Chapter 3

Beliefs as Mental Sentences

At the end of the previous chapter we reviewed some reasons for thinking that the notion of belief is best construed as a relational notion. By and large, relational accounts of belief divide into two categories, the division turning on the sort of thing that fills the second slot in the belief relation. Theories in one of the categories take belief to be a relation between a person and a proposition. These accounts differ among themselves on just what a proposition is and on the nature of the relationship between the person and the proposition that is the object of that person's belief. But there are two points on which propositional accounts agree: propositions are some sort of *abstract* entity, and propositions are not sentencelike creatures—they do not have a *syntactic* structure. Theories in the second category take belief to be a relation between a person and a *sentence*. On this view, to have a belief is to have a sentence token suitably inscribed or encoded in one's brain. In this book I will have very little to say about propositional theories of belief. This for two reasons. First, these theories have made relatively little impact on cognitive science; those who hope that contemporary cognitive science will reunite the scientific and the commonsense world views are almost always advocates of a sentential account of belief. Second, I think it is already abundantly clear that propositional theories, whatever their virtues, simply do not tell the right story about belief as it is conceived in folk psychology. The case for this point has been well stated by others.[1]

The bulk of this chapter will be devoted to answering the two obvious questions about sentential theories of belief: What, precisely, do they claim, and what reason is there to accept them? Before attending to either of these matters, however, a few remarks are in order on the divergent goals of sentential theorists on the one hand and theory-theorists like Lewis on the other.

1. *Protoscience and Conceptual Analysis*

In the discussion of philosophical behaviorism I noted that it brought

in its wake a subtle shift in the sort of questions philosophers have asked about the mind. For Descartes or Hume the central questions were ontological: What sort of thing or stuff is a mind or mental state? How is it related to matter? These are questions about the nature of things and, apart from their generality, they are of a piece with the questions that would be asked by a natural scientist interested in the mind. For the philosophical behaviorist, by contrast, the central questions are those of conceptual analysis: How is the concept of mind (or belief, or pain) to be analyzed? What is the meaning of mental state terms? As philosophical behaviorism has lost its grip on contemporary philosophy of mind, the protoscientific questions have once again come to the fore. Thus most defenders of the mental sentence view do not take themselves to be doing conceptual analysis. Rather, they view themselves as engaged in a process which is continuous with the doing of science. They want to know what sort of thing a belief is, and not merely how our ordinary concept of belief is to be understood. Since they are attempting to build a theory about a part of the natural world, they feel free to marshal arguments and evidence from any quarter which may prove helpful. There is no need for them to restrict themselves to facts about our commonsense concept. Still folk notions are not irrelevant to the protoscientific project. If we want to know the nature of the entities denoted by a certain commonsense term or falling under a certain commonsense concept, then we had better keep a sharp eye on the way that concept is used, lest we end up describing the wrong critters. Conceptual analysis also has a more substantive role to play in the protoscientific project. For it is sometimes possible to argue that, given the contours of our commonsense concept, the entities falling under the concept must have (or are likely to have) certain features. What is more, the features need not be the ones which are, in any sense, *entailed by* or *built into* the folk notion. In effect such arguments attempt to show that, in virtue of various facts, including some about our commonsense concept, the *best hypothesis* about the entities falling under the concept is that they have certain features. I am laboring this point to forestall a possible misunderstanding of mental sentence theories. In arguing that belief is a relation between a person and a sentence token in that person's brain, mental sentence theorists are *not* claiming that this is part of our ordinary *concept* of belief. Rather, they are arguing that this is the best hypothesis *about belief*, given a range of facts including some about our commonsense concept and locutions.

2. *Some Features of the Concept of Belief*

Let us begin our discussion by assembling some of the facts about our

ordinary notion of belief which sentential theorists have used in arguing for their view.[2]

(1) First, and most obviously, beliefs (and related states like fears, wants, hopes, etc.) are standardly named and attributed by linguistic constructions involving an embedded sentence or sentential transform. There are other ways of referring to beliefs, of course. My students sometimes refer to a belief of mine as "Stich's favorite belief." But if asked what that belief is, they would reply, "the belief that Ouagadougou is the capital of Upper Volta." Moreover, as Vendler has shown,[3] there are striking parallels between the syntactic behavior of belief sentence complements and the syntactic behavior of the complements in sentences of the form: 'S said that p' or 'S asserted that p'.

(2) As we have noted, 'believes' seems to express a two-place relation. Ordinary language provides us with the resources for distinguishing between a belief and the second element in the belief relation, which might be called *what is believed* or *the object* of the belief. The distinction emerges clearly in a passage like the following:

> John believes that war builds character. Given his early education, his belief is to be expected. Though, of course, what he believes is utter nonsense.

Ordinary language also provides us with natural ways of expressing the fact that a pair of beliefs share the same object:

> John believes that war builds character, and after his rousing lectures many of his students believe it too.

Beliefs can also share the same object with other attitudes:

> John believes that we have turned the corner on inflation, but Maggie doubts it. I wish it were true.

(3) As they are conceived and spoken of in our folk theory, both beliefs and the objects of beliefs can have semantic properties. Thus we may say either

> Maggie's belief is true.

or

> What Maggie believes is true.

However, the truth value of a belief must be the same as the truth value of the object of the belief. Thus common sense finds nothing but paradox in a claim like

> John's belief is true, though what he believes is false.

Also, the truth value of both a belief and of what is believed must match the truth value of the embedded sentence that would be used in ascribing the belief. So if Maggie believes that there are no pandas in Tibet, then Maggie's belief is true if and only if the sentence

There are no pandas in Tibet

is true.

The situation is analogous for entailment, logical equivalence, and other semantic relations. These relations can obtain both between beliefs:

Maggie's belief entails John's

and between the objects of beliefs:

What Maggie believes entails what John believes.

Here too, the semantic relations between the objects of belief must parallel the semantic relations between the embedded sentences used in ascribing the beliefs. So if Maggie believes that there are no pandas in Tibet, and John believes that there are no pandas in northern Tibet, then what Maggie believes entails what John believes if and only if

There are no pandas in Tibet

entails

There are no pandas in northern Tibet.

The various parallels we have noted can hardly be an accident. A theory about the nature of belief should give us some explanation of the fact that the semantic properties of beliefs, the objects of beliefs, and the sentences embedded in belief ascriptions all coincide.

(4) A bit of terminology will prove useful. Let us call sentences of the form

S believes that p

belief sentences, and let us call the embedded sentence, p, *the content sentence*. What we have just seen is that the semantic properties of beliefs and of the objects of beliefs parallel those of the associated content sentences. When our attention turns to the semantic properties of belief *sentences*, however, the situation is much more puzzling. For, though content sentences are embedded within belief sentences, the semantic properties of belief sentences seem to be quite independent of the semantic properties of their own content sentences. The logical perversity thus engendered has been, until recently, the main focus of philosophical interest in belief. To underscore the semantic puzzles posed by belief sentences, it is useful to compare them with superficially

analogus forms like negations. Syntactically, negation functions as a sentence forming operator. Given any declarative sentence, p, we can form a new sentence by embedding p in

It is not the case that _____ .

The situation for 'believes that' is quite similar. Given any declarative sentence, p, we get a new sentence by inserting p for '_____' and an arbitrary name for ' . . . ' in

. . . believes that _____ .

But when our focus shifts to semantics, the analogy quickly breaks down. In the case of negation, the semantic properties of the compound sentence are determined by the semantic properties of the embedded sentence. 'It is not the case that p' is true if and only if p is false. With belief sentences, the truth value of the content sentence tells us nothing about the truth value of the compound. From the (rather surprising) fact that

Los Angeles is east of Reno

is true, we can conclude neither that

Quine believes that Los Angeles is east of Reno

is true, nor that it is false.

The entailment relations of belief sentences also fail to follow those of their content sentences. Thus, from the fact that p entails q, it does not follow that

S believes that p

entails

S believes that q.

If it did, our beliefs would be closed under entailment, which would make mathematics much less of a challenge, and our fellow citizens much less exasperating. (Or perhaps it would make them much *more* exasperating. It seems likely that each of us has at least one pair of contradictory beliefs. And if we believe everything entailed by our inconsistent beliefs . . . .)

The failure of belief sentences to mirror the entailments of their content sentences deserves special note when the content sentence inference is of the following form:

$$\frac{\begin{array}{ll} Fa & \text{(e.g., Bart is a spy)} \\ a = b & \text{(e.g., Bart is the president of Yale)} \end{array}}{Fb \quad \text{(e.g., The president of Yale is a spy.)}}$$

If we embed the first premise in a belief sentence, we get the following inference:

S believes that Fa	(e.g., David believes that Bart is a spy)
a = b	(e.g., Bart is the president of Yale)
S believes that Fb	(e.g., David believes that the president of Yale is a spy)

And this latter inference is certainly not generally valid. However, a venerable tradition insists that there is a *sense* of 'believes that' on which the second inference is valid. This (alleged) sense is called the *relational* or *de re* sense, and referring expressions (like 'Bart' or 'the president of Yale') are said to occur *transparently* in the content sentences of *de re* belief sentences. In contrast, the sense of 'believes that' on which this last inference is not valid is labeled the *de dicto* sense, and referring expressions are said to occur *opaquely* in the content sentences of *de dicto* belief sentences. One of the less orthodox theses of this book is that the putative distinction between *de dicto* and *de re* belief sentences is a philosophers' myth, corresponding to nothing sanctioned by folk psychology. But that is a story for a later chapter. For the moment we are accumulating uncontroversial facts about our commonsense notion of belief. And a point beyond dispute is that (at least on one reading) inferences like the last one displayed are not generally valid.

(5) Another central theme of folk psychology is that mental states interact causally with one another, producing new mental states and, ultimately, behavior. These causal interactions are not random. Rather, folk theory maintains that the pattern of causal interactions often mirrors various *formal* relations among the content sentences that would be used in ascribing the states. Thus, for example, if Maggie believes that all Italians love pasta, and if she comes to believe that everyone at Sven's party is Italian, she will likely come to believe that everyone at Sven's party loves pasta. The folk generalization of which this is an instance is that a belief of the form:

All As are B

and a belief of the form:

All Bs are C

typically causes a belief of the form:

All As are C

What is important here is the natural, indeed all but inevitable, way

in which our folk theory leads us to talk about the *logical form* of the objects of belief. What is believed, folk psychology seems to suggest, has some sort of logical form, and often it is in virtue of the logical forms of their objects that one belief causally interacts with another. Much the same point can be made about the causal interactions among beliefs and other "contentful" mental states. Suppose Sven wants to visit New Zealand, and suppose that he believes that if he is to visit New Zealand then he must obtain a visa. Here, folk psychology urges, we might typically expect Sven to form the desire to obtain a visa. The folk generalization in this case is (roughly) that a desire of the form:

Do A

along with a belief of the form:

In order to do A it is necessary to do B

leads to a desire of the form:

Do B.

For current purposes, details are not essential. What is important is that in capturing the generalizations of practical reasoning, it is plausible to view the objects of belief and desire as having a logical structure which determines the ways in which they interact.

A caveat should be thrown in here. The mental sentence theorist need not claim that *all* the causal interactions among contentful mental states are dependent on their form. Plainly folk psychology recognizes other sorts of causal interactions, some systematic and others not. All the mental sentence theorist need claim is that some quite central causal patterns recognized by folk theory seem to turn on the logical form of the objects of belief.

3. *Sentences in the Head*

I have been collecting facts about our commonsense notion of (and talk about) belief, with the aim of reconstructing an argument to the effect that the best explanation of all these facts would be the hypothesis that belief is a relation between a person and an internally inscribed sentence. The chore now is to see how that hypothesis would account for the evidence collected. Before beginning I had best address a question that may have been troubling the reader since this talk of mental sentences began, to wit: What does it *mean* to talk of having a sentence in the head?

Let me address the most skeptical group of readers first, those who suspect that talk of sentences in the head is *conceptually incoherent.*

Surely this is too strong a criticism. For consider. We might, on a close examination of the contents of the head, discover there a tiny blackboard on which English sentences are literally inscribed in chalk. Or, if fancy runs to more contemporary technology, we could discover a tiny television monitor or CRT with thousands of English sentences displayed on the screen. Since there is nothing conceptually incoherent about this fantasy and since it would surely count as discovering that there are sentences in the head, we can safely conclude that the sentence-in-the-head hypothesis is not conceptually incoherent.

It is instructive to note that even if, *mirabile dictu*, we all have CRTs covered with sentences inside our heads, this would not be nearly sufficient to establish that beliefs are sentences in our head. The sentences on the screen would, at a bare minimum, have to stand in some plausible correlation to what the owner of the head actually believes. So, for example, if the only sentences to be found in my head are the text of Gibbon's *Decline and Fall of the Roman Empire*, these sentences could hardly be the objects of my beliefs. Suppose, then, that the sentences found in my head are the right ones—that for every true sentence of the form

Stich believes that p

there is a token of p on my internal CRT. Suppose, further, that the screen is regularly updated. When an elephant comes into view, a token of 'there is an elephant in front of me' is added to the sentences on the screen, and when the elephant ambles away, the sentence disappears from the screen. This, no doubt, would make it much more plausible that these sentences could be identified as the objects of my beliefs. But something more would still be required. For suppose that the sentences on my CRT, while keeping an accurate record of what I believe, play no causal role in the dynamics of my mental and behavioral life. Surgically removing the screen has no effect on my behavior or on my stream of consciousness. Here, I think, we would be inclined to treat the sentences not as objects of belief but as some sort of epiphenomenon, a psychologically irrelevant causal product of my beliefs. The situation would be different if the sentences on my CRT played causal roles of the sort attributed to beliefs in our folk psychology. Suppose, for example, that inference is causally dependent on what appears on the screen. Imagine that there is an "inference device" which scans the screen, causing new sentences to appear on the screen depending on what it has scanned. Thus, if the inference device scanned tokens of 'All Italians love pasta', and 'Sven is an Italian', it would cause a token of 'Sven loves pasta' to appear on the screen. To test whether the token on the screen was really *causally* involved in the inference process, we

might obscure the bit of the screen displaying 'Sven is Italian' and see if this breaks the causal chain, so that the inference device now does not cause 'Sven loves pasta' to be added to the display. Analogously, let us imagine that the sentences on the internal CRT play the right sort of causal role in practical reasoning, details here being left to the imagination of the reader. If all that we have imagined were true—if there are internal sentences, if they correspond to what the subject actually believes, and if they play a suitable causal role in mental processes and the production of behavior, then surely we would be on safe ground to conclude that beliefs are relations between persons and internal sentences.[4]

All of this might be enough to establish the *intelligibility* of the mental sentence view. But what about its *plausibility*? We know that there are neither blackboards nor CRTs in the head; indeed there is nothing in the head that *looks* anything like a sentence token. Here the defender of the mental sentence view must insist that appearances are deceiving. We have many familiar examples of sentences encoded in a medium which cannot be detected by direct inspection. I relieve the tedium of my daily commute by listening to books recorded on cassette tapes. These tapes are, in a quite unproblematic sense, covered with sentence tokens, though the tokens are unrecognizable without the aid of a sophisticated piece of electronic apparatus. Similarly, the mental sentence theorist does not expect to find the sentences in the brain inscribed in Roman letters. Rather, if they exist at all, they will be recorded in some neural or neurochemical code.

This leaves us with the nice question of just when some complex neurochemical state is to *count* as being an encoding of a sentence, be the sentence in English or in some other language. My guess is that there can be no satisfying, fully general answer to this question. Certainly none has been offered. Part of the answer will inevitably require that there be a coding or mapping of units in the language (words, morphemes, or whatever) onto types of neurochemical states. A particular instance (or token) of the state to which a word is mapped will then count as an encoded token of the word. The code will also have to map at least some syntactically salient relations among words or units (for example, the relation of being the next word in the sentence) onto relations among neurochemical states. This will enable us to identify a sequence of suitably related neurochemical states as an encoding of a sequence of words. If that sequence of words is a sentence, then the sequence of suitably related neurochemical states will be the encoding of that sentence. This is hardly the last word on the question of coding. Indeed, before we end this chapter we will have reason to puzzle further over just what is to count as the neural encoding of a sentence.

But for now I think enough has been said. Let us grant to the mental sentence theorist that the existence of sentences in the brain is both intelligible and not incompatible with anything we know about the brain.

4. *Explaining the Facts about Belief*

The contention of the mental sentence theorist, recall, is that we can best explain facts about our commonsense notion of belief by hypothesizing that belief is a relation between a person and an internally represented sentence. It is time to see how these explanations are supposed to work. To show the view in its best light, let us assume that in all of our heads there are enormous numbers of sentence tokens which are unproblematic encodings of just those English sentences which express the contents of our current beliefs. Indeed for the sake of vividness we can imagine for the nonce that we really do have a little CRT in our heads covered with sentence tokens. Further, let us assume that these sentence tokens play the causal role of beliefs as this is depicted in folk psychology. Given these favorable assumptions, what can be said about the various facts assembled earlier in this chapter?

First, it is no surprise that beliefs are standardly attributed by a construction involving an embedded sentence. On the mental sentence hypothesis, my belief that Ouagadougou is the capital of Upper Volta is a relation between me and a token of 'Ouagadougou is the capital of Upper Volta'. So the belief sentence

Stich believes that Ouagadougou is the capital of Upper Volta

wears its logical form on its face—well, almost on its face. To say that a relation obtains between a pair of objects, we flank a relational predicate by *names* of the objects, and while 'Stich believes that Ouagadougou is the capital of Upper Volta' has my name to the left of the relational predicate 'believes that', to the right it has a sentence, not the name of a sentence. It would be more accurate to write

Stich believes 'Ouagadougou is the capital of Upper Volta'

thereby making it clear that the second element in the relation is a sentence. It is not surprising that this is not ordinarily done, since ordinary usage is notoriously sloppy about the distinction between use and mention.*

*Actually,

Stich believes 'Ouagadougou is the capital of Upper Volta'

still does not quite capture what the mental sentence theorist intends, since quotation

Once we have made the implicit quotation marks explicit, the various logical peculiarities of belief sentences have a straightforward explanation. They are all just what we would expect, given that the content sentence is mentioned, not used. Changes made within a quoted context produce the name of a *different* sentence, and there is no reason to expect that a person who had an internal inscription of the first sentence will also have an internal inscription of the second just because the two sentences have the same truth value or are logically equivalent. In particular, it is no surprise that

David believes 'Bart is a spy'

may be true, while

David believes 'the president of Yale is a spy'

is false, even though Bart is the president of Yale. This latter identity guarantees that the two content sentences will have the same truth value, but they are still very different sentences, one of which may be internally represented in David while the other is not. So, on the mental sentence view, the inference from

David believes that Bart is a spy

and

Bart is the president of Yale

to

David believes that the president of Yale is a spy

is strictly parallel to the inference from

'Bart is a spy' was written on the blackboard

and

Bart is the president of Yale

to

'The president of Yale is a spy' was written on the blackboard.

Since the latter inference is patently invalid, we should expect the former to be invalid as well.

---

mark names denote types, not tokens. What is needed is something like:

Stich believes a token of 'Ouagadougou is the capital of Upper Volta'

or perhaps even better:

($\exists$t) (t is a token of 'Ouagadougou is the capital of Upper Volta' and Stich believes t).

But let us not quibble over small points.

The mental sentence view has an equally easy time explaining the semantic facts about what is believed (i.e., about what I have been calling the *objects* of belief). Here, recall, the most striking fact was that the semantic properties of the object of a belief coincide exactly with those of the associated content sentence. But this is exactly what would be expected if the object of the belief just *is* a token of the content sentence. We noted that common sense ascribes truth and other semantic properties to beliefs as well as to the objects of belief and that these too matched the semantic properties of the appropriate content sentence. Here, I suppose, the best line for the mental sentence theorist to take is that talk of the truth, consistency, or entailments of a belief is just a shorthand for talk about the truth, consistency, or entailments of the object of that belief.

Finally we saw that many of the most useful generalizations of folk psychology require that beliefs (and other contentful mental states) be categorized according to the logical form of their objects. For this to make sense, the objects of belief must be the sort of things which *have* logical forms. And of course sentences are possessors of logical form par excellence. Glimpsing ahead a bit, we can see why mental sentence theorists are so sanguine about integrating the contentful states of folk psychology into serious scientific psychology. For if beliefs and their kin are relations to sentence tokens whose causal interactions are largely determined by their logical forms, then we have a clear research program. What we want to discover are the forms of various mental representations and the principles governing the devices or processes which manipulate representation tokens in virtue of their forms. While we may have no good leads on the neurophysiology of these token-manipulating mechanisms, we know that there is nothing in principle mysterious about physical devices which manipulate sentence tokens in virtue of their form. For this, near enough, is just what is done by an electronic digital computer.

Some of the defenders of the mental sentence account would insist that their view offers more than a research *program* in cognitive psychology. They contend that the view is "presupposed by the best—indeed the only—psychology that we've got."[5] If this is correct, it is a powerful independent argument for the mental sentence theory of belief. In chapter 9 we will take a careful look at the arguments offered in support of the claim that the best in contemporary psychology presupposes the mental sentence view.

5. *The Language of Thought*

All that we have said so far makes the case for the mental sentence

hypothesis look remarkably good. But before we get carried away we should note that the argument we have been reconstructing rests precariously on a single implausible assumption, viz. that the sentences in our heads which are the objects of our beliefs are *English* sentences. Note that the assumption I am questioning is that the objects of belief are tokens of English sentences, not the fanciful proposal that they are written on an internal CRT. The latter conceit is offered merely as an aid to the imagination, and nothing said in the previous section depends on it. The assumption might be passingly plausible if "our heads" referred only to yours and mine, English speakers both. But interpreted less xenophobically, the assumption is hopeless. Surely we do not expect to find tokens of English sentences inside the heads of monolingual Korean speakers, not to mention the heads of prelinguistic children or the family dog. Yet folk psychology clearly ascribes beliefs to all of these.[6] So something else will have to be said about the objects of belief.

There are two proposals that have been taken seriously. The first, suggested by Harman,[7] and tentatively endorsed by Field,[8] is that the objects of belief (and other contentful mental states) are sentences in the language of the *believer*, or perhaps something a bit richer, like sentences in the believer's language paired with their phrase structure trees. On this view, my beliefs are relations to internal tokens of English sentences, and the monolingual Korean's beliefs are relations to internal tokens of Korean sentences. What to say about beasts and babies is less clear. The other view, championed by Fodor,[9] takes the objects of belief to be sentences in a species-wide mental code, *the language of thought*. On this view, the Korean who believes that dry food is good for man is related to the same sentence in universal mental code as was Artistotle, who held the same belief, though he expressed it in Greek. Since the language of thought is supposed to be species wide and innate, the beliefs of prelinguistic children pose no special problems. Fido presumably holds his beliefs in the language of canine thought. Since Fodor's view has been rather more influential, I will focus my discussion on it. But just about all the points to be made can be made mutatis mutandis if we assume instead that the objects of belief are sentences in the language spoken by the believer.

When we were pretending that all beliefs had English sentences as their objects, we had a relatively easy time explaining the facts we had collected about our commonsense notion of belief. On dropping that pretense, however, many of those explanations begin to come unglued. Consider the story we told about why beliefs should standardly be ascribed by a locution involving an embedded sentence. What could be more natural if the belief simply is a relation between the believer

and an internal token of *that very embedded sentence*? But what are we to say now that we are assuming belief is a relation between the believer and a sentence in the language of thought? Why do we standardly attribute this state using a locution involving an embedded *English* sentence? And what is the relation between the English content sentence and the object of the belief which, ex hypothesis, is a sentence in the language of thought? This last question is perhaps the central one, for until we have some account of the relation between content sentences and the objects of belief, most of the other explanatory stories told in the previous section cannot be reconstructed. It is, for example, a renewed puzzle why the semantic properties of the objects of belief should coincide with the semantic properties of content sentences, a puzzle that will be solved only if the relation between mental sentences in the language of thought and content sentences in English turns out to be one which preserves semantic properties. Similarly, the story we told to explain the opacity of belief sentences rested heavily on the assumption that the relation between content sentence and object of belief was the relation of type to token. With this assumption denied, the explanation no longer hangs together.

Well, what can the mental sentence theorist say about the relation between content sentences and the objects of belief? One idea that seems natural enough derives from the observation that the content sentence is the one *we* (i.e., English speakers) would use to express the belief which we attribute to others by embedding that sentence in a belief sentence. So, for example, we would express the belief that it's raining by *saying*, 'It's raining'. French speakers would express the belief in other words, and babies (if they have such beliefs) would not express them linguistically at all. Following up on this idea, Fodor introduces a function F, "from (e.g. English) sentences onto internal formulae,"[10] that is, onto what we have been calling sentences in the language of thought. He then proposes the following "principle" to connect the object of the belief that it's raining (which, recall, will be a token of an internal formula) with the English sentence 'It's raining':

> 'It's raining' is the sentence that English speakers use when they are in the belief-making relation to a token of F(it's raining) and wish to use a sentence of English to say what it is that they believe.[11]

A bit later, he writes that

> F(it's raining) is distinguished by the fact that its tokens play a causal/functional role (not only as the object of the belief that it's raining, but also) in the production of linguistically regular utterances of 'it's raining'.[12]

This last quote could do with a brief gloss. It is only *chez nous*, among English speakers, that utterances of 'It's raining' count among their causes tokens of the mental sentence F(it's raining). In another possible language utterances of 'It's raining' might typically be caused by tokens of F(there is nothing but cabbage for dinner). There are other shortcomings in Fodor's story about the function from English sentences to mental formulas which are less easy to patch.[13] But I am inclined to think that Fodor's basic idea is on the right track. What is of crucial importance for our current concerns is that if anything much like Fodor's account is accepted, then the problem of the relation between sentence types and their mental or neurochemical tokens must come to center stage once again.

6. Sentence Type and Mental Token

To see just why this issue assumes critical importance, let us trace through the sort of account a Fodor-style theory would have to give of a belief sentence like

Otto believes that it's raining.

What, precisely, does this sentence tell us about Otto? Or, to put the question more fashionably, what are the sentence's truth conditions? Well, first off, the sentence is ascribing a belief to Otto, and ex hypothesis to have a belief is to have a token of a sentence (or formula) of the language of thought properly stored in the head. A long story could be told about "properly stored"—it must be stored in such a way that it functions as a belief, interacts with other beliefs and with desires in the proper way, and so on. But this part of the account is not now in question, so let us take it for granted. What about Otto's mental sentence token itself? What does our belief ascription tell us about it? Here is where the content sentence comes in, with the story going something like this: the content sentence is 'it's raining.' Among us, English speakers, that content sentence is correlated with a particular internal code sentence, viz. the one whose tokens typically cause us to say 'It's raining'. So now we have specified a sentence type in internal code. And what about Otto? What is his relation to this sentence type? Well, it's simple; the sentence token he has in his head is a token of the very same internal code sentence type, though Otto, who speaks only German, would express his belief by saying 'Es regnet'. In short, Otto and the English speaker who sincerely says 'It's raining' have in their respective heads *tokens of the same mental sentence type*. So for a Fodor-style account of belief sentences to hang together, we must have some workable notion of what it is for two distinct people, speaking different

languages, to have in their heads distinct tokens of the same sentence type.

At first blush this might seem to pose no problem. After all it seemed in section 3 that we could make plausible sense of a person having a sentence token in his head. So why not use the same account in talking of a pair of people with tokens of the same sentence? To see where the problem lies, remember the story told in section 3 about the encoding of a sentence. The idea was that the first of a pair of mappings would take sentence parts (like words) onto types of neurochemical states, and the second would take syntactic relations (like the relation of being the next word) onto relations among neurochemical states. The trouble with this, as it were, purely syntactic story about the type-token relation is that it simply will not do the work needed in a Fodor-style account of belief sentences.

A bit of science fiction will help underscore the difficulty. Let us revert to an earlier fancy and suppose that inside each person's head there is a tiny CRT covered with sentence tokens. Suppose further that these sentence tokens are functioning as the objects of belief, playing the causal role assigned to beliefs by folk psychology. Finally, let us suppose that all the sentences on all human belief screens appear to be written in the same language. Indeed to make matters really simple, suppose them all to be written in what appears to be *English*. Granting all this, the obvious mapping from English sentence types to "mental tokens" is simply the one which maps sentence types to their CRT inscriptions. With the generous assumptions we are making, it should be particularly easy to tell Fodor's tale about Otto. Let us assume that when you and I produce "linguistically regular utterances" of 'it's raining', the causally implicated mental sentence is a token of 'it's raining'. When we say 'Otto believes that it's raining', we are saying that Otto has on his CRT a token of the same type as the tokens that are causally implicated in our production of "linguistically regular utterances" of 'It's raining'—that is, that he has a CRT inscription similar in shape to the following:

It's raining.

But, I claim, this purely syntactic story simply cannot be right. Merely having an inscription of the right shape on his belief-CRT can't be all there is to Otto's believing that it's raining. Suppose, for example, that we note the following correlation: Whenever Otto is awake and alert, and it begins to *snow* around him, a token of 'it's raining' appears on his CRT. Suppose further that Otto regularly opens his umbrella when 'It's snowing' appears on his CRT but never does so when 'It's raining' appears. And, looking around at the other sentences on his CRT, we

find tokens of 'Rain is white', 'It generally rains in the winter and snows in the summer', 'In the spring, the flowers need plenty of snow' etc. Finally suppose that when Otto wishes to express the belief whose object is the token of 'It's raining', what he says is 'Es schneit'. Given all of this, what are we to say Otto believes when a token of 'It's raining' appears on his CRT? Clearly the right answer is that Otto believes that it's snowing. But on the purely syntactic story about the type-token relation, he believes that it's raining! For on *that* story, to believe that it's raining simply is to have an 'It's raining' token on the belief CRT.

At this point it might be protested that, with a bit of fiddling, what I have been calling the purely syntactic story could perfectly well attribute to Otto the belief that it's snowing when a shape similar to

It's raining

appears on his CRT. We have been tacitly assuming that the mapping from English sentences to mental tokens is the same in all people, and thus if my mental token of the sentence 'It's raining' is similar in shape to the inscription just displayed, Otto's must be too. But since this leads to absurd consequences, let us give up the assumption that everyone uses the same mapping. Perhaps in Otto's encoding of the language of thought, the word 'rain' does not map onto CRT inscriptions similar in shape to

rain

but rather to CRT inscriptions similar in shape to

snow.

Making this assumption (and assuming that the rest of Otto's mapping works just like mine), we can attribute to Otto the belief that it's snowing since the inscription on his CRT which is shape-similar to

It's raining

is in fact a token of (i.e., is mapped onto) the sentence type 'It's snowing'.

All this is true enough. Unfortunately it leads us out of the frying pan and into the fire. For recall that we started out attempting to reconstruct a Fodor-style account of the truth conditions for 'Otto believes that it's raining'. On the first proposed mapping, the one that mapped 'rain' to shape-similar tokens, we got the wrong truth conditions. To remedy the problem, it was noted that we can map 'rain' to inscriptions shape-similar to

snow.

And so we can. But we can also map 'rain' to inscriptions shape-similar to

sleet

or, for that matter, to inscriptions shape-similar to

dark of night.

Mappings, after all, are just functions, and functions are cheap. So if a brain state (e.g., an inscription on a cerebral CRT) counts as a token of a sentence merely in virtue of there being *some* mapping from the sentence onto states of this type, then any brain state (and thus any CRT inscription) will count as a token of any sentence we choose, and we are left without any coherent account of the truth conditions for 'Otto believes that it's raining'.

It appears then that we are confronting the following dilemma. If we assume that the mapping from English sentence types to mental sentence tokens is a "purely syntactic" one on the lines sketched in section 3, then if the mapping is uniform from one person to the next, we get wildly incorrect truth conditions for belief sentences. However, if the mapping is not uniform from one person to the next, we get no coherent account of the truth conditions for belief sentences. The conclusion to draw is obvious enough. The purely syntactic mapping from English content sentences to mental sentence tokens must be abandoned. What is needed, in its stead, is a mapping which preserves or mirrors some more interesting properties of the sentences mapped.

One defender of the mental sentence account of beliefs who has confronted this issue directly is William Lycan, and it will be instructive to consider his view in some detail. In setting out his view, Lycan adopts Sellarsian dot quote notation, where dot quotes "are common noun forming operators that also serve to ostend the linguistic tokens that they enclose."[14] With this notation,

Jones believes that broccoli causes erysipelas

would be rendered

Jones believes some ·Broccoli causes erysipelas·.

"A slight variation would be to express the force of [this last sentence] as

Jones believes one of *those*. / Broccoli causes erysipelas."

Belief, then, "is construed as a dyadic relation which a person bears to a linguistic or quasi-linguistic token that falls into a certain category."[15] The object of Jones's belief that broccoli causes erysipelas is a mental

sentence token, one falling within the extension of the predicate 'is a 'Broccoli causes erysipelas''. This, of course, is just Fodor's view with a different notation. In Lycan's Sellarsian notation, the question pondered in the last few paragraphs—what mental state would count as a token of a given content sentence—is transformed into a question about the extension of dot quote predicates:

> How are we to determine the extension of the predicate 'is a 'Broccoli causes erysipelas''? Alternatively, how are we to tell when some linguistic or quasi-linguistic token of some quite other shape *is* 'one of *those*'?[16]

Lycan notes a pair of plausible alternatives. The first, a Sellarsian idea, is to group tokens by their *inferential role*: "an item will count as a 'Broccoli causes erysipelas' just in case that item plays approximately the same *inferential role* within its own surrounding conceptual framework that the sentence 'Broccoli causes erysipelas' plays in ours."[17] This idea needs a bit of interpreting, since, as quoted, it seems to be comparing the inferential role of a sentence token (a physical state or object) with the inferential role of a sentence type (an abstract object). But the Sellarsian standard need not make a detour through sentence types. Rather, I think, the proposal is best viewed as comparing foreign mental sentence *tokens* with domestic ones. A state or token in Jones's head counts as a 'Broccoli causes erysipelas' if it plays an inferential role similar to that of the states or tokens that underlie normal assertions of 'Broccoli causes erysipelas' in me, or perhaps in English speakers generally. Inferential role here is simply part of the broader causal role that the tokens in question play in the cognitive dynamics of their respective heads. And it is not entirely clear just how much of this causal network Lycan or Sellars would wish to count under the rubric of inferential role. Practical reasoning, I suppose, would naturally be included. But what about the causal links tying the objects of belief to perception? For example, we would expect that a good glimpse of broccoli in front of him would generally lead to the production of a 'There is some broccoli in front of me' in the appropriate place in Jones's brain. It is less natural to count this strand of the causal network as part of the "inferential role" of the sentence token. However, there is no need here to settle just what Lycan or Sellars may have intended. I raise the issue to serve as a backdrop for the introduction of some terminology of my own. On the Sellarsian view, as I have been interpreting it, what is essential in grouping mental sentence tokens together as being of the same type is some part of their causal network—the way in which the token causally interacts with other states and processes of the organism. If a principle for typing mental sentence

tokens counts a pair of tokens as type identical when (and only when) the tokens have similar patterns of potential causal interaction, then, recalling the terminology of the last chapter, let us call it a *causal* account of typing. If the only relevant potential causal links that a typing scheme recognizes are those obtaining between mental tokens and other mental states, those obtaining between mental tokens and stimuli, and those obtaining between mental tokens and behavior, then let us call it a *narrow causal* account.

Causal accounts of the typing of mental tokens contrast with the purely syntactic shape-similarity account that was invoked in my tale about Otto and his tiny CRT. On a shape-similarity account, Otto has a token of the same type as the one which leads me to say 'It's raining', since we both have mental inscriptions shape-similar to

It's raining.

But Otto's inscription is not in the least *causally* similar to mine. Its causal network is much more closely akin to the one that would be exhibited by a token of 'It's snowing' on my CRT. So on a causal account of typing mental tokens, Otto's inscription, despite its shape, counts as a token of 'It's snowing'. And thus the causal account, along with commonsense intuition, ascribes to Otto the belief that it's snowing.

The second sort of classification scheme for mental tokens that Lycan describes turns not on the causal properties of the tokens, but on their *semantic properties*. "We might count a thing as a 'Broccoli causes erysipelas' just in case the thing has the same truth conditions as does our sentence 'Broccoli causes erysipelas', or if and only if the thing has the same truth conditions computed according to the same recursive procedure."[18] In a similar vein Field suggests that tokens of mental sentences may be classified together if they have the same meaning or content.[19] Just what this criterion of sameness of content comes to is less than obvious. Much of chapter 5 will be devoted to pondering the matter. For present purposes, we need only the roughest of ideas and a label. I will call any scheme for typing mental tokens which relies on their meaning or content or truth conditions a *content account*. Content accounts, like causal accounts, contrast with the shape-similarity story. On the content account, Otto's CRT inscription counts as a token of 'It's snowing' despite its shape and because of its content or truth conditions.

A more interesting question is how causal accounts and content accounts compare with each other. Do they categorize mental tokens differently, or do they inevitably come out with the same categorization? On this issue, opinions divide. According to Fodor the two sorts of classification schemes coincide, "plus or minus a bit." Indeed Fodor

sees this as "*the* basic idea of modern cognitive science."[20] Any thoroughgoing functionalist in the philosophy of mind will also end up on this side of the divide. On the other side, denying that causal and content accounts converge, are Field, Lycan, Perry, McDowell, and the truth.[21]

The distinction between causal and content accounts of mental token typing makes it possible to give a quick and clean statement of three central theses of this book. First, the mental sentence theory of belief, if fleshed out with a narrow causal account of typing, just does not comport with our workaday folk psychological notion of belief—it is not an account of belief, as the term is ordinarily used. Second, the mental sentence theory, if fleshed out with a *content* account of typing, is a good first pass at saying what beliefs are, as they are conceived by folk psychology. But, third and most important, beliefs so conceived have no role to play in a mature cognitive science.