Chapter 4

Some Evidence Against Narrow Causal Accounts of Belief

In this chapter I begin my argument for the claim that accounts of belief which invoke a narrow causal standard of individuation do not characterize a notion of belief plausibly identified with our commonsense notion. My focus throughout the argument will be on the narrow causal version of the mental sentence theory sketched in the previous chapter. With appropriate adjustments of detail, however, an entirely parallel case can be constructed against theory-theory (or functionalist) accounts which are committed to the narrow causal individuation of mental states.[1] Since the case I shall present is complex, I had best begin by sketching a general overview of the argument.

My first task will be to assemble some evidence for the lack of fit between the folk notion of belief and the account of belief offered by the narrow causal version of the mental sentence theory. Basically my theme will be that a narrow causal account dictates judgments about how beliefs are to be characterized and when they count as the same or different, which do not comport with the judgments of folk psychology. There are two sides to this story. Sometimes the problem is that a narrow causal account draws its distinctions too coarsely, because it is incapable of distinguishing between mental states that folk psychology takes to be importantly different. On the other side the problem is reversed. A narrow causal account is sometimes too finicky, forcing distinctions that folk theory does not draw.

The evidence to be marshaled for these conclusions will be evidence about our commonsense intuitions. The use of such intuitions is standard practice in philosophical analysis. Still perhaps a few words are in order on what I take to be the proper use of intuitive data. Intuitions are simply spontaneous judgments. At best intuitions can tell us something about the boundaries or extension of our folk concepts. Thus for example, if we present a subject with pieces of furniture varying along several dimensions and ask for his spontaneous judgments on which ones are couches, we can learn a good deal about what sort of objects do and do not fall under his concept of couch. We might also ask a

subject for his judgments about putative principles of classifications, principles of the form: an object is a couch if (and/or only if) it is F. Intuitions about the principles guiding our judgments on particular cases are notoriously unreliable, however, and I shall generally try to avoid them.[2] What is more, there is no guarantee of accuracy accompanying intuitions about individual instances either. We sometimes are poor judges about what we would say on a particular case, especially when that case is verbally described.[3] And even when we correctly report on what we would say of a particular case, there is no guarantee that this is what we *should* say, that our judgment correctly reflects the extension of our concept.* Despite their fallibility, however, intuitions are often the best and most systematic evidence available in determining the contours of a folk concept. In the absence of an argument that intuitions in some domain are particularly likely to be mistaken or misleading, it would be folly to ignore them.

I would, however, be very uncomfortable indeed to rest my case against narrow causal theories of belief on the sort of intuitive data presented in this chapter. There are two reasons for my reluctance, one general, the other local. The general reason is that the facts about intuitions are negative data, and negative data hardly ever suffice to scuttle a theory. There are endless moves, principled or ad hoc, that a theorist can make to blunt the force of prima facie negative data. Nor are these moves always symptoms of intellectual bad faith.[4] To do in a theory we need another theory, preferably one that does a better job at explaining all the evidence available, including the negative evidence that troubled the old theory. Thus my case against narrow causal theories spills over into the next chapter, where I will develop my own account of the commonsense concept of belief. The local reason for my reluctance to rest my case on intuitive data is that, when it comes to judgments about what hypothetical subjects would properly be said to believe, intuitions are notably labile. It is often possible to evoke conflicting intuitions about a given subject's beliefs simply by framing the question in a somewhat different way. I take this lability as itself a bit of evidence about the folk notion of belief. A virtue of the theory to be developed

*How could it happen that an intuition correctly reporting what we would say of a given case might nonetheless be mistaken about the extension of our concept? The most interesting possibility is that the concept points outside itself, so to speak, in specifying its extension, requiring information about which we are ignorant or misinformed. If we present people with pieces of furniture and ask for their intuitions on which ones count as couches, we are likely to get a good fix on the extension of their couch concept. But if we present people with gemlike stones and ask for their judgments on which are diamonds, we will get a much less accurate indication of the extension of their diamond concept.

in the following chapter is that it explains why intuitions about a given case vary with the setting in which the intuition is evoked.

The evidence to be presented in this chapter falls under three headings. In the section on *Holism* I will present some cases suggesting that our intuitions about how a belief is to be described depend in part on the other beliefs the subject has. This dependency makes for intuitive distinctions that cannot be captured by a narrow causal theory. In the section on *Reference* I will use much the same strategy, this time focusing on intuitive distinctions rooted in the reference of words or concepts. Once again the theme will be that folk psychology draws distinctions more fine grained than those available to narrow causal accounts. Finally, in the section on *Irrelevant Causal Differences*, we will look at the other side of the coin, noting how under suitable circumstances commonsense intuition can be much less demanding than a narrow causal theory, classifying together cases narrow causal accounts must take to be clearly different.

1. *Holism*

Before considering cases, I should elaborate a bit on what I take to be the commitments of narrow causal theories. Recall that the narrow causal version of the mental sentence theory classifies mental sentence tokens as type identical if the tokens have the same (or similar) causal interrelations with other mental states, with stimuli, and with behavior. It is essential to realize that this talk of causal interrelations must be interpreted as pertaining not only to the causal interactions between a given token and other *actual* states or stimuli but also to the interactions that *would* obtain between the token and *possible* further mental states or stimuli. An illustration may serve to make the point. Suppose, as we have been supposing intermittently, that human mental sentences are written on little cerebral CRTs, and that the sentences in the language of thought look, for all the world, like tiny English sentence tokens. Now let us reflect on what is necessary, on the narrow causal account, for an inscription on Otto's CRT to count as type identical with the inscription on my CRT which I express by saying, "All men are mortal." I will assume that the mental token on my CRT looks similar in shape to the one I have just written. Of course on the causal account, looks are irrelevant. Otto may also have a mental token which looks like the one just written, though to count as the belief that all men are mortal it must enter into the same causal interactions. But the same causal interactions with *what*? I believe that Socrates was a man, and that belief (whose object, let us suppose, is a token on my CRT similar in shape to: Socrates was a man) along with my belief that all men are

mortal, once caused me to infer (i.e., add to my CRT) a token of 'Socrates was mortal'. Otto, by contrast, does not believe that Socrates was a man. Perhaps he has never heard of Socrates (and thus no inscriptions containing tokens of 'Socrates' appear on his CRT). Or he might have some quite mistaken belief about Socrates—that he was a deity, perhaps, or a dachshund. Since Otto has no token of 'Socrates was a man' on his CRT, his token of 'All men are mortal' does not interact with a token of 'Socrates was a man' to yield a token of 'Socrates was mortal'. So there is a sense in which the causal interrelations of my 'All men are mortal' token differ from the causal interrelations of Otto's. However, it is alien to the spirit of the causal account of token typing to count *this* difference in reckoning whether Otto's 'All men are mortal' token and mine are of the same type. What matters is not whether Otto's 'All men are mortal' token actually *has* causally interacted with a token of 'Socrates was a man', but rather what *would* happen if Otto were to have such a token on his CRT. To count as type identical with mine, it must be the case that were Otto to have a 'Socrates was a man' it would causally interact with his 'All men are mortal' in the expected way. The causal patterns of interest to narrow causal accounts of typing are not merely those that have obtained among actual states, but also those that would obtain among nonactual though possible states. The essential point is that, for a narrow causal theorist, the type identity of a mental state is determined by its *potential causal interactions* with other mental states, with stimuli, and with behavior. Its type identity does not depend on the other mental states the subject happens to be in at the moment in question. Thus the type identity of a mental sentence token will not depend on the other mental sentence tokens that happen to be in the subject's head at a given moment.

This said, let us turn to cases. What I want to demonstrate is that intuitive judgments about whether a subject's belief can be characterized in a given way and intuitive judgments about whether a pair of subjects have the same belief are often very sensitive not only to the potential causal interactions of the belief(s) in question but also to other beliefs that the subject(s) are assumed to have. The content we ascribe to a belief depends, more or less holistically, on the subject's entire network of related beliefs.

The cleanest case I have been able to devise to illustrate the holism in content ascription turns on the sad fate of people afflicted with progressive loss of memory as the result of the degeneration of brain tissue. Often these people are troubled by the loss of many cognitive faculties, but to make the point I am after, let me ignore these other cognitive problems and describe in a somewhat idealized way the history of Mrs. T, a real person who was employed by my family when I was

a child. As a young woman, around the turn of the century, Mrs. T had an active interest in politics and was well informed on the topic. She was deeply shocked by the assassination of President William McKinley in 1901. In her sixties, when I first knew her, she would often recount the history of the assassination and spell out her analysis of the effects it had had on the politics of the day. As Mrs. T advanced into her seventies, those around her began to notice that, though her reasoning seemed as sharp as ever, her memory was fading. At first she had trouble remembering recent events: who had been elected in the Senate race she had been following; where she had left her knitting. As time went on, more and more of her memory was lost. She could not remember the difference between the Senate and the House, nor the length of the president's term of office. As her affliction got worse, she could no longer remember what the president did, nor could she recall who George Washington was. Some weeks before her death, something like the following dialogue took place:

S: Mrs. T, tell me, what happened to McKinley?
Mrs. T: Oh, McKinley was assassinated.
S: Where is he now?
Mrs. T: I don't know.
S: I mean, is he alive or dead?
Mrs. T: Who?
S: McKinley.
Mrs. T: You know, I just don't remember.
S: What is an assassination?
Mrs. T: I don't know.
S: Does an assassinated person die?
Mrs. T: I used to know all that, but I just don't remember now.
S: Do you remember what dying is?
Mrs. T: No.
S: Can you tell me whether you have died?
Mrs. T: No, I just don't remember what that is.
S: But you do remember what happened to McKinley?
Mrs. T: Oh, yes. He was assassinated.

Neurologists are far from a detailed understanding of what happens to the brains of people like Mrs. T.[5] But to make the point I am after, let us engage in a bit of speculation. Suppose that what afflicted Mrs. T was a more or less pure case of progressive loss of memory, without involving other cognitive faculties. And suppose that it was real memory *loss*, with the memories or beliefs in question simply being erased from the victim's head. The mental sentence theorist can take this erasure metaphor quite literally. Memories and beliefs, on his view, are relations

to sentence tokens, and to lose a memory is to erase the token. For vividness, let us imagine, yet again, that our mental sentences are written on tiny CRTs. Then the effect of Mrs. T's disease was progressively to erase areas of her CRT screen.

Now the question I want to pose for our intuitive judgment is this: Shortly before her death, Mrs. T had lost all memory about what assassination is. She had even forgotten what death itself is. She could, however, regularly respond to the question, "What happened to McKinley?" by saying, "McKinley was assassinated." Did she, at that time, *believe* that McKinley was assassinated? For just about everyone to whom I have posed this question, the overwhelmingly clear intuitive answer is no. One simply cannot believe that McKinley was assassinated if one has no idea what an assassination is, nor any grasp of the difference between life and death.

Consider, now, what the narrow causal version of the mental sentence theory must say about this case. We assume that before her illness Mrs. T had many thousands of sentence tokens on her CRT, including many about McKinley, many about death, many about assassination, and so on. But as the screen is erased, fewer and fewer of these sentences remain. At the time of the dialogue recounted above, almost none of this remains on the screen, apart from the single token of 'McKinley was assassinated'. Still, for a causal theory, the existence of this token entails that Mrs. T does believe that McKinley was assassinated, for the token has lost none of its former causal potential. We have assumed that only Mrs. T's memory is affected, and thus, ex hypothesis, were her belief screen to be restocked as it was in her prime, the isolated 'McKinley was assassinated' token would interact with these (possible) tokens in just the right way. Since in causal potential her 'McKinley was assassinated' token is quite similar to mine, and since mine is the belief that I express by saying 'McKinley was assassinated', it follows that Mrs. T's token is of the same type as mine and that she too believes that McKinley was assassinated. Plainly, here, we have a folk psychological distinction that the narrow causal version of the mental sentence theory just does not capture. On the commonsense view, before her illness Mrs. T believed that McKinley was assassinated. After her affliction had become quite severe, she no longer believed it. But if the causal potential of her few remaining mental sentence tokens has not changed, a causal account must view them as the same beliefs they have always been. Causal accounts do not reflect the holism of belief ascription.

Intuitions reflecting the holism of belief ascription are sharpest in cases like Mrs. T's, where the subject's belief tokens are a radically reduced subset of our own. However, kindred intuitions can be evoked

in cases involving subjects with a much richer set of beliefs. Consider the following case. Let us imagine that some of our descendants have developed a rich and fruitful scientific theory, along with a new set of technical concepts and terms to express them. We may suppose that all of this theory is inscribed on the CRTs of these future scientists. As an aid to the imagination, we can continue the fiction that the language in which their beliefs are inscribed is English, though of course it will not be present day English but some extension of that language enriched with new terms to represent the new scientific concepts. The narrow causal version of the mental sentence view has no problem in conceding that these future scientists have beliefs that we don't, though it would add that we cannot ascribe any content to these beliefs since our language lacks suitable content sentences. Now let us suppose that via some strange set of circumstances which can here go unspecified, one of our contemporaries, Paul by name, comes to have imprinted on his belief screen a token of one single sentence which, in the heads of our scientifically sophisticated descendants, represents some deep strand of their new doctrine; let it be the mental sentence they express by saying, "Hydrogen atoms are single-petaled superheterodyning negentropy flowers."[6] In saying that Paul's belief is a token of the same mental sentence whose tokens appear in the heads of our future scientists, I am invoking the standard of the narrow causal account. Paul's token has the same causal potential as the tokens in future heads. Of course the token in Paul's head is largely inert, for lack of further theoretical sentences to interact with. But no matter. On the causal account Paul and the future scientists share the same belief. Here again commonsense intuition surely disagrees. From the fact that Paul has this odd, isolated sentence in his head, even if, as we may suppose, it sometimes leads him to *say*, "Hydrogen atoms are single-petaled superheterodyning negentropy flowers," it does not follow that Paul and the fabled scientists share the same belief. Indeed from the perspective of folk psychology, Paul's isolated mental token hardly counts as a belief at all. On a narrow causal account of belief, these folk intuitions are simply an unexplained anomaly.

Though I have told the tale about Paul in rather fanciful terms, other, more familiar situations prompt parallel intuitions. A majority of literate Americans over the age of seven have a belief which they express by affirming "$E = mc^2$." And it is not implausible to suppose that, by the standards of the causal account, this belief is identical to the one that a sophisticated physicist expresses with the same sentence. For scientifically unsophisticated people, however, the belief underlying their affirmation of '$E = mc^2$' is a largely isolated one. In reflecting on these cases there is a substantial intuitive pull in the direction of denying

that the scientist and the man in the street are expressing the same belief. When my eight-year-old son asserts that $E = mc^2$, I am strongly inclined to think that the belief underlying his assertion is radically different from the one which he will express with the same sentence if he retains his avid interest in science for a decade or two. But the changes that will make the child's belief intuitively type identical with those of the scientist need not involve significant changes in the causal potential of the underlying mental state. Rather, what is needed is the addition of a rich set of further beliefs in which his current belief will be embedded.

An interesting variant on these cases arises when we consider our intuitions about beliefs we share with people whose belief systems elsewhere diverge from our own. Consider, again, the future scientists. It is plausible to suppose that they will have mental tokens which are, by the standards of the causal account, type identical to the mental token I express by saying, 'The Hindenburg exploded because it was filled with hydrogen'. But we have also supposed that many of their beliefs about hydrogen are radically different from mine. Question: Is the belief they express by saying, 'The Hindenburg exploded because it was filled with hydrogen' the same as the one I express with the same words? I find that to most ears not previously contaminated by philosophical theory, the question has a distinct "yes and no" feel to it.[7] It is sort of the same belief but also sort of not. This ambiguous intuition is yet another anomaly for a causal account, since unless we suppose that the two beliefs differ in causal potential, that account straightforwardly entails that the two beliefs are the same.

This is a convenient point for a brief detour to caution against an all too common confusion. When I discuss cases like the ones we have been considering, philosophers often ask just what, on my view, is the content of the child's belief, or the future scientist's belief, or Mrs. T's belief. My reply, in all these cases, is that the question invokes an undefined notion, and, absent a definition, it is senseless. Let me elaborate. On the mental sentence view as I have portrayed it and also on my own view to be developed in the next chapter, no mention need be made of an entity called a *content*, which beliefs can have (or lack). There are of course content *sentences*, which are simply sentences embedded in belief sentences. On the mental sentence view, there are also *objects* of belief, which are sentences in the mental code or language of thought. Finally, we *do* something that might be described as "ascribing content" to a belief; that is, we attribute a belief to a person by using a belief sentence containing a content sentence, or we talk about a belief using related locutions also containing embedded content sentences (e.g., 'Otto's belief that snow is white'). But in all of this

there is no need to suppose that there are things, contents, which beliefs have. *To ascribe content to a belief is simply to describe it in a certain way.* I do not wish to deny that some reasonable sense might be *given* to the notion of there being a thing which is a belief's content. For example, it might be urged that sentences in the language of thought express propositions, and these propositions might be stipulated to be the content of the belief which has the associated mental sentence as its object. But, and this is the essential point, any such move requires some terminological stipulation. There is no preexisting notion of a content-entity to be found in folk psychology. Talk of there being some thing which is "the content" of a belief is a theorist's term of art, often used though rarely explained. What is more, it is a term of art which finds no natural place in mental sentence theories, unless the theorist wants to pursue the proposition gambit. And most don't.

So what is to be said about cases like the future scientists' belief? Well, if the mental sentence theory is correct, then it has as its object some sentence in the language of thought. There is also presumably some sentence in the language of the scientists which they use to express this belief. But there is no sentence in *our* language that is used to express that belief. Thus we cannot ascribe content to it; we cannot attribute this belief by uttering a sentence of the form: 'S believes that p', for the unexciting reason that we have no suitable p.* If the belief has truth conditions, and there is no reason to insist that it doesn't, then these are not specifiable in contemporary English either. But does the belief have a content, and if so what is it? My answer here is that pending some terminological stipulation, the question makes no sense.

To return from the detour, what I have been arguing is that intuitions about sameness of belief and about the appropriateness of describing a belief as the belief that p do not match up with the judgments mandated by the causal version of the mental sentence theory. My diagnosis of the mismatch is that, in all these cases, our intuition seems to be taking into account not only the causal potential of the mental sentence token but also the network of further beliefs that the subject has or lacks at the time in question. Sometimes, as in the case of Mrs. T, intuition seems to give a clear answer which is just the opposite of what a causal account requires. But it is interesting to note that sometimes intuition renders a less settled judgment. It is not clearly wrong to

*If we know the sentence scientists use to express the belief, we can use *that* sentence as a content sentence in a belief ascription. But this is a puzzling move, since we are assuming that we do not know what the content sentence means! Still, we often do invoke the words of the subject in ascribing a belief to that subject, even though we don't understand the words. (Cf. 'Heidegger believed that nothing noths'.) The theory developed in chapter 5 will have an explanation of why this sometimes works.

describe a certain belief as the belief that p, but it is not clearly right either. These yes-and-no intuitions play a prominent role in the following section, and any account of the folk notion underlying our intuition should explain their prominence. Cases evoking conflicting intuitions count as evidence against the narrow causal account not because intuition and theory are in direct conflict, but because intuition pulls in both directions, while the causal account pulls in only one.

2. *Reference*

The common strand in cases considered under this heading is that they all involve some aspect of reference. My theme will be that our intuitions about the appropriate description of a belief and about the sameness or difference of beliefs are affected by the reference of the terms that a subject would use to express the belief. Since reference is generally not fixed by the pattern of potential causal interconnections among a subject's beliefs, other mental states, stimuli, and behavior, however, it often happens that the narrow causal version of the mental sentence theory dictates judgments about a belief that conflict with our intuition. For convenience, I will group the cases to be considered under three headings: Proper Names, Kind Terms, and Indexicals.

Proper Names

A traditional view about proper names holds that their denotation is determined by the beliefs the speaker would express using the name, or by the sentences he would accept in which the name occurs. During the last fifteen years a new view has come to be widely accepted, however, growing out of the work of Donnellan, Kaplan, Kripke, Putnam, and others.[8] Stripped to its barest essentials, this new "causal" account holds that a name denotes a person on a given occasion of use if there is a suitable causal chain linking the person and the use of the name. I am inclined to think that the causal story is only part of the truth about the denotation of proper names. For present purposes, however, I propose to stay as far as possible from tangled issues in the philosophy of language. The story I want to tell focuses on beliefs and what we would find it intuitively natural to say about them.

Consider the following example involving the beliefs of two subjects, Tom and Dick. Tom is a contemporary of ours, a young man with little interest in politics or history. From time to time he has heard bits of information about Dwight David Eisenhower. We can assume that most of what Tom has heard is true, though there is no need to insist that all of it is. Let us also assume that each time Tom heard something about Eisenhower, Eisenhower was referred to as 'Ike'. Tom knows

that this must be a nickname of some sort, but he has no idea what the man's full name might be and doesn't very much care. Being little interested in such matters, Tom remembers only a fraction of what he has heard about Ike: that he was both a military man and a political figure; that he played golf a lot; that he is no longer alive; that he had a penchant for malapropisms; and perhaps another half dozen facts. He has no memory of when or where he heard these facts, nor from whom. The mental sentence theorist will hold that all these beliefs are stored sententially in the appropriate place in Tom's brain. For vividness, I will assume that the appropriate place is a tiny CRT, and that the sentences are stored in English.

To tell you about Dick, I must indulge in a bit of fiction, though I suspect that readers with a richer knowledge of history could produce a more realistic example. Dick, in my story, is a young man in Victorian England. Like Tom, he is bored by politics and history. Dick has heard some anecdotes about a certain Victorian public figure, Reginald Angell-James, who, for reasons that history does not record, was generally called 'Ike'. And (the plot thickens) in all the stories that Dick has heard about Angell-James, the gentleman was referred to as 'Ike'. Angell-James and Eisenhower led very different careers in different places and times. However, there were some similarities between the two men. In particular, both were involved in politics and the military, both liked to play golf, and both had a penchant for malapropisms. Moreover, by a quirk of fate, it happens that the few facts Dick remembers about Angell-James coincide with the few facts Tom remembers about Eisenhower. What is more, of course, Dick would report these facts using the very same sentences that Tom would use, since the only name Dick knows for Angell-James is 'Ike'. Indeed, we can suppose that the only mental sentences about Angell-James to be found in Dick's head are, on the narrow causal standard, exact duplicates of the sentences about Eisenhower to be found in Tom's head. That is, each of Tom's sentences about Eisenhower can be paired with one of Dick's sentences about Angell-James, and the sentences so paired are identical in point of their causal potential. (It will do no harm to suppose that the sentences on their CRTs look the same too.)

Now let me try to evoke some intuitions about these cases. Suppose that one fine day in 1880 one of Dick's friends asks him what he knows about Ike. Dick replies, "He was some kind of politician who played golf a lot." A century later, one of Tom's friends asks him an identically worded question, and Tom gives an identically worded reply. First intuition probe: Were Tom and Dick expressing the same belief? Most people on whom I have tried this case are initially inclined to say no. They buttress this view by noting that Tom's belief was about Eisen-

hower, while Dick's was about Angell-James. This intuition tends to be even stronger if we alter the story by adding that, in fact, Angell-James did not play golf at all, though it was widely believed that he did. On this version, Dick's belief is false, while Tom's is true. On reflection, however, some people concede that they can "sort of see" why someone might want to say that Tom and Dick held the same belief, though they insist that it is odd or misleading. (In fairness, I should report that I have found one informant—a noted psychologist—who claims his intuitions run just the other way.) When asked how they would describe what Tom and Dick believe, many people give an interesting response. Tom, they say, believes that Eisenhower was a politician who played golf a lot, whereas Dick believes that Angell-James was a politician who played golf a lot.

Plainly these intuitions pose a problem for the narrow causal version of the mental sentence view. For on that view Tom's belief and Dick's ought to count as identical, since Tom's mental sentence and Dick's have identical causal potential. On the narrow causal account there is no explanation of the strong intuitive pull toward counting Tom's belief and Dick's as different.

Three observations about this case will prove useful when it comes time to build a better theory. First, the case is quite distinct from the holism cases discussed above. The network of beliefs about Eisenhower in which Tom's belief is embedded is entirely parallel to the network of beliefs in which Dick's belief about Angell-James is embedded. Indeed, by the narrow causal standard these networks are identical.* So the intuitions in this case cannot be attributed to the effects of related beliefs. Second, there is a striking parallel between our intuitions about our subjects' beliefs and our intuitions about the denotations of the names they use to express their beliefs. Both Tom and Dick might utter 'Ike was a politician'. However, there is a strong intuitive pull toward saying that their utterances of 'Ike' *refer* to different men; when Tom uses 'Ike' he is referring to Eisenhower, while when Dick uses 'Ike' he is referring to Angell-James. Were the story retold in such a way that intuition would find their utterances of 'Ike' referred to the same man, then the intuition that they had different beliefs would dissolve. Third,

*In the philosophical literature there is a quaint tradition of making this point by imagining a separate planet, Twin Earth, in some far off corner of the universe, on which are to be found doppelgangers, molecule for molecule replicas of people on earth. Having frequently used such examples myself, I offer a pair of sociological observations. First, nonphilosophers often find such cases so outlandish that they have no clear intuitions about them. Second, Twin Earth examples drive psychologists up the wall, reinforcing the widespread conviction that the concerns of analytic philosophy are frivolous. To avoid these problems, I will refrain from using Twin Earth examples whenever possible.

it is at least prima facie problematic for the narrow causal account that we find it so natural to describe Tom's belief as "the belief that Eisenhower was a politician." For presumably you and I have mental sentences containing tokens of both 'Ike' and 'Eisenhower', and these tokens are causally distinct. Tom, by contrast, has only 'Ike' tokens in his beliefs about Ike. And if we were to assert, "Tom believes that Eisenhower was a politician," Tom would deny it. The theory in chapter 5 provides a framework for explaining why we nonetheless feel comfortable describing Tom's belief as we do.

Kind Terms

My second case unfolds against the background of a curious difference between American and British English. The facts, as I understand them, are as follows. In American supermarkets one can buy a number of green salad vegetables, including a rather bitter curly leafed variety called 'chicory' and a smooth leafed, tightly packed conical variety called 'endive'. What appear to be exactly the same vegetables are available at British greengrocers; the labels are reversed, however. What Americans know as 'chicory', Englishmen call 'endive', and vice versa.*

The beliefs I want to focus on are those of a pair of subjects, one American, the other English. John, the American, is a "meat and potatoes man" with no taste for vegetables or salad. He has heard mention of salad greens called 'chicory' and 'endive' though he cannot remember ever seeing or tasting either of them. Nor could he tell which was which. However, he has heard, and remembers, that chicory can be rather bitter. The action in our story takes place on a day when John is a guest for dinner at the home of friends. After the main course the hostess asks John whether he would like to try some chicory salad. Tact not being John's strong suit, he declines, saying, "No thanks, chicory is bitter." The other protagonist in my tale is Robin, an Englishman. Robin shares John's dislike of vegetables and his limited knowledge of them. Like John, Robin has heard of the existence of salad greens called 'chicory' and 'endive', but he has never seen either. Robin has heard some misinformed fellow Englishman say, "Chicory is often rather bitter," and he believed it. By now the rest of the story will be obvious. Robin, too, is invited to a dinner party (this one in England) and, on being offered "a chicory salad," declines saying, "No thanks, chicory is bitter."

*Since first writing this, a number of cosmopolitan friends have assured me that I am oversimplifying. The accounts they have given of trans-Atlantic diversity in salad terms are considerably more complex and patently inconsistent with one another. No matter. For the purposes of my example, please simply *assume* that the facts are as I have stated them.

To evoke the intuitions relevant to this case, reflect on the following two questions. First, was the belief John was expressing when he said "Chicory is bitter" the same as the belief that Robin was expressing when he uttered the same words? Second, how are John's belief and Robin's best described? Most of the people on whom I have tried this case report conflicting intuitions, with a strong pull in the direction of saying that John and Robin do not have the same belief, and a somewhat weaker pull in the direction of saying that they do. American informants are inclined to think that "John believes that chicory is bitter" would be the natural way of saying what John believes. However, American informants tend to be more cagy about Robin. "What he believes," one student told me, "is that the stuff he calls 'chicory' is bitter." Many feel it is natural to say that Robin believes *endive* is bitter, though adding a gloss to the effect that he calls endive 'chicory'. Interestingly, though, just about everyone agrees that if they had to say in Chinese what Robin believes, they would use the Chinese word for endive, not the Chinese word for chicory. After reflecting on how they would describe Robin's belief in Chinese, most people I've asked are considerably less reluctant to describe Robin's belief as "the belief that endive is bitter."

As in the proper name case, these intuitions pose problems for the narrow causal version of the mental sentence theory. As I have told the story, it would be plausible for the mental sentence theorist to suppose that John and Robin have type identical sentences in their store of beliefs about endive and chicory. More specifically, the narrow causal account would classify the mental sentence that John expresses with 'Chicory is bitter' as type identical with the mental sentence that Robin expresses with the same words. The intuitive inclination to say that these are different beliefs is left unexplained. Note also that reference seems to be implicated in our intuitions, since when uttered by John 'chicory' refers to chicory (the curly, bitter stuff), while when uttered by Robin, 'chicory' refers to endive (the smooth leafed, conical shaped one).*

Indexicals

The beliefs that a person would express with a sentence containing an indexical like 'I', 'here', or 'now' pose what is perhaps the most obvious problem for narrow causal theories. For in these cases we need no

*This example is inspired by those of Putnam (1975) and Burge (1979), though it avoids difficulties that trouble some of their examples. See, in particular, the discussion in Fodor (1982). In focusing directly on intuitions about sameness of belief and about how beliefs are appropriately described, I think the intricate puzzles Fodor raises can simply be sidestepped.

contrived examples to see that common sense sometimes invokes a standard of sameness of belief different from the one urged by the causal account. If two sincere and confident presidential candidates each asserts, "I will be the next president of the United States," there is a strong intuitive inclination to say that they have different beliefs. Indeed given the facts about American presidential elections, their beliefs are incompatible; they cannot both be true. But there is also a discernible tug in the other direction. If two lottery ticket holders both show up at the pay-off window each thinking that he has won the lottery, it would be natural enough to explain their behavior by saying that they have the same belief. If, as seems inevitable, the mental sentence view will include indexicals in the language of thought, then presumably each presidential candidate has a token of 'I will be the next president of the United States' in his brain. And since these tokens will be causal isomorphs of each other, the narrow causal theory cannot account for that strand in our intuition which insists that the two candidates have different beliefs. As in our two previous cases, reference seems clearly implicated, since 'I' in the mouth of one candidate has a different reference from 'I' in the mouth of the other.

Cases involving indexicals also pose some difficulty for the Fodor-style account of how content sentences are related to mental sentences.[9] Recall that for Fodor the mental sentence being ascribed to the believer is identical to the one that the speaker would ordinarily express by using the content sentence. But now consider the case proposed by Perry,[10] in which I see some unfortunate fellow attacked by a bear. Following the recommended defensive strategy, the victim rolls up like a ball and remains motionless. In explaining why he does this, it would be natural to begin by saying, "He believes that he is being attacked by a bear." But the belief that *I* would express by saying, "He is being attacked by a bear" can hardly be identical to the one the victim has, *by the standard of the narrow causal account.* For this belief causes me to run and seek help, not to roll up like a ball. The belief which would lead me to behave as the victim does is the one which I would express by saying, "*I* am being attacked by a bear." When I am the observer and he is the victim, however, I obviously cannot attribute the correct belief by saying, "He believes that I am being attacked by a bear."

One other curiosity is worth noting about the way in which self-referential beliefs are ordinarily described. In the case of the two presidential candidates, if one is named 'Smith' and the other is named 'Jones', it would be natural enough to describe the situation by saying, "Smith believes that Smith will be elected, while Jones believes that Jones will be elected." It is possible, though, to embellish the story in a way which makes this a much less natural description. If, for example,

Smith does not know that he is named 'Smith', and if this fact looms large in the tale, then the intuitive acceptability of 'Smith believes that Smith will be elected' declines.[11]

3. *Irrelevant Causal Differences*

Under the headings of Holism and Reference we have looked at cases where the discriminations drawn by common sense were finer than any that would be available on the narrow causal version of the mental sentence theory. In this section I want to look at a few cases where the situation is reversed, the narrow causal account draws distinctions which our intuition ignores.

Consider first the beliefs of people who are afflicted by some minor perceptual handicap like color blindness. The potential causal pathways leading from stimulus to belief in these people differ in a clear way from the pathways available to normal subjects. If, for example, a normal subject is shown a copy of one common color blindness test, he will see (and thus come to believe) that there is a figure 8 on the sheet. A color-blind subject will not see the figure 8 and will not come to believe that there is a figure 8 on the sheet. It follows from the narrow causal account that the belief a normal subject expresses by saying, "There is a figure 8 on this sheet" differs in type from the belief that a color-blind subject would express using the same words. But in most circumstances common sense recognizes no such distinction. If a pair of subjects, one color-blind and the other normal, are both *told* that a certain sheet of paper (which neither of them can see) has a figure 8 on it, and if both subjects trust their informant, it would be intuitively bizarre to insist that they must have formed different beliefs.

It might be thought that in the color blindness case our intuition is being guided by the fact that the difference in potential causal paths leading to the two beliefs is relatively minor. Perhaps the causal account can be saved from embarrassment in cases like this if it counts beliefs as type identical when they are *approximately* the same in their potential causal interactions. I think there is an important insight lurking in this suggestion. But as it stands it clearly will not do, since common sense will often count a pair of beliefs as the same even though their difference in causal potential is enormous. Consider, for example, what we are inclined to say about the beliefs of people who are totally blind. Suppose that I tell both a blind person and a sighted person that there is a cat in the next room, and they both believe my report. It seems intuitively natural to say that they both come to have the same belief—the belief that there is a cat in the next room. Our intuitions remain the same if we change the example by replacing the (merely) blind subject with a

person like Helen Keller whose perceptual deficiencies are staggering. Under most circumstances we would find it intuitively natural to say that if Ms. Keller and a normal person are both told there is a cat in the next room, and if they both accept the report as accurate, then they both come to believe the same thing. Yet here, surely, the narrow causal theory is in trouble, since the difference between the potential causal pathways leading to Ms. Keller's belief and those leading to the normal subject's belief are vast by any measure.

I have been writing as though our intuitions about the beliefs of perceptually handicapped subjects were unequivocal and quite stable. But this oversimplifies the actual situation. Many people are distinctly uneasy with the claim that a blind subject forms the same belief we do in response to verbal reports, particularly when the visual deficiencies of the subject become salient to the situation. Suppose, for example, that I tell both a sighted subject and a congenitally blind subject that in the next room there is an auburn cat with light blue eyes and a dark brown tail. Assuming that both subjects take my report to be accurate, do they both form the same belief? Intuition wavers.

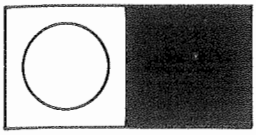
I noted earlier that intuitions about how beliefs are to be described and whether they count as the same or different are often very labile. Very similar cases may provoke quite different intuitions depending on the context in which the question about the subject's belief arises. Unlike the theory developed in chapter 5, the narrow causal account has no explanation for this phenomenon. Let me illustrate the phenomenon with a pair of examples focusing on the beliefs of a color-blind subject. People suffering from one sort of color blindness report that red and green look very similar to them, and when asked to characterize the color of objects they are looking at, they often perform very poorly. However, color-blind people learn to use other visual cues to make essential distinctions—the red traffic light is the one on the top, the green one is the one on the bottom. As a result, color blindness often goes undetected. Indeed, the very existence of color blindness was unknown until the late eighteenth century, when it was discovered by Dalton.[12] Now imagine the case of Peter, a store clerk, who suffers unknowingly from severe red-green color blindness. One day, after the Christmas sales are over, Peter and a fellow worker are asked to take down the Christmas decorations and box them, with a separate box for each color. Peter's fellow worker, Greg, notes with some concern that Peter often puts red decorations in the green box and vice versa. Finally, after watching Peter take down a bright green Christmas ball, examine it carefully, and put it into the red box, Greg says, "Peter, you can't really believe that ball is red. There must be something wrong with your eyes." Here, I think, we are inclined to agree that Greg's

characterization of Peter's cognitive state is quite appropriate. It would seem more than a bit intuitively odd to explain Peter's action by saying he believed the ball was red and leaving it at that. Contrast this case with the following. Peter has just been hired as a night attendant at a chemical factory. On the first day of work he is given very simple instructions. In addition to sweeping the floors, he has only one duty. If the alarm bell rings it means that something has gone wrong with the automatic equipment and there might be an explosion. Should this happen, he is to go into the control room and throw the red lever. There are many other levers of different colors in that room, and these he is not to touch. As it happens, during his first night on the job the alarm bell sounds, and Peter rushes into the control room. Unfortunately one of the engineers has left a pile of papers on top of the red lever, completely obscuring it from view. There is, however, a large *green* lever in the center of the control panel. Peter rushes to throw this lever, with disastrous results. Why did Peter throw this lever? Here, I think, it would be entirely natural to say that he did it because he believed the lever was red.

Thus far in this section we have been attending to cases in which the possible *perceptual* causes of a belief are altered. Analogous intuitions can be evoked if we look, instead, at the patterns of *inference* that may lead to and from a given belief. An intriguing example can be built around the beliefs of subjects participating in Wason and Johnson-Laird's so-called selection task experiment.[13] In one version of the experiment subjects are shown four cards like those in figure 1. Half of each card is masked. Subjects are asked to look at the cards and to decide which masks it is essential to remove in order to know whether the following claim is true about these cards:

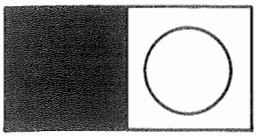
If there is a circle on the left, there is a circle on the right.

Since the task is so straightforward and the directions so simple, it is hard to imagine that subjects don't all end up having the same beliefs about what they are being asked to do. Yet, surprisingly, though they start with the same beliefs about the setup and the task, subjects infer very different answers. A relatively small number conclude (correctly) that the masks on (a) and (d) must be removed. Many more subjects come to believe that the masks on (a) and (c) must be removed; still others conclude that only the mask on (a) must be removed. Moreover, subjects who end up with a mistaken belief often defend their conclusion with a perverse vigor. For the narrow causal theory these subjects are something of an embarrassment. Both the subjects who get the right answer and those who get the wrong answer start with what common sense classes as the same beliefs about the problem. But since they



(a)

(b)



(c)

(d)

Figure 1

draw different inferences from them, the narrow causal version of the mental sentence theory must classify these initial beliefs as different in type.

Here, as in the analogous case of minimal perceptual difference, it might be objected that I am construing the narrow causal standard too strictly, since the difference between the subjects who get the right answer and those who get the wrong answer marks only a very minor difference in the causal potential of their beliefs. And as in the perceptual case, I think there is some merit in the protest. For if we consider cases in which the inferential patterns exhibited by a subject's beliefs grow increasingly different from our own, it becomes increasingly intuitively uncomfortable to describe their beliefs with the same content sentences we use in describing our own beliefs. For a particularly clear illustration of this phenomenon, let me again indulge in a bit of science fiction. Suppose, as we did earlier, that beliefs are stored on a tiny cerebral CRT and that various mental processes, including inference and practical

reasoning, depend on what is written on the CRT. Let us imagine that for each of these cognitive processes there is a separate device which scans the CRT and makes appropriate modifications to the store of beliefs, the store of desires, or whatever. Now what I want to consider is a sequence of cases, each involving a subject who suffers a sudden breakdown in his inference-making device. Let us call our subjects Dave[1], Dave[2], . . . , Dave[9], and let us imagine that as we proceed down the list of Daves the breakdowns in the inference-making device get more severe. For example, we might suppose that Dave[1] simply loses his capacity to infer via transitivity of the conditional. So, though he may have tokens of both

If my wife works late, then my wife does not feed the dog.

and

If my wife does not feed the dog, then I must feed the dog.

in his belief store, he does not infer that he must feed the dog if his wife works late. Dave[2], let us suppose, shares Dave[1]'s deficiency and adds to it an incapacity to reason normally with disjunctions. Though he has a token of the sentence

Either the car keys are on my desk or the car keys are in the kitchen

on his CRT, adding

The car keys are not on my desk

does not lead him to infer

The car keys are in the kitchen.

Dave[3] adds yet another quirk, and Dave[4] still another. However, in each case it is assumed that the inferential breakdown is quite sudden and does not alter the set of sentences stored on these men's CRTs. What are we to make of the beliefs of these subjects immediately after their breakdowns? The answer, I think, is that as we proceed down the list it becomes increasingly unnatural to characterize a given belief the way we would have before the breakdown. Suppose, for example, that along with his other problems Dave[8] has lost his ability to infer by modus ponens. Though his CRT continues to display a token of

If it's raining, then I should take my umbrella

and though he continues to affirm this sentence when asked, this belief no longer leads him to infer

I should take my umbrella

when a token of

It's raining

appears on his CRT. Does he still believe that if it's raining then he should take his umbrella? My intuition here is strongly inclined to give a negative verdict.

It appears that real mental breakdowns are never quite so neatly restricted to a single cognitive capacity. But inferential peculiarities are often part of the clinical profile. One sort of schizophrenic typically infers from

Napoleon was a great leader

and

I am a great leader

to

I am Napoleon.

Does a person who has come to affirm 'I am Napoleon' as a result of this pattern of inference really believe that he is Napoleon? To my intuition, the answer is neither clearly yes nor clearly no.

The example of inferential breakdown seems to suggest that our intuitions are guided by the degree of causal similarity between a subject's belief and our own. And this is not entirely uncongenial to the narrow causal version of the mental sentence theory, though advocates of that account generally pay little attention to degrees of similarity. But there are further complications to be confronted. A number of writers have noted that it is often overwhelmingly intuitively plausible to ascribe beliefs to animals using some of the very same content sentences we would use in ascribing beliefs to quite normal people.[14] There is, for example, no intuitive awkwardness in Russell's story of the Christmas turkey rushing toward its executioner with the mistaken but inductively well-supported belief that it was about to be fed. But the turkey is a singularly stupid bird, and it is likely that despite Dave^{8}'s enormous cognitive deficits, normal human inferential capacity is closer to his than to the turkey's. The narrow causal account has no explanation of why we should be more ambivalent about Dave^{8} than we are about the turkey.

Before bringing this chapter to a close, I should remind you of its relatively modest aim. What I hope to have established with these examples is that there are many different cases in which the judgments of intuition are not what would be expected if we were relying on a commonsense notion of belief which could be analyzed along the lines

of the mental sentence theory with tokens typed on the narrow causal standard. None of my examples suffices to show that narrow causal accounts of belief are beyond prudent patching. But collectively perhaps they are enough to convince you that the fishing might be better somewhere else.