[10] 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.[11] 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.[12] 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*.

3.2. The Pragmatic Half of the Causal-Pragmatic Theory

This modification to the simple causal theory, however, is not enough to solve another problem: many states of agents indicate things but 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.[13]

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.[14]

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

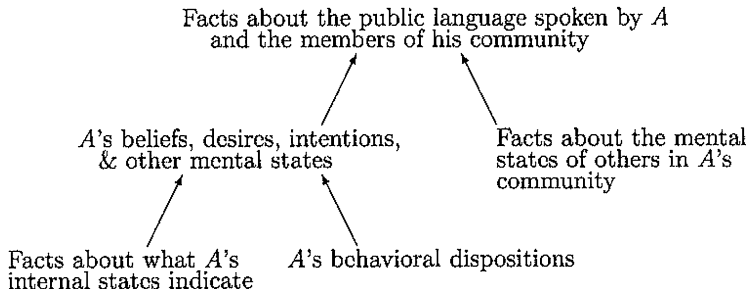
> To desire that *P* is to be disposed to act in ways that would tend to bring it about that *P* in a world in which one's beliefs, whatever they are, were true. To believe that *P* is to be disposed to act in ways that would tend to satisfy one's desires, whatever they are, in a world in which *P* (together with one's other beliefs) were true.[15]

For Stalnaker, then, what it is for an agent to have a certain belief is for that agent both 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 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.

Were some account of this sort correct, its philosophical interest would be considerable. For, as Stalnaker points out, we would then have not only an account of belief given solely in terms of causal relations and dispositions to action, but also an account of desire in terms of the same class of facts; and, using these, it is not entirely implausible to think that we might be able to give an account of what it is for an agent to have a certain sort of intention and to ascend from there via a broadly Gricean strategy to an account of the meanings of expressions and gestures in public systems of communication.[16] We would then have constructed, using relatively meager building blocks, an account of the nature of and relations between a whole class of concepts fundamental to the philosophies of language and mind. Letting '*A*' stand for an arbitrary agent, this version of the mentalist picture might be represented as follows:

THE MENTALIST PICTURE



This completes the argument for the conclusion that *if* mental content is prior to public language meaning, *then* there are strong reasons for thinking that Stalnaker's view of the nature of the contents of beliefs and other mental states must be correct.

In the next section, I shall present five arguments for the conclusion that Stalnaker's causal-pragmatic account of belief and desire cannot play this foundational role.

4. Five Problems for the Causal-Pragmatic Theory

4.1. The Conjunction Problem

It is widely agreed that Stalnaker's appeal to optimal conditions makes room for false beliefs, and so solves the disjunction problem. 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 p. On Stalnaker's view, there must be some belief state b of a which indicates p, so that, if we let 'O' abbreviate the predicate 'is in optimal conditions,' the following claim is true:

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

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

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

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 p is the case. But, of course, p is not the whole explanation for a's being in state b. It could have been the case that p 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 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 *p* is true *and a* is in optimal conditions. 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*,

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

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.[17] 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.[18]

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 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.[19] 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 both p and the proposition that the agent is in optimal conditions are true, 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 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.[20]

4.2. 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 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.

An internal state x indicates a proposition p (for an agent) iff in the nearest possible world in which that agent is in optimal conditions *and in that state*, the agent is in the state because p is the case. Evidently, this definition of indication presupposes a relation of sameness of an internal state across possible worlds. This raises the question: When we are asked to imagine the causes of *this state* in another possible world, what exactly are we being asked to imagine? A natural thought is that we are asked to imagine the causes of the agent in question coming to be in a state of the same *physical type* as the internal state which underlies the relevant belief in the actual world. We cannot, for example, say that sameness of belief states across possible worlds is determined by the *contents* of those states; on the present view, the indication relation defines content, and the indication relation relies on rather than explains the intended relation of sameness of internal states across possible worlds.

But this point brings out a problem with the appeal to counterfactuals in the theory of content. We know that physical states have their content only contingently; so the assumption that one of my belief states has the same content in the actual world as in some possible world is a substantive one. Indeed, if the differences in the agent's cognitive system or environment are sufficient, we should expect this assumption to be false. The indication theory of content, however, assumes that for any belief state x of an agent, the content of x will be the same in the actual world as in the nearest possible world in which the agent in question is in x and in optimal conditions. But there is no reason to think that this will in general be true; and indeed, we can come up with cases in which it pretty clearly fails.

All we need do is consider an example of an agent with a false belief about her own cognitive system; consider, for example, an agent's belief that her brain is made of silicon. Let b be the belief state which underlies this belief. On the assumption that the causal-pragmatic theory is true (and hence that b indicates the proposition that her brain is made of silicon), we know that in the nearest possible world w in which the agent in question is in b and in optimal conditions, the agent must be in b because her brain is made of silicon. But there is no such possible world w; tokens of b are of the same physical type as an actual state of the agent's brain, and the actual agent's brain is not made of silicon. So there is no possible world in which the agent is in b and it is true that her brain is made of silicon. 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.[21]

Now, presumably it is possible for the agent to both be in *b* and be in optimal conditions; the example only shows that in the nearest world in which this is the case, *b* does not have the same content as it has in the actual world. 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.

4.3. 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 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*.[22]

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.

The *prima facie* problems for this thesis about belief are well known; here I'll rehearse them briefly.[23] 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 believes $P \equiv a$ believes P & Q)

from which it follows, given the distribution of belief over conjunction,[24] that

$\Box$ (a believes $P \rightarrow a$ 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 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.[25]

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 the objects of belief as sets of possible worlds, 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

> [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.

seems clearly false. 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 it 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[26]

> [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 make room for 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 seems to be enough to negate

any advantage gained by the appeal to meta-linguistic propositions.[27] 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

> [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 is 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].[28] 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;[29] 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.[30]

4.4. Indeterminacy and the Pragmatic Account of Belief States

So far we have focused on the indication theory of content; but there are important reasons for doubting the pragmatic theory of belief states as well.

The causal-pragmatic theory 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 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."[31] 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. The causal aspect of the causal-pragmatic account 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 rule out states like the reflectance properties of a bald man's head from being belief states. But an 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.

This suggests a modification of the causal-pragmatic theory. Perhaps the two halves of the theory should be more tightly bound; rather than just conjoining the indication theory of content with the pragmatic theory of belief states, perhaps we should require that, for an agent to believe *p*, she

should be in some internal state which both indicates *p* and is (part of) the causal basis of her dispositions to act in ways which would satisfy her desires in a world in which *p* and her other beliefs are true. This seems to help with the above case, since it seems plausible that the reflectance properties of the bald man's head are not a part of the causal basis of his running toward the outfield wall.

But, for two reasons, this apparently promising reply is not satisfactory.

First, any counterexample to the original causal-pragmatic theory can be turned into a counterexample to the strengthened theory by considering the fusion of the state which indicates the proposition in question with some other state which is a part of the causal basis of the relevant dispositions. Such 'gerrymandered' states will always be available; in the case above, we might consider the state which is the fusion of the reflectance properties of the bald man's head with some other state directly involved in the production of the relevant behaviors. No doubt, this seems like cheating; but there is a serious point here. It is tempting to try to get round this new version of the problem by excluding such gerrymandered states from consideration by stating further constraints on what sorts of states can be belief states. By doing so, however, the proponent of the causal-pragmatic theory, like the MR-theorist, risks running afoul of the modal constraint. It is far from obvious that any constraints on belief states can be both strong enough to exclude different sorts of gerrymandered states and weak enough to avoid ruling out the possibility of agents with very different cognitive systems than ours having beliefs.

Second, we can find counterexamples similar in kind to the 'bald man' example in which the state which does the indicating also plays a role in the production of the relevant behavior. 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.

If cases of states which indicate something but are not belief states 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, as the example of skin temperature shows, 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.

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.

4.5. Belief and Language Use

Each of the four preceding sections have developed a problem for the causal-pragmatic picture of belief. A final objection to the account, I shall argue, goes some way toward showing why each of the preceding four arise. The objection, put simply, is that there are systematic connections between the linguistic behavior of agents, the meanings of expressions in the public language which they speak, and their beliefs; accordingly, any picture of belief which, like Stalnaker's, denies linguistic meaning a role in constituting facts about belief is bound to go astray. Such theories, if I'm right, are simply looking for belief-constituting facts in the wrong place.

The connection between language and belief I have in mind is just this: if one sincerely accepts a sentence that means p, and one understands the sentence, one thereby comes to believe p. This disquotational principle may be expressed as follows:[32]

> Necessarily, if an agent accepts a sentence that means p, then the agent believes p.

Now recall our statement of the causal-pragmatic account of belief:

> 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

Together, the disquotational principle and the causal-pragmatic account entail that it is a necessary truth that, whenever an agent accepts a sentence that means p, then that agent is in an internal state which, under optimal conditions, he would be in 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, were he in that state and in optimal conditions, he would be in that state because of the fact that (in the optimal conditions world under consideration) the 14th president of the United States was Franklin Pierce? Note that the state in question can't be a disposition to accept a sentence with a certain meaning; part of the point of the causal-pragmatic account of belief was to give a foundational account of what it is for an agent to believe *p* which can serve as the most basic level of a mentalist picture of intentionality on which beliefs are constitutive of, and hence independent of, facts about public language meaning.

Aside from this disposition, it seems 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.[33]

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."[34] 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.[35] 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.

The proponent of the mentalist picture of intentionality is likely to respond by claiming, first, that these cases—in which the contents of an agent's beliefs do seem to depend on the meaning of which public language sentences he accepts—are exceptional, and, second, that in these exceptional cases, there are mechanisms — which can themselves be explicated in terms of facts about mental content—which explain away this seeming dependence of beliefs on facts about public languages. This is likely to take the form of an appeal to *deference* or *the division of linguistic labor*. As Mark Greenberg has pointed out in an important paper, this amounts to giving a disjunctive account of the nature of thought; the theorist in question is claiming that one can have a belief *p either* by satisfying the causal-pragmatic theory (or some variant thereof) *or* by deferring to other agents.[36] We can express this as the claim that we should give separate theories of beliefs of which we have 'full grasp' and beliefs which are 'deference-dependent'.

But, as Greenberg argues, there is reason to doubt both the plausibility of the division of beliefs into cases of full grasp and of deference-dependence, and the efficacy of the appeal to deference.

Pre-theoretically, this division of cases seems to have limited appeal. The claim that cases like that of Bob should be set to one side as cases of only partial grasp of the proposition believed presupposes that we should not count competence with a sentence as a sufficient condition for grasping the proposition expressed by that sentence. But there is good reason to doubt whether any distinction between full and partial grasp of a proposition which does not count competence with a sentence as sufficient for full grasp should

have any special theoretical role to play. Among the agents to whom we may truly attribute beliefs involving the concept expressed by "is president," 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 the various propositional attitudes characteristic of mental content, 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 this concept and those to whom we cannot.

Furthermore, as Greenberg emphasizes, there seem to be cases in which we are willing to attribute beliefs to agents which do not fit comfortably into either of the categories of 'full grasp' or 'deference-dependent' beliefs.[37] In one such case, discussed by Burge, an agent (who is otherwise like standard examples of agents having deference-dependent thoughts) might develop 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.[38] 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, as Greenberg points out, 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. This seems to be a case which evades both of the disjuncts of the modified mentalist account we have been considering. But if one allows linguistic meaning to play a role in the determination of mental content, such cases pose no serious problem: the agent's belief may be constituted by her disposition to accept a sentence of her public language which means that, for example, sofas are religious artifacts.

Furthermore, even if we suppose that some principled and exhaustive division of beliefs into cases of full grasp and of deference-dependence is possible, we still need an account of what deference is. This task is more pressing than is often realized. Deference is often invoked as a kind of unexplained explainer; but, as Greenberg's discussion shows, this kind of reliance on deference is far from innocent. While 'deference' does have two obvious interpretations in this context, neither is well-suited to the defense of a mentalist picture of intentionality.

On one hand, talk of deference or the division of linguistic labor might just amount to the claims (i) that the meanings of sentences in public languages are social facts typically determined by factors outside the control of any one agent and (ii) that, for this reason, normal membership in a linguistic community involves a significant extension of an agent's ability to have certain kinds of thoughts and form certain kinds of beliefs. This kind of platitudinous reading of the appeal to deference fits well with the fact that theorists often treat deference as an unexplained explainer not in need of much analysis; however, it is not an understanding of deference to which a proponent of

the priority of mental content over linguistic meaning can appeal, since it explains the contents of some beliefs of agents in terms of social facts about public languages.

More often, the appeal to deference is offered as a way of saving the mentalist picture in the face of apparent counterexamples. On this second interpretation, the aim would be to explain the content of Bob's new belief *without* appeal to the meaning of the sentence Bob came to accept. But if the content of this belief is not to be explained by facts about what Bob's internal states indicate, and is not to be explained by the meaning of the sentence he accepts, then it seems that it must be explained by the contents of some other of Bob's mental states—in this connection, it is standard to speak of *deferential intentions*. But what intentions could these be? One is tempted to rely on an intention to form the belief that Franklin Pierce was the 14th president of the United States; but this is just to push the bump in the rug. We are looking for an explanation of how Bob could have a belief with a certain content; but, by appealing to an intention whose content includes the content of the belief to be explained, we make use of a fact which requires the same sort of explanation. Other candidate deferential intentions raise problems of their own.[39] For these reasons, it seems to me unlikely that the appeal to deference provides the defender of the priority of belief over language with the resources necessary to handle the necessary connections between language use, linguistic meaning, and belief.

Suppose, then, that we set aside the appeal to deference and take seriously our ordinary attributions of thoughts and beliefs to agents. One might still find the idea that we can arrive at counterexamples to the causal-pragmatic theory (or other versions of functionalism, broadly construed) on the basis of facts about language use a bit mysterious; how *could* something like language use cause facts about beliefs of agents to float free of facts about what internal states of agents indicate? 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 and 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. Kripke's examples indicate that all that is required for understanding an expression is satisfaction of minimal communal standards of use; if this is true, then 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 should 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 a mentalist to explain their common belief.

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 and 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—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.[40] The present point is just a generalization of Kripke's claims about the contents of proper names to the contents of public language expressions more generally, and from there, via the disquotational principle, to the contents of thoughts.[41]

5. An Alternative Picture of Intentionality

We began by presenting the causal-pragmatic account of belief as a consequence of four intuitively plausible theses about belief. If the conclusion of the preceding section is correct, then we must reject at least one of these theses: the priority of the contents of beliefs over public language meaning. To reject this is already to depart from the mentalist picture; but, by itself, this appeal to public language meanings does not solve all the problems with the causal-pragmatic account of belief. Even if it goes some distance in solving the problems with the causal account of content, it does not remove the difficulties with the pragmatic account of the attitude of belief, discussed in §4.4 above.

Consider in a very schematic way how an account of belief broadly similar to Stalnaker's could be revised to make room for the constitution of beliefs by facts about public language meanings. According to the causal-pragmatic account, for an agent to believe p, that agent must be in some state which bears a certain relation R to p. If we relax this requirement so that what is required is that the agent be in some state which either bears R to p or stands in a relation R' to a sentence S which means p, our account would then look something like this:

> Necessarily, an agent believes p iff
>
> (i) there is some state x of the agent such that either $R(x, p)$ or $(R'(x, S)$ & S means $p)$, &
> (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

Recall that the reason why clause (ii) of this account was needed in the causal-pragmatic account was that some states can indicate a proposition without being belief states; indeed, virtually *all* states of an agent will meet

this description. The problem is that, absent any more information about the relation R' between internal states and sentences, it is reasonable to think that it will also turn out to be possible for a state to bear R' to a sentence which means p without thereby being a belief state. If this is so, then the problem of §4.4 will recur; namely, the fact that the criterion for a state's being a belief state makes use of facts about the desires of agents will lead to an implausible and widespread indeterminacy of facts about what agents believe and desire.

One response to this is to jettison the pragmatic account, and its attempt to give an account of the attitude of belief states in terms of dispositions to behavior, altogether. In its place, one might put an account which gives, for lack of a better word, more 'psychological' conditions on belief states; perhaps an account which says more about the properties of internal states of agents, and what is required for such a state to be a belief state, than does the pragmatic account. Such a view, however, would face the same sort of dilemma as did the appeal to mental representations in §2 above: it would have to be both specific enough to count only belief states as such, and at the same time be general enough to, in accord with the modal constraint, apply not only to humans but to all possible agents capable of forming beliefs. The prospects for meeting both of these constraints do not seem good.

A partial solution to this problem is to build into our account, not only facts about linguistic meaning, but also facts about linguistic actions. We should then not require that an agent be in some state which bears some, presumably causal, relation to a sentence which means p, but simply that the agent be disposed to accept a sentence which means p. The idea is to find an attitude toward *sentences*—here, the attitude of acceptance—which is the analog of an attitude toward *propositions*—here, the attitude of belief. It can then be a sufficient condition on having the relevant attitude toward a proposition p that the agent have the corresponding attitude toward a sentence which means p. This both shares with the pragmatic theory of belief states its greatest virtue, and avoids its most serious vice. Because it is stated purely in terms of behavioral dispositions, as an account of the attitude of belief it shares the virtue of being independent of contingent psychological claims, but does so without entailing that the beliefs and desires of agents are radically indeterminate. If this is right, then we should reject not only the thesis of the priority of belief over public language, but also the functionalist thesis that facts about belief are constituted by second-order properties of internal states rather than by dispositions to perform certain sorts of actions.

I said that this use of sentential attitudes is only a partial solution to our problems; and so it is, for two reasons. First, though we might by this route secure sufficient conditions for an agent believing p, it is not at all plausible that we also give necessary conditions. Agents—animals and infants, for example—can have beliefs without being members of linguistic communities, and it is plausible that agents who are members of linguistic communities can have beliefs that are not expressed by any sentence which they are disposed

to accept. Second, an account of thought more generally will be interested, not only in what it is for an agent to have a given belief, but also in what it is for an agent to desire, intend, think, or imagine something; and, while in the case of belief the attitude of acceptance toward sentences is ready to hand, it is not at all clear that, in the case of other propositional attitudes, corresponding sentential attitudes are anywhere to be found.

The moral of each problem is that an account of this sort cannot, in explicating various kinds of thought, rely on linguistic behavior alone. Any plausible constitutive account of mental states of the sort being suggested will have to include the non-linguistic actions into which linguistic behavior is integrated.

An account of this kind faces a number of serious challenges; here I can only briefly discuss what I take to be the most fundamental of them:

1. The suggested account employs facts about public language meaning in the explanation of the contents of beliefs; but this leaves public language meaning unexplained.
2. Any account of a mental state in terms of dispositions to action faces a dilemma: are the dispositions to which the theory appeals dispositions to certain ('non-intentional') bodily movements, or to intentional actions? If the former, then the resources of the theory are surely too sparse; if the latter, the theory presupposes the very mental states it aims to explicate.

I think that these two challenges are related. The most plausible response seems to me to be to take the second horn of the dilemma about intentional action, and use it to respond to the first challenge. The idea that we can give an account of the meanings of expressions in public languages in terms of dispositions of speakers to use those expressions in various ways is much more plausible if the dispositions in question are dispositions to assert sentences, ask questions, etc., rather than bare dispositions to emit sounds in various circumstances.[42]

But this leaves the worry about intentional action expressed in the second challenge unanswered: how are we to explain what it is for an agent to undertake a certain intentional action, such as asserting or accepting a sentence, if not by appeal to bodily movements caused, accompanied, or rationalized by certain beliefs, desires, and intentions? If there is no alternative to explanation of the nature of intentional action in terms of these propositional attitudes, then an account of mental states of the kind being suggested seems straightforwardly circular.

One response is to accept the circularity and call it 'interdependence'; but I think that we can do better. For we can pose a version of the second challenge about intentional action to the causal-pragmatic version of functionalism we have been discussing. According to this theory, a state which indicates *p* will be the belief *p* only if the agent in question is disposed to act so as to satisfy

her desires in a world in which *p* and her other beliefs are true. So this theory too makes use of facts about the behavior of agents, and we can ask: is this 'behavior' a matter of the intentional actions undertaken by agents, or merely their bodily movements?

It seems to me that, in order for the theory to be tenable, the behavior made use of by the theory must be restricted to intentional actions of the agents in question. Consider a stick, the surface of which indicates the temperature of the surrounding air. If a wind comes along and blows the stick a bit to the left into some shade, there is no obvious way, short of pointing out that the stick has not *acted*, to block the unwanted result that the stick 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 non-intentional 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. All of this indicates that the proponent of the causal-pragmatic theory, like the proponent of the view I have been developing in this section, must take facts about the intentional actions of agents to be prior in the order of explanation to facts about beliefs and other propositional attitudes. Insofar as this argument generalizes to other forms of functionalism, and insofar as one thinks that there must be some substantive story to be told about what it is for an agent to have a belief with a given content, this amounts to an argument for the priority of action over belief. And this, if true, is enough to show that the second challenge above is not decisive.

This communitarian picture opposes the mentalist picture on each of the three general issues raised at the outset of this essay: the relative priorities of thought and language, of individual and society, and of the importance of behavior and of internal states in understanding the nature of intentionality. To be sure, this is only a picture, and vague even so far as pictures go. The aim of this paper has not been to argue for its correctness, but only to argue that the problems faced by the more well-worked out mentalist picture of thought and language suffice to show that the communitarian picture is an alternative which deserves to be taken seriously.[43]

Notes

[1] Prominent versions of this mentalist picture of intentionality may be found in Lewis, *Convention*; Schiffer, *Meaning*; Loar, *Mind and Meaning*; Evans, *Varieties of Reference*; Fodor, *A Theory of Content and Other Essays*; Peacocke, *A Study of Concepts*.

[2] Stalnaker, *Inquiry*.

[3] This is functionalism in a broad sense; it includes as a special case the stronger thesis which identifies the relevant second-order properties with functional roles.

[4] One might want to reply that dispositions to accept mathematical sentences with certain meanings might be constitutive of having these beliefs; while, as will become clear, I think that this is a plausible view, it is not open to a theorist who adopts the thesis of the priority of mental content over public languages.

[5] Stalnaker, 'Mental Content and Linguistic Form', p. 230.

[6] One might doubt this 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 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? 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. (See, e.g., Soames, *Beyond Rigidity,* Ch. 10, "What do Natural Kind Predicates Have in Common with Proper Names?", especially pp. 281 ff.) No doubt this is 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.

[7] 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 no 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 the present objection is right, then one interested in questions like 'What is the nature of belief?' or 'What is it for an agent to represent the world as being a certain way?', answers to which must meet the modal constraint, should not look to mental representations and their second-order properties for answers.

[8] I owe the idea that the notion of 'tokening a mental representation' might be a problematic one to Mark Greenberg, and his seminar on Mental Content in the Fall of 2000 at Princeton.

[9] The following quote from Jerry Fodor is representative:

> 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*. (Fodor, 'A Theory of Content, II: The Theory', p. 91)

Of importance for now are not the details of the mind-world relations in terms of which Fodor explains mental content, but rather the mental events which stand at one end of this relation. 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.

[10] See Fodor, *Psychosemantics*.

[11] See especially Stampe, 'Toward a Causal Theory of Linguistic Representation'. Relevantly similar views may be found in Dretske, *Knowledge and the Flow of Information;* Fodor, 'Psychosemantics or Where Do Truth Conditions Come From?'

[12] 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. For some skepticism about the possibility of giving a non-circular specification of optimal conditions which will meet this constraint, see Schiffer, 'Stalnaker's Problem of Intentionality'.

[13] Stalnaker, *Inquiry*, p. 18.

[14] Stalnaker, *Inquiry*, p. 19.

[15] Stalnaker, *Inquiry*, p. 15.

[16] Though it is clear that Stalnaker endorses the Gricean strategy of giving an account of public language meaning partly in terms of intentions (Stalnaker, *Inquiry*, pp. 32–33), it is not

clear how he thinks intentions fit into his causal-pragmatic picture of belief and desire. It seems likely that he would be attracted to the idea of giving an account of intentions either in terms of beliefs and desires, or in terms of beliefs and desires along with behavioral dispositions.

[17] 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.

[18] Note that, because the proposition that an agent is in optimal conditions is not typically a necessary consequence of values of 'p', this is not a version of the well-known 'problem of deduction' for possible worlds semantics. More on this problem in §4.3 below.

[19] 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. It is hard to imagine what optimal conditions would have to be like in order to satisfy this requirement.

[20] Yet 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. One might think, however, that this intuition 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.

[21] This counterexample is similar to the cases of altering discussed in Appendix 2 of Johnston, 'Objectivity Refigured: Pragmatism without Verification'. 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.

[22] Stalnaker, *Inquiry, p. 24*

[23] The objections are drawn from Soames, 'Lost Innocence' and Soames, 'Direct Reference, Propositional Attitudes, and Semantic Content'. 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; more on this below.

[24] For a brief discussion of the distribution of belief over conjunction, see §4.1 above.

[25] 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.

[26] There is some question what the nature of meta-linguistic propositions is supposed to be. Sometimes, Stalnaker takes them to be about "the relation between a proposition . . . and its content" (Stalnaker, 'Replies to Schiffer and Field', p. 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.

[27] Field first made this point in 'Mental Representation', pp. 38–9; he develops it further in 'Stalnaker on Intentionality', p. 111.

[28] 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.

[29] See, for example, Stalnaker, *Inquiry*, p. 76.

[30] 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 compartmentalization 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.

[31] Stalnaker, *Inquiry*, p. 19.

[32] 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.

[33] 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.

[34] This is an extension of the well-known thought-experiments of Burge, 'Individualism and the Mental'.

[35] 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 cannot do.

[36] See Greenberg, 'Incomplete Understanding, Deference, and the Content of Thought'. Many of the central ideas of this paper can also be found in Greenberg, *Thoughts Without Masters: Incomplete Understanding and the Content of Mind'*. For a useful discussion by a proponent of this kind of disjunctive theory, see Peacocke, *A Study of Concepts*, pp. 27–33 though Peacocke focuses on possessing concepts rather than believing propositions.

[37] For a much fuller discussion of issues involving the appeal to deference, see Greenberg, 'Incomplete Understanding, Deference, and the Content of Thought'.

[38] Burge, 'Intellectual Norms and the Foundations of Mind'. The use to which I put Burge's example here is due to Greenberg rather than to Burge.

[39] One can't appeal to the intention to accept a sentence with its usual meaning, since this again relies on social facts about sentence-meaning in the explanation of mental content. One can't appeal to the intention to form a belief with the same content as someone else's belief state, since most people have no intentions about belief states at all. Again, for fuller discussion, see Greenberg 'Incomplete Understanding, Deference, and the Content of Thought'.

[40] See, for examples, the discussions of Schmidt and Gödel, and Peano and Dedekind, in Kripke, *Naming and Necessity* pp. 83–85.

[41] It is worth briefly mentioning a different sort of response to the arguments against the causal-pragmatic account. One might respond by limiting the ambitions of the theory, and conceding that the theory cannot account for *all* the beliefs of agents. For example, the above arguments show that it cannot account for linguistically mediated beliefs, and for beliefs about the states which underlie one's own beliefs; but the causal-pragmatic theory might still be a good account of some subset of our beliefs or more primitive information-bearing states, and these

might yet be well-suited to play the foundational role required by the mentalist picture. Three points indicate that this is not a promising route for the mentalist to take: (i) The discussions of the conjunction problem, the problems associated with the view of the objects of belief as sets of possible worlds, and the problem of the indeterminacy of beliefs and desires show that the causal-pragmatic theory not only cannot account for all of our actual beliefs, but also entails that we have all sorts of beliefs which we do not in fact have. (ii) The class of linguistically mediated beliefs is, in the case of language-using adult human beings, extremely large; plausibly, these cases are too central to be set to the side by any account of belief. (iii) The foundational role given to belief and other mental states within the mentalist picture requires that these states should be suitable to give an account of what it is for an expression in a public language to have a certain meaning. But the complex literature which has grown up around such Gricean and neo-Gricean accounts of meaning shows that this is no easy task even if we take on board all of our beliefs and intentions; there is every reason to believe that the task will prove impossible if we limit ourselves to some primitive subset of our beliefs.

A different way of modifying the causal-pragmatic theory within the mentalist picture which I have not discussed is to take *both* internal relations between states of an agent and relations between those states and facts external to the agent to be constitutive of the contents of those states, as in Loar, *Mind and Meaning*. (This might also be done in the context of a theory which takes mental representations as more fundamental than belief states, as in Harman '(Nonsolipsistic) Conceptual Role Semantics'.) Such theories deserve a fuller discussion than I can give them here. However, two critical points are worth making: (i) Many such theories give accounts of perceptual beliefs which make use only of relations to the external world; for this class of beliefs, the objections in the text will hold even for this sort of mixed theory. (ii) Such theories are no better positioned to account for the necessary connections between linguistic behavior, linguistic meaning, and belief than are the sorts of accounts mentioned in the text.

[42] In my view, this idea is made more plausible by the failure of competing attempts to explain the nature of meaning in public languages in terms of the beliefs and intentions of speakers. I hope to develop this point in future work.

[43] Thanks for helpful comments on previous versions of this paper to Scott Soames, Mark Greenberg, Gideon Rosen, Jim Pryor, Paul Benacerraf and the members of Princeton's Dissertation Seminar in the Spring of 2002, and an anonymous reviewer for *Noûs*.

References

Burge, Tyler, 'Intellectual Norms and the Foundations of Mind', *Journal of Philosophy*, 83:12 (1986), pp. 697–720.

Burge, Tyler, 'Individualism and the Mental', in: Ludlow, Peter and Martin, Norah, editors, *Externalism and Self-Knowledge*, (Stanford, CA: CSLI Publications, 1998), pp. 21–83.

Dretske, Fred, *Knowledge and the Flow of Information*, (Cambridge, MA: MIT Press, 1981).

Evans, Gareth; McDowell, John, editor, *Varieties of Reference*, (Oxford: Oxford University Press, 1982).

Field, Hartry, 'Mental Representation', in: Stich, Stephen and Warfield, Ted, editors, *Mental Representation: A Reader* (Cambridge, MA: Basil Blackwell, 1978), pp. 34–77.

Field, Hartry, 'Stalnaker on Intentionality', *Pacific Philosophical Quarterly*, 67:2 (1986), pp. 98–112.

Fodor, Jerry, 'Psychosemantics or Where Do Truth Conditions Come From?' in: Lycan, William, editor, *Mind and Cognition*, (Oxford: Basil Blackwell, 1980), pp. 312–338.

Fodor, Jerry, *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*, (Cambridge, MA: MIT Press, 1987).

Fodor, Jerry, *A Theory of Content and Other Essays*, (Cambridge, MA: MIT Press, 1990).

Fodor, Jerry, 'A Theory of Content, II: The Theory', in: Fodor, *A Theory of Content and Other Essays*, pp. 89–136.

Greenberg, Mark, *Thoughts Without Masters: Incomplete Understanding and the Content of Mind*, (Dissertation, Oxford University, 2000).
Greenberg, Mark, 'Incomplete Understanding, Deference, and the Content of Thought', (unpublished ms.).
Harman, Gilbert, '(Nonsolipsistic) Conceptual Role Semantics', in: Harman, *Reasoning, Meaning, and Mind*, pp. 206–232.
Harman, Gilbert, *Reasoning, Meaning, and Mind*, (Oxford: Clarendon Press, 1999).
Johnston, Mark, 'Objectivity Refigured: Pragmatism without Verificationism', in: Haldane, John and Wright, Crispin, editors, *Reality, Representation, & Projection*, (New York: Oxford University Press, 1993) pp. 85–130.
Kripke, Saul, *Naming and Necessity*, (Cambridge, MA: Harvard University Press, 1972).
Lewis, David, *Convention*, (Cambridge, MA: Harvard University Press, 1969).
Loar, Brian, *Mind and Meaning*, (Cambridge: Cambridge University Press, 1981).
Peacocke, Christopher, *A Study of Concepts*, (Cambridge, MA: MIT Press, 1992).
Schiffer, Stephen, *Meaning* (Oxford: Oxford University Press, 1972).
Schiffer, Stephen, 'Stalnaker's Problem of Intentionality', *Pacific Philosophical Quarterly*, 67:2 (1986) pp. 87–97.
Soames, Scott, 'Lost Innocence', *Linguistics and Philosophy*, 8:1 (1985), pp. 59–72.
Soames, Scott, 'Direct Reference, Propositional Attitudes, and Semantic Content', in: Salmon, Nathan, and Soames, Scott, editors, *Propositions and Attitudes*, (Oxford: Oxford University Press, 1988), pp. 197–239.
Soames, Scott, *Beyond Rigidity: The Unfinished Semantic Agenda of* Naming and Necessity, (Oxford: Oxford University Press, 2002).
Stalnaker, Robert, *Inquiry* (Cambridge, MA: MIT Press, 1984).
Stalnaker, Robert, 'Replies to Schiffer and Field', *Pacific Philosophical Quarterly*, 67:2 (1986) pp. 113–123.
Stalnaker, Robert, 'Mental Content and Linguistic Form', in: his *Context and Content*, (New York: Oxford University Press, 1990), pp. 225–240.
Stampe, Dennis, 'Toward a Causal Theory of Linguistic Representation', in: French, Peter, Uehling, Theodore and Wettstein, Howard, editors, *Contemporary Perspectives in the Philosophy of Language*, (Minneapolis: University of Minnesota Press, 1979), pp. 81–102.