Chapter 5

A Content Theory of Belief

At the end of chapter 3, I drew a distinction between two sorts of mental sentence theories, a distinction which turned on how they handle the problem of typing mental sentence tokens. The burden of chapter 4 was that things do not look promising for theories invoking a narrow causal account of typing. In the present chapter I will look at the other alternative, the one which types tokens according to their content. As indicated earlier I think this is the right place to look. What I propose to do is to set out in some detail a theory in the spirit of the mental sentence theory, with tokens typed by content, and to argue that this theory provides plausible explanations for all the data we have assembled about our folk psychological concept of belief. In saying that the theory I will defend is in the spirit of mental sentence theories, I intend "spirit" to be interpreted rather broadly. For there are many points, some of considerable importance, on which I part company with these theories. Indeed it will prove convenient to set out my theory by first reviewing the essential points of the content version of the mental sentence theory and then noting in detail where I differ and why.

Before setting out on this task, a word is in order on the intended scope of my analysis. Just about all of the recent philosophical discussion of belief has been wedded to the view that there are two distinct senses of the word 'believes': the *de dicto* or notional sense, and the *de re* or relational sense. On this view, belief sentences are generally ambiguous, since they may be understood as invoking either the *de dicto* or the *de re* sense of 'believes'. A great deal has been written on how these senses are to be characterized and on whether one can be defined in terms of the other. However, it is my contention that no matter how the putative senses are characterized, the ambiguity thesis is simply mistaken. *In their ordinary usage 'belief', 'believes', and related terms are not ambiguous.* And though belief sentences may sometimes be ambiguous, the ambiguity is always to be traced to some other component of the sentence. The verb 'believes' introduces no systematic ambiguity. Thus I intend the account developed below to describe the single,

univocal ordinary meaning of 'believes'. If I am right in rejecting the claim that ordinary language belief sentences are systematically ambiguous, however, then the vast majority of philosophers who have written about belief in the last three decades are wrong, and some explanation is in order on just how they were led astray. In the next chapter I will take a look at the evidence and arguments that have been offered for the ambiguity thesis, and I shall argue that the data can be better explained by the theory developed below.

1. Content Mental Sentence Theories

Let me begin by sketching the essential points of mental sentence theories which type tokens according to content. Figure 2 depicts two people: A (the attributor) is uttering 'S believes that p' and thereby attributing to S the belief that p. S is the person about whom A is speaking. The aim of the theory is to explain just what A is saying about S—to analyze the meaning of his assertion.

The first point to note is that A is saying that S has (or is in) a certain sort of psychological state, viz., a belief. The mental sentence theory makes two rather different claims about states of this sort. First, to count as a belief at all, a state must play a certain sort of causal role in the mental dynamics of the subject. This causal role can be recounted in terms of characteristic causal interactions with stimuli and with other categories of mental states, which are themselves characterized by their interactions with beliefs, with each other, with stimuli, and with behavior. Thus to attribute any belief to a subject is to presuppose that the subject's psychology exhibits an overall pattern or global architecture along the lines depicted in figure 2. (A, of course, is presumed to have the same global psychological architecture, though details have been omitted from my sketch.) The second thing that mental sentence theories tell us about beliefs is that they are sentence tokens, either in the language of the subject or in a species wide language of thought. A bit more precisely, mental sentence theories claim that to have a belief is to have a sentence token inscribed in the brain in such a way that it exhibits the causal interactions appropriate to beliefs. Generally mental sentence accounts tell a parallel story about desires: to have a desire is to have a sentence token represented in the brain in such a way that it exhibits the causal interactions appropriate to desires.

In saying 'S believes that p', of course, A is not merely telling us that S has some belief or other. He is using the content sentence, 'p', to specify which belief it is. In explaining how this specification works, the content version of the mental sentence theory invokes a pair of relations to link the content sentence to S's belief. The first of these

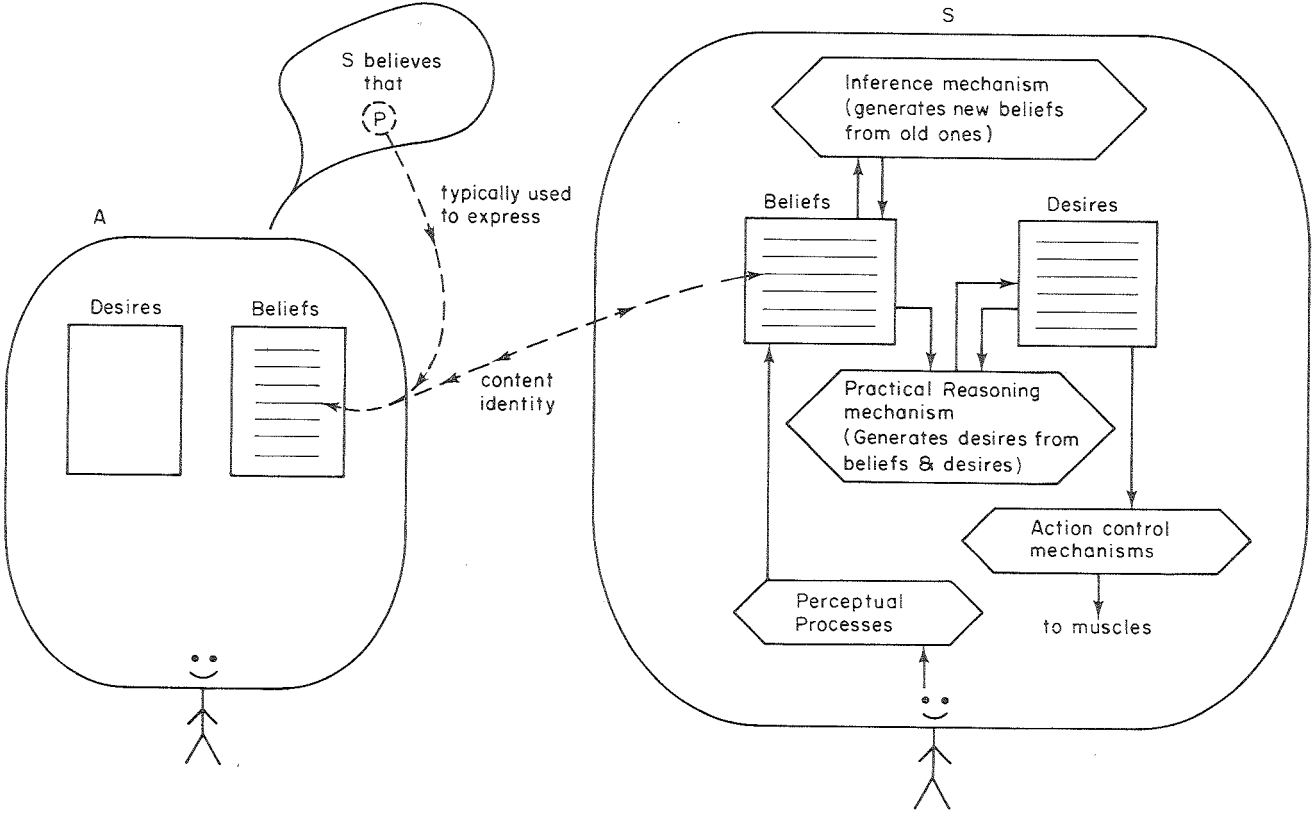
S believes that
P
typically used to express
A
Desires
Beliefs
content identity
S
Inference mechanism (generates new beliefs from old ones)
Beliefs
Desires
Practical Reasoning mechanism (Generates desires from beliefs & desires)
Action control mechanisms
Perceptual Processes
to muscles

Figure 2

relations links the content sentence to an actual or possible belief state of the attributor. Though details are generally left sketchy, the core idea is that the belief in question is the one which the attributor would typically express by uttering the content sentence. Or, as Fodor has put it, it is the belief which plays "a causal/functional role in the production of linguistically regular utterances" of the content sentence.[1] The second relation is identity of content, and it serves to link the attributor's belief to the subject's. So, stripped to essentials, here is what the theory claims: When A says, 'S believes that p', he is saying that S has a mental token of a sentence stored in the way characteristic of beliefs, and this token is content-identical to the one which he (A) expresses by uttering 'p'.

2. *Elaborations and Revisions*

Though the theory sketched in figure 2 is a plausible first approximation, considerable tinkering will be required before we have an account which does justice to our commonsense concept of belief. In this section I will indicate where and why I think modifications are in order.

Conceptual Analysis and Protoscience

As the attentive reader will have noted, I have already smuggled in one modification of the mental sentence view as characterized in chapter 3. Mental sentence theorists generally take themselves to be doing protoscience, not conceptual analysis. Theirs is a theory about the nature of the psychological states spoken of in our folk psychology, not an attempt to analyze the concepts of folk theory. By contrast the account I shall offer aspires to be a descriptive analysis, cleaving as closely as possible to the contours of our commonsense concept. Given the broader purposes of this book, I have no choice but to aim for a descriptive conceptual analysis. If one is convinced that a term in common use denotes some real entity or class of entities, one is free to build a theory about the nature of those entities exploiting whatever evidence or argument may prove useful. But to assume that stance here would be to beg what I take to be a fundamental question. On my view it is up for grabs whether the terms of folk psychology denote anything at all. Perhaps the concepts of folk psychology, like many of those in folk medicine or folk cosmology, will turn out to be true of nothing; perhaps there are no such things as beliefs. If one takes this possibility seriously, then the inevitable strategy is to get as clear as possible on the workings of our folk concept and then assess the prospects of pressing this concept (or some elaboration of it) into service in a serious empirical theory.

Since conceptual analysis has had a bad press of late, I should say

something about how I conceive of the endeavor. There is a long philosophical tradition, stretching back to Socrates, which sought to give an analysis by giving a definition, a set of necessary and sufficient conditions for the application of the concept. If this exercise is to have any point, the terms used in the definition will have to be simpler or more perspicuous than the term or concept being defined. Thus the project of providing necessary and sufficient conditions takes on a reductivist cast. But things have not gone well for this reductivist project. A plausible case could be made that in two and a half millennia of trying, we haven't succeeded in defining *anything* of any interest.[2] So it might be thought that it is time to give up conceptual analysis as a lost cause. But I am inclined to think that this pessimism is warranted only if we follow the central philosophical tradition and insist that a conceptual analysis must give us reductivist necessary and sufficient conditions. Outside philosophy, where the demand for reductive definition has less sway, descriptive conceptual analysis seems to be flourishing. In social anthropology it has long been common practice to describe a set of exotic folk concepts by detailing their interrelations with one another, with community practices and rituals, and with the society's ecology and tradition. Such an account, when well done, can give us deep insight into the conceptual web of exotic folks. But it does not yield reductivist definitions of native concepts. In more recent years a rather analogous attempt at conceptual analysis has been launched by people working in artificial intelligence and cognitive simulation. Their focus has been on domestic concepts rather than exotic ones. Typically their goal has been to describe some interrelated set of commonsense concepts in sufficient detail to assemble a computer simulation which can process information invoking these concepts in ways similar to the ways people process them. Now on my view, philosophical conceptual analysis, when done properly, ought to be continuous with the project of the cognitive simulator. The philosophical analyst can be viewed as giving a rather coarse-grained discursive characterization—a sort of sketchy verbal flow chart—for the more detailed program that the cognitive simulator is trying to write. This is how I would have the project of the current chapter construed.[3]

The move away from reductivist definition as a goal in philosophical analysis is not really all that new. Many writers in the last two decades have embraced the view that our commonsense concepts are folk-theoretic concepts, best analyzed by making explicit the largely tacit structure of folk theory.[4] However, I am inclined to think that something new and interesting is added by viewing philosophical analysis as continuous with cognitive simulation. Although cognitive simulation is hardly past its infancy, it has already established one important fact

beyond serious argument. Folk conceptual systems have again and again shown themselves to be unexpectedly rich, complex, subtle, and interconnected. They are so complex that it is hard to imagine them explicated in detail without exploiting the resources of computer languages and couching one's account as a complex program. If this is right, then philosophers concerned with conceptual analysis have a pair of options. They can bite the bullet and become programmers, or they can aim at informal, often imprecise description which leaves much of the detail unspecified. I think both strategies are respectable. While the informal philosophical strategy must sacrifice specificity, it can often offer greater scope and perspective. This informal strategy is the one I shall be pursuing in the pages to follow.

Mental Sentences and Complex Mental States

In practice the shift from the protoscientific stance of the mental sentence theorist to the strategy of conceptual analysis that I adopt will make surprisingly little difference in my account, save on one point. Mental sentence theorists hold that the objects of belief are sentence tokens, though they do not argue that any such view is entailed by our folk theory. Rather, they maintain, the hypothesis that beliefs are relations to internally inscribed sentence tokens is compatible with folk theory and provides a natural bridge between folk psychology and contemporary research in cognitive science.[5] I have no quarrel with any of this. But since my aim is to describe our folk concept rather than to elaborate on it, I am uncomfortable with the claim that the objects of beliefs are sentence tokens. There is ample reason to suppose that our folk theory takes beliefs to be internal *states* whose role in the psychological economy of the subject fits the pattern sketched in figure 2. Also, it seems clear that folk psychology takes beliefs to be *complex* states whose components, often labeled 'concepts', can recur in distinct beliefs, as well as in desires and in other intentional psychological states. For example, it is intuitively natural to suppose that my concept of water is a component in my belief that water is H_2O, and in my belief that (most of) the stuff in Lake Michigan is water, and in my desire for a glass of water. It is also noteworthy that there is a conspicuous correlation between the concepts we would ordinarily say are involved in a given belief and the words we would use to express the belief. As far as I can see, however, there is nothing *in our folk theory* to motivate the move from viewing beliefs as complex internal states to viewing them as sentence tokens. Nor do I see much reason to think that folk psychology is committed to a systematic language of thought. Thus I will presume, in my account, that a belief is a relation between a subject and a complex internal state of the sort depicted in figure 2.

I should emphasize though that my departure from the mental sentence theory on this point may be more apparent than real. For mental sentence theorists typically leave the notion of an internalized sentence token as little more than a *metaphor*. And it may well turn out that when the metaphor has been unpacked, it claims no more than that beliefs are relations to complex internal states whose components can occur as parts of other beliefs.

The Relation between the Content Sentence and the Attributor's Belief

In figure 2 the link between the content sentence and the attributor's belief has been labeled "expresses," and in my gloss of the sketch I followed Fodor in explaining that a content sentence expresses the attributor's belief if the belief plays a causal role in the production of "linguistically regular" utterances of the sentence. I think the basic insight here is a sound one. In saying what someone else believes, we describe his belief by relating it to one we ourselves might have. And we indicate this potential belief of our own by uttering the sentence we would use to express it. But while the central idea is sound, the details cry out for more careful elaboration. Two problems are especially conspicuous. First, there is reason to suspect that the convenient appeal to "linguistically regular" utterances is actually circular. It is clear enough why some such qualification was called for, since sometimes, as when we aim to deceive, the belief that p plays no role in the causal history of an utterance of p. The belief that p plays the intended causal role only when our assertion of p is sincere. So it looks suspiciously like 'linguistically regular utterance of p' is being used as a synonym for 'sincere assertion of p'. However, it is hard to see how this latter notion could be unpacked without invoking the idea of an utterance *caused by the belief that p*. And now we are standing on our own tail. For the whole idea was to characterize the belief that p in terms of an utterance to which it stands in a special causal relation. In explaining that special relation we find ourselves invoking the very notion it was to be used to explain.

A second problem with our explanation of the "expresses" relation echos a difficulty noted in chapter 2 in our discussion of the theory-theory. To put it starkly, there are just not enough sincere assertions to go around. Adopting an earlier example, it is probably the case that no one has ever sincerely asserted that Buckingham Palace is filled with pickles. And since there are no such sincere assertions, there is no belief which has played a causal role in the production of such assertions. But this bodes ill for belief sentences like 'S believes that Buckingham Palace is filled with pickles'. For absent sincere assertions of the content sentence, our proposed analysis cannot get off the ground.

To patch these problems, I propose to use a pair of tools: first, the idea of a typical causal pattern linking belief states to utterances, and second, a counterfactual. Let me begin by explaining my notion of a typical causal pattern. The basic thought is that there is a typical or characteristic sort of proximate causal history that underlies most of our assertions, and this pattern is the hallmark of straightforward sincere assertion. This is not of course to deny that many of our assertions have causal histories that do not fit the pattern. We sometimes assert what we do not believe, and we sometimes have the most devious of reasons for asserting what we do believe. Yet it remains true that most of our assertions are sincere and straightforward expressions of belief. Now what I am assuming and what I think commonsense psychology assumes is that there is a common causal pattern underlying most cases in which we sincerely and straightforwardly express a belief. Since sincere, straightforward assertion is the preponderant mode of speech, the presumed common causal pattern is the *typical causal pattern* underlying our assertions. No doubt the pattern, like all patterns, admits of some variation from case to case. Yet I think that we ordinarily suppose that the proximate causal histories underlying sincere, straightforward assertions are distinguished by some important common features—that they constitute a single *kind* of psychological process. It is perhaps a symptom of this tenet of folk psychology that we find the idea of constructing a generally reliable lie detector well within the bounds of conceptual possibility. Indeed I am inclined to think that the impressive success of state-of-the-art lie detector technology is some evidence that our folk theory is correct in this assumption.

Granting the notion of a typical causal pattern underlying our assertions, it might be thought that we could identify the belief that p simply as *the* belief which can play a role in a typical causal history leading to an assertion of p. But on reflection, this pretty clearly won't do, since many beliefs in addition to the belief that p may play a role in a typical causal history leading to a sincere assertion of p. Thus for example, on one occasion I may assert 'Ouagadougou is the capital of Upper Volta' because I believe it and because I believe you have just asked me what the capital of Upper Volta is. On another occasion I may make the same assertion because I believe Ouagadougou is the capital of Upper Volta and because I believe that you have just asserted that Abidjan is the capital of Upper Volta. On other occasions still other beliefs may play a role in a typical causal pattern leading to an utterance of 'Ouagadougou is the capital of Upper Volta'. However, in all these cases I think we assume the belief that p plays a special central role in typical causal histories leading to the utterance of p. If pressed to specify this central role, I would proceed as follows. First, concentrate

on those sentences that Quine calls "eternal"—i.e., sentences whose truth values remain fixed despite variation in the identity of the utterer, or the time or place of utterance. Now in the class of eternal English sentences there are some which have been uttered with considerable frequency. Further, there will generally be only one belief which is part of the causal history of *each* typically caused utterance of a frequently uttered eternal sentence. It is the sort of role played by this belief that I am calling the *central* role. However, I am inclined to think that in elaborating this way of picking out the central causal role a belief can play in typical causal histories, we are going well beyond what is embedded in our commonsense psychology. I suspect folk theory gets by with the mere assumption that *there is* a special central role for the belief that p to play in typical causal histories of utterances of p, without ever detailing necessary and sufficient conditions for this role.* To characterize the role and to recognize cases of deviant causal chains linking a belief with an utterance in a nontypical way, my guess would be that folk theory relies on a few *exemplars* or *prototypical* examples to which new cases can be compared.[6]

With the notion of a central role in a typical causal history in hand, we can piece together a better account of the relation between an utterance of 'p' and the belief that p. The basic idea is this: We are to imagine that (contrary to fact) the ascriber is uttering the content sentence with a typical causal history. The belief which he aims at identifying is the one which would play the central role in the typical causal history leading to his imagined utterance of 'p'. So on my account, a rough paraphrase of what A is saying would run as follows: S is in a belief state identical in content to the one which would play the central causal role if I were now to produce an utterance of 'p' with a typical causal history.

Ambiguity and the Logical Form of Belief Sentences

In my account as it has been developed thus far, I have been following the Carnapian strategy of treating belief ascriptions as though they expressed a relation between the believer and the content sentence; though, of course, the relation I have sketched is at best a distant kin

*Grice (1975) notes that folk theories are likely to be "law-allusive," alluding to the existence of laws which they do not themselves specify or have need to specify. My suggestion here is of a piece with Grice's, though "theory-allusive" might be a better term, since what I am suggesting is that our folk theory alludes to the existence of some detailed theory which specifies both the nature of a typical causal pattern and the details of the special causal role. For its purposes, folk theory need not and thus does not specify the theory it alludes to.

of what Carnap had in mind.* There is reason to think we should give up this vestige of Carnapian doctrine, however. For on the Carnapian account it is a sentence *type* to which the believer is related. And if we take this tack, we are left with no explanation of the fact that the ambiguity of belief ascriptions is not nearly so great as the ambiguity of sentence types whose tokens are embedded within them. An example will serve to make the point. Suppose that you and I are walking down the street and we spot a mutual acquaintance, Charlie, pacing back and forth in front of the building across the street. Since you know that Charlie should be in his office at this hour, you ask me, "What's Charlie doing over there?" I reply, "He believes his girl friend has gone to the bank, and he is waiting for her to come out. They had a nasty quarrel this morning, and he wants to patch things up." Now the first thing to note about this example is that the sentence *type*, 'His girl friend has gone to the bank' is ambiguous. The word 'bank' may mean either a financial institution or the edge of a river, and the phrase 'his girl friend' might be used to refer to any of an indefinitely large number of women. Thus there are *many* beliefs which, under suitable circumstances, I might express by sincerely uttering a token of the content sentence. Or, to put the matter the other way around, there are many beliefs which might play the central causal role in a typical causal history leading to an utterance of (a token of) the content sentence. And according to my analysis, as we left it in the previous section, to say that Charlie believes that his girl friend has gone to the bank is to say that Charlie is in a belief state content-identical to the one I would express by earnestly uttering 'His girl friend has gone to the bank'. But, as we have just seen, there are indefinitely many such beliefs, all quite distinct. So if my analysis were correct as it stands, we should expect the belief ascription in my little tale to be comparably ambiguous. Yet of course we find no such thing. It was perfectly clear that the belief I was attributing to Charlie was a belief about *his* girl friend and the financial institution in front of which he was pacing. The ambiguity in belief ascriptions that my account leads us to expect just is not there.

In general when a belief ascription is used to attribute a belief to a person, the referring expressions in the content sentence have the denotation they would have were the content sentence alone to have been issued in the identical setting. This is true when the referring terms are pronominal expressions like 'his girl friend', and also when the referring expressions are names. Thus if I were now to tell you

*Actually, the account as we left it in the last section portrays belief as a three-place relation, with the relata being the believer, the content sentence, and the attributor. For Carnap's view, see chapter 2, section 1.

that former President Ford believes Nixon erased the famous gap in the Watergate tapes, I would be unambiguously ascribing to Mr. Ford a belief about his predecessor, no matter how many people there may be named 'Nixon'. Similarly a token of a *semantically* ambiguous content sentence generally renders a belief attribution in which it is embedded ambiguous only to the degree that a token of the content sentence would be semantically ambiguous if uttered alone in the same setting. Some strategy must be found to take account of these facts in the analysis of belief attributions.

One attractive solution is to borrow an idea due to Davidson.[7] In his account of indirect discourse, Davidson urges that we abandon the Carnapian strategy of analyzing sentences like

(1) Galileo said that the earth moves

as expressing a relation between Galileo and a sentence type, viz., 'The earth moves'. Instead, Davidson proposes that we view each use of a sentence like (1) as expressing a relation (he calls it the relation of "samesaying") between Galileo and the utterance following the word 'that'. More precisely, Davidson urges that the word 'that' be construed as a demonstrative, referring to the speech act which follows.[8] The speech act is, in the nonphilosophical sense, an *act;* it is a sort of skit produced not as an assertion but as a demonstration. Its function is analogous to that of a rude gesture that might accompany my utterance of

> When the teacher's back was turned, the boy whom she had just reprimanded went like that.

The analogy with a gesture is an illuminating one, since the type of gesture performed and its import are very much a function of the surrounding context. If I say,

> After chewing out the recruit, the sergeant went like that.

and accompany my utterance with a sweeping motion of my open palm, the gesture I am attributing to the sergeant will vary dramatically, depending on whether the motion of my hand stops in midair or adjacent to the cheek of a man in my audience. Similarly, if I report,

> The scout paused, examined the tracks carefully, and pointed like that

the information I convey may be critically dependent on the direction in which I point. The context of the act plays an analogous role in belief ascriptions. Questions of ambiguity and reference in the speech act (i.e., the skit or demonstration) following the demonstrative 'that'

will be largely resolved by the setting in which the act takes place, just as they would be were the act to have been performed in earnest.

When we weave this Davidsonian idea into my account of belief ascription, the picture that emerges looks like this. In saying

> Andrea believes that lead floats on mercury.

(under ordinary circumstances) I am doing two rather different things. First, I am claiming that Andrea has a belief. Second, I am performing a little skit, doing a bit of play acting. The words following the demonstrative 'that' in the sentence I utter constitute the script for the play. The two components of my belief ascription are related because the skit is designed to exhibit the typical or characteristic effect of the belief I am attributing to Andrea. A bit more precisely, I am saying that Andrea is in a belief state content-identical to one which would play the central role in the causal history leading up to my play-acting assertion, were that assertion to have been made with a typical causal history. This view of belief ascription is very much in the spirit of an account once advocated by Quine. Consider the following passage:

> In indirect quotation we project ourselves into what, from his remarks and other indications, we imagine the speaker's state of mind to have been, and then we say what, in our language, is natural and relevant for us in the state thus feigned. An indirect quotation can usually expect to rate only as better or worse, more or less faithful, and we cannot even hope for a strict standard of more and less; what is involved is evaluation, relative to special purposes, of an essentially dramatic act. Correspondingly for other propositional attitudes, for all of them can be thought of as involving something like quotation of one's own imagined verbal response to an imagined situation.
>
> Casting our real selves thus in unreal roles, we do not generally know how much reality to hold constant. Quandaries arise. But despite them we find ourselves attributing beliefs, wishes and strivings even to creatures lacking the power of speech, such is our dramatic virtuosity. We project ourselves even into what from his behavior we imagine a mouse's state of mind to have been, dramatize it as a belief, wish, or striving, verbalized as seems relevant and natural to us in the state thus feigned.[9]

Content Identity and Content Similarity

The central feature of content-style mental sentence theories as sketched in figure 2, the feature that distinguishes them from narrow causal theories, is the relation of content-identity that links the belief state of

the believer with the hypothetical or counterfactually characterized belief state of the attributor. This is also the feature about which I have so far had the least to say. In this reticence I have followed the lead of the literature, which is singularly vague and uninformative about the relation of content-identity. I once thought that the way to remedy this problem was to seek a definition of the content-identity relation along familiar analytic lines. This would involve seeking a set of necessary and sufficient conditions for content identity, then testing the proposal against intuitions about cases. If the proposed definition rules that a pair of beliefs are content-identical and intuition agrees or if the definition (along with the rest of our analysis) agrees with intuition in describing a given belief as the belief that p, than we have some evidence that our definition has captured our intuitive concept. If the definition departs from the dictates of intuition, then it must be reworked. What I am describing of course is standard practice, the stock in trade of analytic philosophy. However, I am now convinced that for all its comfortable familiarity, the route I have been describing is not the way to get a handle on the intuitive notion I have been calling 'content-identity'. Not to put too fine a point on the matter, the reason is that *there are no necessary and sufficient conditions for the application of this intuitive concept*, at least not along the lines traditionally conceived.

Several lines of thought converge in leading me to this conclusion. Consider first the fact that many of the intuitions we noted in the previous chapter seemed to lie along a continuum. Recall for example the case of Mrs. T, the woman suffering from gradual, progressive loss of memory. Before the onset of her illness Mrs. T clearly believed that McKinley was assassinated. By the time of the dialogue reported in chapter 4 she clearly did not believe it. But at what point in the course of her illness did her belief stop being content-identical with mine? The question is a puzzling one and admits of no comfortable reply. What we are inclined to say is that her belief gradually became less and less content-identical with mine; it became less and less the belief that McKinley was assassinated. The same apparent impossibility of drawing a natural boundary, even a rough one, can be seen in the case of Paul, who somehow comes to accept a single sentence of a radically new theory. To see this, suppose we imagine the process repeated, with Paul coming to have internal inscriptions of the fundamental sentences of the new theory one at a time. At what point would we be prepared to say that Paul and the future scientists *first* have content-identical beliefs? Well, at no *point*. Their beliefs simply become more and more identical in content. Note the parallel here with children learning some new concept or theory. How much physics must my son know before it is appropriate to say he believes that $E = mc^2$? The

more the better, of course; but there are no natural lines to draw. The case of Dave[1], . . . , Dave[9] makes the point in a different way. Where, as we proceed from Dave[1] to Dave[9] do the beliefs of these imagined subjects stop being content-identical with the beliefs *we* would express using the same sentences? The answer, surely, is that there is no such point. Their beliefs and ours simply become less and less content-identical. What these cases suggest is that the relation I have been calling 'content-identity' is actually a *similarity* relation, one which admits of a gradation of degrees.

The suggestion that content-identity (so called) is actually a similarity relation is reinforced by the context dependence of many intuitions about how beliefs are appropriately described. We saw an example of this in the previous chapter in the case of the color-blind store clerk turned night watchman. Whether or not we were prepared to say without qualification that he believed the object he was looking at was red depended on the focus of our interest. Much the same phenomenon could be illustrated with many of our other examples. Consider, to choose just one case, the child who accepts and repeats a few isolated sentences of a complex scientific theory. If we tell a story in which it is important that the child *assert* $E = mc^2$ or *deny* $E = mc^4$ it can seem quite natural to say that the child does indeed believe (or know) that $E = mc^2$. (Here is the outline of such a story: The quiz show producers are pondering what questions to set for little Alice. Their plan is to allow her to win a small sum, but to be sure that she does not win the grand prize. "For the grand prize, how about asking her what E equals in Einstein's famous equation," proposes one producer. "No," protests the other, who has interviewed Alice at length, "she knows that $E = mc^2$".) If, by contrast, we tell a story in which it is important that Alice be able to use the belief to solve some contextually important problem, our inclination is to deny that she believes $E = mc^2$. This shift in judgment, as a function of the focus of our interest in the context where the judgment is called for, is quite characteristic of similarity judgments. If we are discussing climate and terrain and I ask, "Is the USSR more similar to Canada or to Cuba?" Canada is the clear intuitive choice. But if our discussion has been about political systems and I raise the same question, the natural answer is Cuba.

Another consideration encouraging me to believe that what I have been calling 'content-identity' is actually a similarity relation is that analogous similarity relations have recently come to play a large role in the empirical study of the mental representation of concepts and the use of concepts in categorization. The work I have in mind has developed in reaction to the traditional view that concepts are represented by definitions—lists of properties or features which are indi-

vidually necessary and jointly sufficient for the application of a concept. The theories that have been proposed in opposition to this tradition are still very much in flux.[10] Typically they claim that a concept (the concept of a bird, say, or of a robin) is represented by a set of features few if any of which are necessary. In deciding whether a given object falls under a concept, the theories claim that a subject performs a similarity match, determining the extent to which the features in his representation of the object coincide with the features in his representation of the concept. A variation on this theme suggests that concepts are stored not as a set of features but, rather, as one or more stored *prototypes* or *exemplars*. On this version, we determine the applicability of a concept to an object by matching the features in our representation of the object with those in our prototypical representation(s). The effect of context may shift our focus from one prototype to another or affect the weight that is given, in the similarity measure, to one sort of feature or another. These latter complications are only beginning to be studied, however.

It turns out that feature matching or prototype similarity theories explain an impressive range of data that are problematic for the traditional view that concepts are represented by mental definitions. Here are a few striking illustrations. First, people find it a fairly natural task to rate various objects or subcategories falling under a concept with respect to how typical or representative of the concept each is.[11] For example, people rate robins and sparrows as more typical birds than eagles, and eagles as more typical than chickens. Moreover, the typicality rankings obtained correlate with a surprising range of further phenomena. They predict both the speed and the accuracy with which subjects will respond when asked to determine whether an item or a subcategory falls under a category. People are faster to judge that a robin is a bird than to judge that a chicken is a bird. Typicality rankings also predict which concepts will be learned first by children and which subcategories will be mentioned first when subjects are asked to list items falling under a category.[12] Perhaps most suggestive to philosophers is the finding that all of the above mentioned typicality-correlated phenomena also correlate with what has been called the "family resemblance measure." This latter measure is determined by asking subjects to list features characteristic of members of various subcategories of a given category. For example, if the category is furniture, subjects are asked to list features of a table, a chair, a rug, a vase, etc. The more mentioned features a subcategory has in common with other subcategories, the higher is its family resemblance measure.[13]

Now none of the results I have been describing bear directly on the analysis of belief ascription. The reason I find them relevant is that

they suggest that context-dependent judgments of similarity may play quite a fundamental role in the mental mechanism underlying our intuitions on how things are to be categorized or described.* And, as we have already seen, there are ample hints suggesting that this is afoot in the case of our judgments about the appropriateness of belief ascriptions. So let me mold these various hints into an explicit hypothesis. What I am proposing is that in figure 2 we replace the content-identity relation with a notion of context-dependent similarity. If this is right, then we will have the following rough paraphrase of what we are saying when we say 'S believes that p':

> p.
> S is in a belief state similar to the one which would play the typical causal role if my utterance of that had had a typical causal history.

The 'that', of course, is a demonstrative, referring to the play-acting utterance of 'p' that preceded. I offer this account as an hypothesis to be tested against our intuitions. But before seeing how well it does, some elaboration is needed on just how the notion of similarity is supposed to work.

Basically I conceive of the similarity measure as a feature-matching measure along the lines sketched above and characterized in greater detail by Tversky.[14] Unlike Tversky's account of similarity, however, the one needed in my account will have to discriminate among features in a context-dependent way, sometimes giving heavier weight to features of one sort, sometimes to another. I have no detailed proposal to make on how this sort of context sensitivity is best built into a feature-matching similarity tester. It is one of those (many and important) problems of detail that I will foist off on the cognitive simulator. Without actually building a model or writing a program, we have no choice but to sketch the forest while ignoring the trees.

Still, even at this relatively coarse-grained level of description, something must be said on how beliefs are represented—what the features are in virtue of which similarity among beliefs is to be assessed. I think that the features which are most salient in our mental characterization of beliefs can be grouped under three headings. The first and in a sense the most central is *functional* or *causal-pattern* similarity. A pair of belief

*An important caveat: Though I think the prototype-cum-similarity-match theory is an important advance in the study of concepts and categorization, I am inclined to think it is much too *simple* a story. Rather than a prototype or a cluster of features, I suspect our concepts are represented in structures more akin to Minsky's frames (Minsky 1975, Winograd 1981). Though I adopt the prototype theory as a metaphor for my view, the account I am developing is comfortably compatible with a more complex framelike picture of conceptual representation.

states count as similar along this dimension if they have similar patterns of potential causal interaction with (actual or possible) stimuli, with other (actual or possible) mental states, and with (actual or possible) behavior. A strong causal-pattern similarity is the single standard for sameness of belief proposed by the narrow causal version of the mental sentence theory. In addition to the global causal-pattern similarity stressed by causal accounts, however, there are various dimensions along which a pair of belief states (or other mental states) can be *partially* causal-pattern similar. For example, a pair of belief states may interact similarly with other beliefs in inference but may not have terribly similar links with stimuli. These beliefs would count as highly similar when the context primes for inferential connections but as rather dissimilar when the context focuses interest on the connections between belief and perception. Causal-pattern similarity is the basic sort of similarity since it is presumed to some degree by the other features used to gauge similarity of belief. What we do, in effect, is to locate a pair of belief states which are (or are assumed to be) passingly similar in causal pattern, then, when relevant, attend to other features in our representation of belief.

The second sort of feature which we use to assess similarity of beliefs underlies what might be labeled *ideological similarity*. The ideological similarity of a pair of beliefs is a measure of the extent to which the beliefs are embedded in similar networks of belief. In effect, ideological similarity measures the similarity of the doxastic neighborhood in which a given pair of belief states find themselves. As in the case of causal-pattern similarity, partial ideological similarity is often much more important than global ideological similarity. Since belief states are compound entities, ideological similarity can be assessed separately for the several concepts that compose a belief. And context can determine which concepts are salient in the situation at hand. For example, under some circumstances, if Boris and Marie both say, 'Abstract art is bourgeois', we may count them as having similar beliefs if their other beliefs invoking the bourgeois-concept are similar, even though they have notably different beliefs invoking their abstract-art concept. But if the difference in their conception of abstract art looms large in the context, our judgment will be reversed, and they will not count has having similar beliefs.

The third sort of feature used in assessing belief state similarity leads to what I will call *reference similarity*. The core notion of reference similarity is straightforward enough. A pair of beliefs count as reference similar if the terms the subjects use to express the beliefs are identical in reference. Complications come quickly, however, since there is considerable dispute over just how the reference of a term is determined

for a given speaker. The dominant view a decade or two ago was that reference is determined by the set of statements involving the term that the speaker takes to be true. If this were the whole story, then reference similarity would reduce to ideological similarity. But recent work has made it pretty clear that other factors are involved in determining the reference of a term. One prime candidate is the causal history of the use of the term, a causal chain stretching back through the user's concept, through the concept of the person from whom he acquired the term, and so on to the person or stuff denoted.[15] A second candidate, persuasively defended by Burge, is the use of the term in the speaker's linguistic community.[16] Neither the causal nor the linguistic community story is free from problems, and neither is a paradigm of clarity.[17] It would be out of place in this book to spend substantial time improving matters (though I would do it anyhow, if I could). Since there are different and potentially competing factors contributing to our judgment about the reference of a term, my theory would predict that context may single out one or another of these for special emphasis in the assessment of belief state similarity. And in fact I think that cases illustrating this sort of context sensitivity can be constructed.[18] It might be thought that since reference similarity is defined by appeal to the expression of a belief in language, this component of belief state similarity would play no role in our judgments about the beliefs of animals or prelinguistic children. By and large I think this is true. However, the causal history strand in the determination of reference has an analogue in the case of beastly concepts, since a dog's concept, if not its use of a term, may have a causal history linking it to some specific object or stuff. And, as we shall see below, these quasi-reference-fixing causal links will sometimes influence our intuitions about how a dog's belief is to be described.

On the theory I am proposing, the three sorts of features sketched above are the principal determinants of belief similarity and thus the principal determinants of our intuitions on sameness of belief and on the appropriateness of a content sentence in characterizing a belief. However, I make no claim about the exhaustiveness of these three sorts of features. It is no doubt the case that other sorts of features can play a role in our characterization of a belief state. And given a suitable context, one or another of these further features may play a dominant role in determining our intuition about the appropriateness of a description of the belief in question. For example, in a fascinating paper John Haugeland argues that computer simulations, current ones at least, do not really understand a text about which they can answer questions, because they lack an existential sense of themselves, a concern about who they are "as some sort of enduring whole."[19] I think that Haugeland

succeeds in at least tempting our intuitions to agree that the computer does not really believe or understand. His strategy is to note an undeniable difference between the computer's "cognitive state" and our own, then to focus on contexts involving embarrassment, guilt, and threats to one's self-image where the difference is of central importance.

Before we turn to other matters, I want to consider two closely connected objections to the idea of embedding a notion of similarity in the analysis of belief ascription. The first objection protests that introducing similarity seriously distorts the picture of our intuitive notion, since similarity is a matter of degree, while believing that p is not something which admits of degrees. There is an initial plausibility to this charge. But on a closer look, I think its plausibility can be traced to a misleading grammatical quirk. Certain properties or attributes are expressed in English by adjectives admitting of comparative endings. 'Tall', 'fat', 'bald', and 'blue' are examples, and the properties they express are the ones we think of as admitting of degrees. Where a comparative form is not readily available, there is a tendency to think that the notion does not admit of degrees. But the example of such concepts as *cup*, *couch*, or *eggbeater* should convince us that this tendency is to be resisted. In each of these cases it is easy to imagine constructing a sequence of objects whose early members clearly fall under the concept and whose later members become increasingly less cuplike, couchlike, or eggbeaterlike. As we saw in the case of Mrs. T, analogous sequences can be constructed with clear examples of the belief that p at one end and increasingly less belief-that-p-like cases further along. In none of these cases can a natural boundary be drawn separating the instances which fall under the concept from those which do not. Despite appearances, then, both the *couch* concept and the *belief that p* concept do admit of degrees.

The second objection protests that the use of similarity in our analysis entails that belief ascriptions are always rather vague. But, the objection continues, this is simply not true. Many belief ascriptions, particularly *de dicto* belief ascriptions, say something exceptionally precise about the believer, so precise, in fact, that even the subtlest change in the content sentence may turn a true belief ascription into a false one.[20] My reply here is that the critic is simply mistaken in thinking that importing a notion of similarity entails that belief ascriptions are always vague. Perhaps the best way to make the point is to look at the use of similarity in another setting. If you and I are wandering through the National Gallery and I comment casually that the Valázquez portrait of Pope Innocent is very similar to Rembrandt's Portrait of a Polish Nobleman, my remark would no doubt be intended as a rather vague one which would remain true even were the Rembrandt notably dif-

ferent from the way it is. Take away the Polish Nobleman's earrings, for example, and the two paintings still count as similar by the contextually indicated standards. But now suppose that we are in the laboratory of the National Gallery watching an expert examine the Rembrandt and an excellent forgery. Looking up from her microscope she says, "It's remarkable. They are very similar." Note that here a vastly different standard is contextually indicated. The expert would not have said what she did if the forger had forgotten to paint in the earrings! The point I want to extract from these two cases is what while the sentence *type* 'A is similar to B' may be quite vague, individual tokens uttered in context can be very precise indeed. It is the context which makes clear the standard and the aspects of similarity that are appropriate. Now it is my contention that the alleged precision of so-called *de dicto* belief ascriptions is simply a consequent of focusing on a certain sort of context. What is needed is a case involving a compatriot, someone whose beliefs and language are very similar to our own. We then make it a matter of special interest exactly what the believer would be prepared to say or assent to, making sure that slight differences in words make a big difference in that context. In this setting a minor change in the content sentence can radically change our intuition about the propriety of a belief ascription. But when the believer is less like us—a child, perhaps, or a person from an exotic culture—or when the exact words he would utter or accept are of small importance, minor differences in the content sentence are of less moment.

3. *Testing the Theory Against the Facts*

In this section and the two that follow I want to indicate how the theory I have been sketching handles the facts of intuition. Since my theory is billed as a descriptive conceptual analysis, these facts constitute the principal data which the theory must explain. I will divide my effort into several parts. In the current section I will go through the cases described in the previous chapter, showing how the intutions these cases evoke are explained by the theory. In the following two sections I will look at two particularly vexing cases, the apparently absurd beliefs of exotic folk and the beliefs of animals, and I will argue that in these cases too the theory does a plausible job of explaining our intuitions.

I begin with those cases discussed toward the end of chapter 4, under the heading Irrelevant Causal Differences. The point that was being made in that section was that sometimes belief states exhibiting substantial differences in their potential causal patterns will nonetheless be ascribed with the same content sentence and intuited to be the same

belief. My strategy for generating the cases was to start with a normal or near normal subject and make progressively larger changes in the potential causal patterns leading to or from his belief states. In the first sequence of cases, the changes were perceptual, leaving the inferential parts of the causal network unchanged; the second sequence worked just the other way around. What my theory predicts in these cases is as follows. First, if we keep the context fixed, then as the difference between the causal pattern of a subject's belief and the causal pattern of our own gets greater, our willingness to ascribe both beliefs with the same content sentence should diminish. This is particularly clear in the Dave[1], . . . , Dave[9] cases. As inference gets progressively less normal we get progressively less willing to use the content sentences we would have used before the subject's inferential breakdown. The theory also predicts that in cases where it is not of great contextual importance, significant differences in causal pattern will be ignored. This explains why we are generally willing to characterize the beliefs of color-blind or totally blind people with the same content sentences we use to characterize the beliefs of normal subjects. Finally the theory predicts that in contexts which prime for one or another feature of belief, differences in that feature will be of particular importance in our assessment of similarity and thus in our intuitions about the appropriateness of a content sentence. This, I would urge, is what is going on in the case of Peter, the store clerk turned night watchman. In the store clerk half of our tale, the emphasis is on his anomalous visual perception, and it sounds natural to say that he does not really believe that the ball is red. In the night watchman half of the story, the emphasis is on practical reasoning. How could he have behaved so stupidly, given that he believed that the red lever was the one which would avoid an explosion and that he wanted to avoid an explosion? Here, since the belief state in question has *inferential* interactions which are much the same as the ones that would be exhibited by our own belief that the lever is red, we are prepared to use the same content sentence for both.

I turn next to the four cases recounted in the section on Holism. There the strategy was to vary the surrounding beliefs, the doxastic neighborhood, in which a belief is embedded, thus altering its ideological similarity to our beliefs or to the beliefs of some other person to whom the subject was being compared. The theory predicts that, keeping the context constant, ideological similarity should be positively correlated with sameness-of-belief intuitions and with intuitive willingness to describe the subject's belief and our own with the same content sentence. The case of Mrs. T fits the prediction perfectly; as she loses more and more of her memory, we become increasingly reluctant to say that her

doxastically isolated belief counts as the belief that McKinley was assassinated. Though none of the cases in the Holism section attempted to manipulate context, this is easily done in the case of Mrs. T. Let us imagine that at about the time of the dialogue reported in that section, Mrs. T was the subject in a somewhat gruesome Milgram-style experiment. She is led into a room where she can see someone she cares for, say Mr. T, sitting in a booth wired to a chair. She is told that Mr. T is receiving painful electric shocks, and Mr. T goes along by screaming convincingly. We then tell Mrs. T that she can stop the shocks by pushing one of two buttons. If McKinley was assassinated, she must push the red button to stop the shocks; if McKinley was not assassinated, she must push the green button. On hearing this Mrs. T rushes to the red button. Why? Here, I think, it is all but impossible to resist the inclination to say that she pushes the red button because she believes that McKinley was assassinated. This of course is just what would be predicted by my theory, since the context emphasizes the importance of the (ex hypothesis normal) inferential connections of Mrs. T's belief and deemphasizes the importance of her belief's doxastic isolation.

The case of Paul, who mysteriously comes to have a doxastically isolated belief, is comfortably explained by my theory, as is the case of the child who accepts '$E = mc^2$' though largely innocent of physics. In both instances, there is little ideological similarity between the subject's belief and the belief of those we are comparing him to—the future scientists in Paul's case; ourselves in the case of the child. So we count Paul's belief as different from the scientists', and the child's belief as different from ours. There is one peculiarity in the future scientist case that merits special attention. We noted that in talking about the future scientists' beliefs it seems natural to use, as a content sentence, the sentence they would use in expressing their belief, even though that content sentence would be quite incomprehensible to us. Why should this be? The explanation I would offer turns on the counterfactual locution built into our analysis of belief ascription. As I have told the story, what we assert when we say 'S believes that p' is, roughly, that S is in a state similar to the one which would have caused me to utter 'p' just now, *had I uttered it with a typical causal history*. But just how are we to make sense of the italicized counterfactual clause? What possible world do we imagine when we imagine ourselves asserting the scientists' sentence in earnest? The one that comes first to mind, I think, is the world in which we have learned and come to believe their theory. And of course in *that* world the scientists' belief state would be ideologically similar to mine.

The final case in the Holism section concerned the future scientists' belief that the Hindenburg exploded because it was filled with hydrogen.

Is it properly so characterized? Does it count as the same belief as the one we would express with the same words? Here the theory predicts intermediate intuitions, and these, I think, are just what we have. The theory also predicts that by emphasizing or deemphasizing the importance of the ideological gap we should be able to urge our intuitions in one direction or another. Cases illustrating this prediction are relatively easy to construct, though I will not pause to elaborate one here.

Let us now consider the set of cases collected under the heading Reference. The first of these concerned beliefs whose expression invoked the proper name 'Ike'. In the mouth of Tom, a contemporary American, 'Ike' referred to Dwight Eisenhower, while in the mouth of Dick, a Victorian Englishman, 'Ike' referred to Reginald Angell-James. This difference in reference can, I think, readily be explained by the causal account of the reference of proper names. The causal history of Tom's Ike-concept traces back to Eisenhower, while the causal history of Dick's Ike-concept traces back to Angell-James. Apart from this difference, the two beliefs were portrayed to be very similar indeed. The theory predicts that when the reference of the names is of particular salience, as it is in this case, our intuitions will count the beliefs as different despite their similarities in other respects. And this, it appears, is just what most people are inclined to say. We also noted that it would be entirely natural under most circumstances to say that Tom believes *Eisenhower* was a politician, while Dick believes *Angell-James* was a politician. This choice of content sentences has a ready explanation on the theory I am defending, since Tom's belief is very similar (referentially and in other respects) to the one I would express by saying 'Eisenhower was a politician', and Dick's is very similar to the one I would express by saying 'Angell-James was a politician'. There are differences between my belief and Tom's, however. Tom would neither say nor assent to the sentence 'Eisenhower was a politician', since he has never heard the name 'Eisenhower'. And if this difference between the causal ties linking belief and behavior were of contextual importance, as it might be, for example, in a story recounting Tom's failure to give the right answer on an exam or a quiz show, we would find it much odder to characterize Tom's belief with this content sentence. Again this is just what my account would lead us to expect.

The second reference case turned on the difference between the American and British uses of 'chicory' and 'endive'. It is not clear to me whether the difference in reference in this case is due to a difference in causal history (à la Putnam and Kripke) or to a difference in accepted patterns of usage in the linguistic communities (à la Burge). But no matter. My theory predicts that in contexts where reference is of some importance, we will take the Englishman's belief to be different from

the American's, though they both express their belief by saying, 'Chicory is bitter'. And, as we saw, this prediction is correct. The theory also explains why, in many circumstances, we would describe the Englishman's belief as the belief that *endive* is bitter. Though when the context focuses attention on the exact words the Englishman would use or accept (think of the quiz show again) the theory correctly predicts that this content sentence will seem intuitively inappropriate.

The reference cases we have looked at thus far illustrate the way in which reference similarity and causal-pattern similarity can pull in opposite directions, with context determining which will make the greater contribution to our intuitive judgments. It might be wondered whether, as the theory leads us to expect, there are analogous cases which pit components of reference similarity against *ideological* similarity. I think that with a bit of imagination such cases are readily constructed. Here is one. It involves the dastardly deeds of one Boris A. Nogoodnik, an agent of the KGB. In the course of his dirty work, Boris decides he must do in a minor American spy named Grimes. He devises a plot which will not only get rid of Grimes, but will also put the blame for the crime on a certain troublesome, distantly related Russian émigré named Boris B. Nogoodnik. To pull it off, KGB man Boris plans to leave a trail of incriminating clues, all pointing to émigré Boris. On the day of the crime there is a slight slip-up however. Grimes recognizes his attacker as his KGB contact Boris Nogoodnik. What is worse, Grimes does not die immediately from his wounds. He is still alive, though just barely, when FBI agent Edgar finds him. Grimes whispers "Boris Nogoodnik did it," and then promptly expires. Now Grimes, let us suppose, has never heard of Boris B. Nogoodnik, the émigré. He was referring to Boris A. whom he had recognized as his attacker. FBI man Edgar, for his part, has never heard of either Nogoodnik. But he quickly picks up the trail of clues, all pointing to poor Boris B. From the clues he comes to believe that the killer lives on Front Street, frequents the *Mauve Gloves Lounge*, drives a Fiat, smokes French cigarettes, etc. And, of course, Edgar believes that the killer's name is 'Boris Nogoodnik'. All of this is true of Boris B. In short order Edgar arranges a stakeout at the townhouse on Front Street. Question: Does Edgar believe that Boris A. Nogoodnik killed Grimes or that Boris B. Nogoodnik killed Grimes? The clear intuitive answer is that Edgar believes Boris B. did it, despite the fact that the causal history of Edgar's utterances of 'Boris Nogoodnik' trace through the expiring Grimes to Boris A., not to Boris B. What has happened, I would contend, is that ideological similarity has overwhelmed similarity of causal history in determining our intuition. It is interesting to note, however, that although we have a case of ideological similarity winning out over a *component* of reference

similarity, it is not a case of ideological similarity winning out over reference similarity *simpliciter*. For in this case it seems most natural to say that when Edgar asserts, "Boris Nogoodnik killed Grimes," he is *referring* to Boris B. The case poses no problem for my account of belief ascription, though it does raise serious questions about the general adequacy of a causal theory of proper names.

Let us turn, finally, to those reference cases involving indexicals. Part of the story, but only part, is quite straightforward. We noted that if two political candidates both say, 'I will be the next president of the United States,' we have conflicting intuitions on whether the two men have the same belief. My theory explains this on lines analogous to those used in the *Ike* case. The candidates' beliefs are similar in causal pattern but different in reference, thus the intuitive conflict. What my theory, as so far developed, cannot explain is the fact that we ascribe both of these beliefs by saying, 'Jones (or Smith) believes that *he* will be the next president'. Similar problems arise with our saying, in the case of the man attacked by the bear, 'He believes that *he* is being attacked by a bear? Let us focus on the latter case. Presumably what we mean here is that the victim has a belief which he might express by saying, 'I am being attacked by a bear'. This sentence is *reference* similar to the content sentence I used in ascribing the belief (viz., 'He is being attacked by a bear'); so all is well there. But the belief I would express using the content sentence in earnest is radically dissimilar in causal pattern from the belief we wish to ascribe. The belief I would express using the content sentence leads me to run for help, while the belief I am attributing to the victim leads him to roll up like a ball. What is more, this case cannot be treated as one in which partial similarity and partial dissimilarity lead to potentially conflicting intuitions. It would simply be wrongheaded for me to say the victim believes that *I* am being attacked by a bear, in virtue of the causal-pattern similarity between his belief and the one which I would express using this content sentence. He believes no such thing.

I think the lesson to be learned from this case is that an additional wrinkle must be added to our analysis of belief ascription. When belief ascriptions use indexical expressions like 'he' (or 'himself'), 'her' (or 'herself'), 'you' (or 'yourself'), 'then', 'there', and some others, referring to the believer, the moment of belief, or the believer's location, there seems to be a linguistic convention mandating *normal* reference similarity and what might be called *transposed-causal-pattern similarity*. By this I mean simply that the belief which is causal-pattern similar to the believer's is not the one I would express using the content sentence in the belief ascription but, rather, the one I would express using the content sentence with the indexical pronoun replaced by one referring

to myself, the current time, or my present location (generally 'I', 'now', or 'here'). So, for example, if I say the victim of the bear attack believed that he would die then and there, I am claiming that he is in a belief state which he would express using terms similar in reference to the terms I used (presumably he would say, 'I will die here and now' if he speaks English) though it is similar in causal pattern to a belief that I would express using a transposed content sentence (viz., 'I will die here and now'). An added indication that these cases involving indexicals are rather special ones is the fact that, to all appearances, ideological similarity drops out of the picture almost entirely in these cases. The bear attack victim's concept of himself, or, more accurately, the set of beliefs he has about himself that he would express using a self-referential pronoun, need not have much resemblance to either the set of beliefs I would express about him using 'he' or to the set of beliefs I would express about myself using 'I'.

4. *Absurd Beliefs*

From time to time we hear reports of people who assert, with apparent sincerity, some claim which strikes us as so patently false, so hopelessly beyond belief, so absurd, that it boggles the mind how anyone could possibly believe it. The beliefs that these people are expressing are what I will call *absurd beliefs*. A rich source of examples of such beliefs can be found in the anthropological literature describing belief systems of so-called primitive peoples. As a paradigm case, consider a belief which, according to Evans-Prichard is widely shared by the Nuer people: the belief that a sacrificial cucumber used in certain rituals is an ox![21] Other examples can be found closer to home among the avowed beliefs of people who accept political, religious, metaphysical, or "scientific" doctrines wildly different from our own.

In general the cases that concern us will be found among the beliefs of people who are, in one or more domains, very ideologically dissimilar from us. But it is important to realize that not all beliefs embedded in doxastically exotic surroundings will strike us as absurd. We saw two rather different examples of this in our discussion of the imagined future scientists who espouse a theory quite unlike anything known to us. Some of their beliefs, like the belief that the Hindenburg burned because it was filled with hydrogen, are neither absurd nor even unfamiliar. Though when context focuses attention on the links between this belief and the perplexing newfangled theory, our intuitions on whether the scientists' belief is the same as ours may waiver. Other beliefs held by our imagined scientists, the ones they express in unfamiliar theoretical vocabulary, will not strike us as absurd, though we

may find them incomprehensible. In these cases, as we saw, there is some intuitive appeal to the strategy of borrowing the scientists' own sentence to use as a content sentence in reporting their belief. In contrast with these cases, absurd beliefs are ascribed using perfectly familiar vocabulary. But the content sentence is one whose falsehood, we think, should be unmistakably obvious to anyone.

There is a certain relativity in my characterization of absurd belief, since what one person takes to be screamingly false, another may take to be debatable, or even true. Consider, for example, the situation of the average freshman philosophy student when first told of Bishop Berkeley's curious belief that chairs and tables are made up of ideas. To the freshman, Berkeley's doctrine seems not merely false, but patently so. Indeed the reaction of many freshmen on first hearing of Berkeley's belief is to insist (or at least suspect) that "Berkeley must have been some kind of nut." This is an intriguing response, and it is one which I think a theory of belief ascription should try to explain. Moreover, it is not all that different from the reaction of rather more sophisticated people on first hearing reports about the Nuer belief. In this case people are unlikely to suspect madness, since the belief is purportedly shared by an entire culture. But many are tempted to conclude that if reports about the Nuer belief are correct, then the tribe must, as a group, exhibit a prerational mentality. Their minds must work very differently from ours. Oddly, though, these suspicions of madness or prerationality often decline when we learn more about the system of beliefs in which the prima facie belief is embedded. After a good introductory philosophy course which sets out Berkeley's system, his arguments, and the intellectual background of his thought, most students no longer suspect that Berkeley was mad. Indeed some come to believe he was right. Analogously, on reading a careful anthropological account of the Nuer, some people find themselves significantly less inclined to suspect that the Nuer mind is vastly different from our own. Rather, they come to see the Nuer rituals and beliefs, including the one about the cucumber and the ox, as a sensible systematic attempt to deal with the natural and social environment the Nuer confront.[22] What I have been reporting about reactions and change of reactions, in the face of absurd belief, is no doubt familiar enough. But it raises a perplexing question for a theory about our commonsense notion of belief. What is there about our folk notion that leads us to react as we do?

The explanation I would offer starts with the counterfactual embedded in the analysis of belief ascription. If you tell me S believes that p, I am to imagine that you have asserted 'p' in earnest, since you are claiming that S's belief state is similar to the one that would be centrally responsible for your utterance, had the utterance had a typical causal

history. But how am I to imagine you asserting 'p' in earnest if 'p' is absurd? In what possible world might you (or I, for that matter) seriously say that a cucumber is an ox? How is the freshman to imagine a possible world in which his instructor earnestly asserts that a chair is made of ideas? In all likelihood, what will come first to mind is a world in which the speaker has taken leave of his senses—"gone bonkers" as one of my students put it. For the content sentences in question are not merely false, they are patently incompatible with many further beliefs which we hold and which we know are shared by the person making the belief ascription. So a possible world in which he succeeds in inserting such a belief into his store of beliefs must be one in which his cognitive mechanisms are no longer functioning properly; the mental subsystem responsible for filtering out blatant contradictions must be out of order. Now the explanation I would offer for the initial reaction to reports (and avowals) of absurd belief, the reaction which insists that the believer is prerational or mad, is simply this: In attempting to imagine a possible world in which the speaker might make his assertion in earnest, the world that comes most readily to mind is one in which the speaker's mind no longer filters out obvious contradictions as our mind does. In short, the most easily imagined world in which the speaker asserts the content sentence in earnest is a world in which he is no longer functioning rationally.

Some of the time, however, some of us can conjure a more flattering possibility. The problem with absurd beliefs is that they so obviously contradict so many other beliefs. But if we can imagine replacing these other beliefs with a new set, a set which is more or less consistent and not glaringly incompatible with the belief being ascribed, we can then imagine the speaker sincerely asserting the content sentence without suspecting him of any deep cognitive failing. If we can also imagine how such an alternate set of beliefs might be acquired and/or maintained in the believer's ecological setting, we will have imagined a possible world in which a speaker with a mind quite as rational as our own might sincerely utter a sentence which, on first hearing, strikes us as absurd. We now have an explanation for the fact that as we learn more about a person's system of beliefs, his practices, his traditions, and his environment, we sometimes find that what had seemed an absurd belief no longer strikes us as indicative of prerationality or madness. Rather than imagining the speaker's mind to be defective in function, we succeed in the more challenging task of imagining a radically different doxastic network.[23]

Unfortunately however, the story about absurd beliefs does not stop here. For there is a sense in which either of the two possible worlds we might conjure in evaluating the report of an absurd belief is un-

congenial to the whole business of describing beliefs by appeal to content sentences. Consider first the case in which we imagine that the speaker has lost the capacity to detect and eliminate contradictions. What we are imagining is a person with serious inferential shortcomings, fully on a par with some of the more serious cases in our sequence from Dave[1] to Dave[9]. Given the marked causal-pattern dissimilarity, we are reluctant to characterize this person's nonabsurd beliefs with the content sentences we would use were he not inferentially impaired. When a person is *that* different from us, we are inclined to think that there is just *no saying* what he believes. But this global reluctance to ascribe content to the belief states of an inferentially peculiar person comes full circle and undermines our willingness to ascribe the absurd belief to him. So there is a sort of tension built into these cases. On the one hand we are inclined to say that a suitably mad person might perfectly well come to believe that a cucumber is an ox or some other patently absurd claim. On the other hand we are inclined to say that if a person's inferential capacities are that far gone, then no content sentences in our language will adequately characterize his belief states. There is a whiff of contradiction here. It seems that when the processes underlying our intuitive judgments are confronted with cases like this one, they urge two intuitions which are simply incompatible. What is more, this is not, as far as I can see, an inconsistency that can be written off by appeal to different contexts of judgment. But perhaps it is not all that surprising that our folk psychology gets into trouble in these cases. For folk psychology presumably evolved as an aid to our workaday dealings with one another. The fact that it renders contradictory judgments about exotic cases does not diminish the practical utility of folk psychology, since *real* cases like this are few and far between.[24] The moral, I suppose, is that where nature doesn't itch, folk theory doesn't scratch.

It would be a mistake, however, to conclude that cases like the ones we have been considering are of interest only as devices for probing the contours of our folk concepts. The clinical literature is full of reports of people who say absurd things and whose reasoning capacities are clearly impaired, though the problems are never quite so clear-cut as the ones we have been imagining.[25] In reading these cases one becomes acutely sensitive to the fact that the descriptive apparatus provided by common sense just is not up to characterizing the patient's cognitive states. When a man insists, apparently in all sincerity, that he is Jesus Christ and that he is a heap of dung,[26] there is no comfortable characterization of the content of his beliefs.

Let us consider now the second possibility that we might imagine in evaluating the report of an absurd belief, viz., the possibility that

the speaker's doxastic network is radically different from our own. The situation here is parallel to the case of inferential breakdown. In order to accommodate the absurd belief without contradiction, we must imagine a doxastic network so different from our own that we undermine our willingness to use familiar content sentences to characterize any of the speaker's beliefs in the doxastically exotic domain. Once again a contradiction looms. Having projected ourselves into an exotic belief network, we are inclined to say both that the subject believes a cucumber is an ox and that English content sentences are not up to characterizing his beliefs at all. This difficulty plagues anthropology much as the problem of inferential failure plagues clinical psychology. One way in which anthropologists have attempted to deal with the problem is to invoke native terminology in their descriptions of native beliefs, in effect labeling the native beliefs as they are labeled by the natives themselves. This will work, of course, only when supplemented by an elaborate gloss explaining how these beliefs, characterized by native content sentences, relate to one another, how they integrate with native practices, and how they enable the natives to deal with their environment as they understand it. When native belief systems and "forms of life" differ radically from our own, there is no shorter way to characterize their cognitive world. The descriptive apparatus of folk psychology is not designed to deal with the beliefs of exotic folk.

I think these reflections on the inadequacy of folk locutions in characterizing exotic inferential patterns and exotic doxastic networks throw considerable light on a pair of arguments that have been widely discussed. In *Word and Object*[27] Quine argues against the "doctrine of 'prelogical mentality'." The thrust of his argument is that if a person makes absurd assertions we should suspect our translation or interpretation of his words, rather than impute to him an absurd belief. "The maxim of translation underlying all this," Quine writes, "is that assertions startlingly false on the face of them are likely to turn on hidden differences in language."[28] This is certainly good advice for the translator or the literary critic. But what course would Quine have us follow if diligent effort can do no better, if any interpretation we can come up with either imputes absurd beliefs or is hopelessly ad hoc? Surely we do not want to insist, a priori, that everyone's inferential patterns must approximate some decent standard of rationality, since we can perfectly well imagine that our own inferential patterns might alter radically as the result of injury or illness. I think the right thing to say here (and, reading between the lines, I suspect Quine would agree) is that in these cases we should simply give up on translation. Translation is inextricably linked with saying what a person believes when he asserts the sentence being translated. And when a person's inferential patterns are sub-

stantially different from our own, we have no content sentence which will comfortably characterize his belief.

In a similar argument, focusing on the case of a radically exotic doxastic network, Davidson has challenged the coherence of the very notion of an alternative conceptual scheme.[29] If we can translate the native's claims and thus specify the content of his beliefs, then, Davidson urges, we are not dealing with the case of a *radically* different conceptual scheme. But if we cannot translate or assign content, then why should we assume we are dealing with a conceptual scheme at all? Why should we think that the untranslatable utterances of these exotic humans constitute speech behavior? Davidson's conclusion is that we should not, that if we cannot translate their language and assign content to their beliefs, then we have no grounds for thinking that these creatures *have* belief. But as other writers have noted, there is a plausible line of argument which leads to exactly the opposite conclusion.[30] We can readily imagine ourselves gradually turning into such doxastically exotic creatures by altering the elements of our doxastic network one belief at a time. Or, to change the image, we can imagine a sequence of people each ideologically similar to his immediate neighbors, with ourselves at one end and the unfathomable native at the other. What is more, in assuming the native to be ideologically exotic we need not assume that the global architecture of his cognitive apparatus has departed from the pattern in figure 2. He can still have belief-like states and desire-like states, even though we have no content sentences to characterize them. Further, these belief-like states may serve the native in good stead, enabling him to deal with his world quite successfully. Indeed, he may do a rather better job than we do. After all, if our race survives, our own scientifically sophisticated descendants may well have doxastic networks as different from ours as are those of our imagined natives. Surely, it is urged, it would be perverse parochialism to deny that these folk have an alternative conceptual scheme and quite a snazzy one at that.

In response to this looming antinomy, I would offer three observations. First, it is misleading to assume, as we have been throughout this section, that extreme ideological dissimilarity will always undermine content ascription. In an appropriate context, causal-pattern similarity may sustain the use of a quite ordinary English content sentence despite quite radical ideological dissimilarity. Second, in those contexts where ideological differences are both important and too great to warrant comfortable use of content sentences drawn from our own language, we are not forced into silence about the native's doxastic world. We can opt for an anthropological description, using their labels for their beliefs and detailing how their beliefs fit into their own form of life.

Finally, to the extent that Davidson's argument and the argument to the opposite conclusion constitute at antinomy, its source should by now be apparent. Our folk psychology evolved and earns its keep in settings where exotic folk were few and far between. When pushed to accommodate such cases, the judgments it renders are often neither clear nor consistent.

5. *Animal Beliefs*

There are many ways in which animals, particularly higher animals, are (or are thought to be) rather like people. They have needs and desires. They perceive the world around them and draw inferences from what they perceive. (Master is reaching for the leash, so we must be going for a walk. I smell food so it must be time to eat.) They form plans, albeit primitive ones, exploiting their beliefs to satisfy their desires. In short, it is plausible to assume that higher animals have a psychology whose global organization fits the pattern of figure 2. But there are also many ways in which animals are (or are thought to be) quite different from people. They are causal-pattern dissimilar, since neither their perception nor their inferential capacities work quite they way ours do. And they are ideologically dissimilar, since their doxastic network differs markedly from our own. Since they have no language, reference similarity is out of the question, though the causal history component of reference similarity may have an analogue in the causal history of animal concepts. These similarities and dissimilarities lead us to expect that we will have conflicting intuitions about the appropriateness of using everyday English content sentences to characterize the cognitive states of animals. In a context where nothing more than a rough causal-pattern and ideological similarity is called for, intuition will find such characterizations of beastly belief quite unexceptional. Suppose, for example, that Fido is in hot pursuit of a squirrel that darts down an alleyway and disappears from view. A moment later we see Fido craning his neck and barking excitedly at the foot of an oak tree near the end of the alley. To explain Fido's behavior, it would be perfectly natural to say he believes that the squirrel is up in the oak tree. But suppose now that some skeptic challenges our claim by focusing attention on the differences separating Fido's belief from ours. "Does Fido really believe it is a squirrel up in the oak tree? Are there not indefinitely many logically possible creatures which are not squirrels but which Fido would treat indistinguishably from the way he treats real squirrels? Indeed, does he believe, or even care, that the thing up the tree is an *animal*? Would it not be quite the same to Fido if he had been chasing some bit of squirrel-shaped and squirrel-smelling ma-

chinery, like the mechanical rabbits used at dog-racing tracks? The concept of animal is tied to the distinction between living and nonliving, as well as to the distinction between animals and plants. But Fido has little grasp of these distinctions. How can you say that he believes it is a squirrel if he doesn't even know that squirrels are animals?" Confronted with this challenge, which focuses attention on the ideological gap that separates us from Fido, intuition begins to waiver. It no longer sounds quite right to say that Fido believes there is a *squirrel* up the oak tree.

The ideological gap between animals and ourselves is not the only factor capable of undermining our confidence in ascriptions of content to animal beliefs. As we saw earlier, the use of a term in a subject's linguistic community serves an important role in nailing down the reference of a term and thus in determining the content sentences we find appropriate in characterizing beliefs which are expressed using the term. It was the influence of community usage that inclined us to say that John, the American, believes that chicory is bitter, while Robin, the Englishman, believes that endive is bitter.[31] And it is in some measure a consequent of community usage that you would characterize one of my beliefs as the belief that wallabies are marsupials, even though I cannot distinguish wallabies from other vaguely kangaroo-shaped creatures. But, of course, Fido does not express his belief verbally and is not a member of a linguistic community. So the fact that he does not distinguish squirrels from other (actual or possible) squirrel-like things generates puzzles about how his belief is to be characterized. If he treats squirrels and various kindred species indistinguishably, shall we say when he is chasing one of these non-squirrels, that he (mistakenly) believes it is a squirrel? Or should we say that he correctly believes it is a furrel, where 'furrel' is a new term denoting the heterogeneous collection of animals and artifacts that Fido treats indistinguishably from ordinary squirrels? Folk psychology, it would seem, provides no comfortable reply. And of course similar qualms might be directed at Fido's concept of a tree. If there is a range of tree-shaped artifacts which Fido cannot distinguish from real trees and if in fact the squirrel has run up one of these man-made ersatz, then does Fido mistakenly believe that the squirrel is up a tree? Or, rather, does he correctly believe that the squirrel has run up the tree-or-canine-deceiving-tree-shaped-artifact? Conundrums like this abound when the context makes it important to say just exactly what it is that an animal believes.[32]

Most of the writers who have addressed the question of animal belief have focused either on contexts where only rough-and-ready similarity is appropriate or on contexts which stress more fine-grained similarity.

Those who focus on contexts of the first sort insist that the attribution of beliefs to animals is unproblematic.[33] Those who focus on the second sort of context dwell on the impossibility of characterizing the content of an animal's belief and go on to urge that in the absence of finely discriminated content sentences, we should not apply the notion of belief to animals at all.[34] In one of my own papers I gave examples of both contexts, noted the conflict of intuitions, and concluded that the issue of whether animals have beliefs is moot.[35] But I would now argue that none of these three positions gets the matter quite right. If the question being asked is whether claims of the form

S believes that p

are ever true, when S is an animal, and 'p' is replaced by some quite ordinary English sentence, then the answer is clearly yes. But there are other conversational contexts in which the very same sentence, referring to the same animal and time, would be false. The apparent paradox here dissolves once we recognize that belief ascriptions are similarity claims, and similarity claims are context dependent.

6. Some Conclusions

The work of part I is not yet finished, since I still owe the reader an account of why I reject the putative distinction between *de dicto* and *de re* belief sentences. However, I think this would be an appropriate place to draw together some of the conclusions that follow from what has been said thus far. On the positive side we have an account of our commonsense concept of belief which appears to do quite a good job at capturing the data of intuition. As advertised, our account is in the spirit of the mental sentence theory, with mental sentences typed by content, though the relation of "content-identity" is ultimately replaced by a context-sensitive multidimensional similarity measure. On the negative side we have an argument against those accounts of the notion of belief which individuate beliefs along narrow causal lines. There are three reasons for rejecting accounts of belief which invoke a narrow causal standard. The first is the holism of belief ascription and individuation. Our intuitions about whether a belief-like state counts as the belief that p and our intuitions about whether a pair of beliefs count as identical in type are sensitive to the doxastic neighborhood—the network of further beliefs—in which the state resides. By altering the surrounding beliefs we can change a paradigm case of the belief that p into a state which common sense would not count as the belief that p at all. However, since these changes are quite independent of the state's causal potential, the narrow causal standard records no change.

The second reason for rejecting narrow causal accounts of belief individuation is the link between the ascription and individuation of belief on the one hand and reference on the other. The reference of a term is often determined in part by its distant causal ancestry and by the use of the term in the speaker's community. But neither of these factors need be reflected in the narrow causal profile of the belief state which would be expressed using the term. Thus it is possible to have a pair of belief state tokens which count as type identical by the narrow causal standard though intuition judges them to be quite different. The final reason for rejecting narrow causal theories is the context sensitivity of commonsense belief ascription and individuation. In one context intuition may judge that a given belief token is appropriately describable as the belief that p, while in another context intuition may find that label inappropriate. Judgments about whether a pair of belief tokens count as type identical are similarly context sensitive. If the narrow causal standard requires that a pair of belief tokens must have *identical* narrow causal profiles to be type identical, then it cannot begin to explain the context sensitivity of our judgments. If a narrow causal theory requires only *similarity* of narrow causal profiles then it has more room for maneuver, since similarity is a context-sensitive notion. But even so, the narrow causal account will be unable to explain the effect of those contexts which prime for ideological similarity or for elements of reference similarity.

Granting that the narrow causal version of the mental sentence theory does not fully capture our folk psychological notion of belief, it is worth pondering just how badly it misses the mark. Fodor, as we have seen, anticipated some slippage between the narrow causal standard of individuation and the content-based scheme of "aboriginal, uncorrupted, pretheoretical intuition."[36] However, on his view, the two classification schemes will coincide "plus or minus a bit."[37] The issue assumes considerable importance if, as Fodor maintains, the close (if imperfect) correspondence between narrow causal (or "formal") taxonomy and content-based taxonomy "is perhaps *the* basic idea of modern cognitive theory."[38] As will emerge in part II, there is reason to doubt that cognitive science need be much concerned about the size of the gap between these two taxonomic schemes. But if Fodor is right on this point, then cognitive science is in deep trouble. For it is quite clear that by any reasonable measure the divide between folk taxonomy and narrow causal taxonomy is *enormous*.

The best way to appreciate the difference between the two is to ask how far we could get in ascribing content to a belief-like state if our knowledge about the state is restricted to a complete specification of its narrow causal profile. It is my contention that if this is all we know,

then we can hardly begin to judge whether a given content sentence correctly characterizes the belief. Writing on a different theme, Fodor himself began to make this point very nicely. In his delightfully irreverent paper, "Tom Swift and His Procedural Grandmother,"[39] Fodor asks what we know about the semantic properties of a computer machine language (ML) sentence when we know the program specifying how that ML sentence interacts with other ML sentences. Does a knowledge of the potential causal connections among ML sentences enable us to determine their truth conditions, or the reference of their terms, or how they are to be interpreted? Fodor's answer is that it does not. To make the point, he asks that we

> Imagine two programs, one of which is a simulation of the Six Day War (so the referring expressions designate, e.g. tank divisions, jet planes, Egyptian soldiers, Moshe Dayan, etc., and the relational terms express bombing, surrounding, commanding, capturing, etc.), and the other of which simulates a chess game (so the referring expressions designate knights, bishops, pawns, etc., and the relational terms express threatening, checking, controlling, taking, etc.). It is a possible (though, of course, unlikely) accident that these programs should be *indistinguishable when compiled;* viz., that the ML counterparts of these programs should be identical, so that the internal career of a machine running one program would be identical, step by step, to that of a machine running the other.[40]

Fodor is quite right of course. The formal computational properties of ML sentences do not tell us whether they are about a chess game or the Six Day War. So if this is all we know about these sentences we can hardly begin to assess their semantic interpretation or their content. But now there is a clear analogy between ML sentences and belief states or mental sentence tokens. If all we know is the "program"—the potential causal interactions among belief states and other mental states—then we are in no position to assign content to the belief state. We cannot tell whether it is a belief about a chess game or about the Six Day War. Here it might be protested with some justice that the analogy is not a perfect one. For in the case of the computer we are assuming that we know only interstate causal links, while in the case of mental sentences we would have a complete narrow causal profile which includes potential causal links to stimuli and behavior. To make the analogy more exact, we should have to imagine our computer connected to sensory transducers and a robot, and we would have to know the potential causal links between transducer inputs and ML sentences, on one side, and between ML sentences and robot behavior

on the other. Would this additional information enable us to determine the semantic properties of ML sentences? Or, putting the question the other way around, does the information about causal links from stimuli to mental states, and from mental states to behavior, which is included in a narrow causal profile enable us to assign content to belief states or sentences in a mental code? In certain cases the answer is clearly no. Information about stimuli and behavior will not tell us that Tom believes Eisenhower was a politician, while Dick believes that Angell-James was a politician.[41] Nor will it tell us that John's belief is about chicory while Robin's is about endive.[42] Moreover, I think a plausible case could be made for a much more massive underdetermination of content by narrow causal profile. The key here is to imagine a community of people who from birth onward are embedded in devices which systematically alter the input to their natural sensory apparatus. Thus, for example, we might imagine that just after birth children in this society have tiny TV systems attached to their eyes and that the images projected into their eyes by these systems systematically alter the true color of objects in the environment. Facing a ripe tomato, children in this society would be subjected to, say, a green spherical stimulus. If we fill in the details in the right way it will be clear that the belief they acquire from this green stimulus is that the object before them is red. From this small beginning, aficionados of philosophical science fiction will see how we could construct a much more elaborate case which would in effect duplicate the old chestnut example of the brain in the vat. The difference is that what we end up with is not a bodyless brain in a vat but a full, normal body whose sensory apparatus is provided with synthetic input and whose bodily movements are monitored to provide appropriate synthetic feedback. The conclusion to be drawn from this rather outlandish case is that a full narrow causal profile will not enable us to characterize the content of the subject's belief states nor to determine the semantic properties of sentences in his mental code. To assign content we must know something about the history of his concepts, the linguistic practices prevailing in his community, and the way in which his mental states are causally related to actual objects in his environment. In short, we must know how the subject is embedded in the world.

One brief final point must be added. In arguing against accounts of belief which invoke a narrow causal standard, my thesis has been that these accounts do not capture our folk psychological notion of belief. I have not argued, nor do I believe, that psychological states which are individuated along narrow causal lines should be eliminated from cognitive science. Indeed in part II, I will argue that the narrow causal

standard or something very close to it ought to be (and often is) respected in serious cognitive science. But if this is right, then since the folk notion of belief is at odds with the narrow causal standard, it has no place in a mature cognitive science.