1

Introduction: The Persistence of the Attitudes

A Midsummer Night's Dream, act 3, scene 2.
Enter Demetrius and Hermia.

Dem. O, why rebuke you him that loves you so?
Lay breath so bitter on your bitter foe.

Herm. Now I but chide, but I should use thee worse;
For thou, I fear, hast given me cause to curse.
If thou hast slain Lysander in his sleep,
Being o'er shoes in blood, plunge in the deep,
And kill me too.
The sun was not so true unto the day
As he to me: would he have stol'n away
From sleeping Hermia? I'll believe as soon
This whole earth may be bor'd; and that the moon
May through the centre creep, and so displease
Her brother's noontide with the antipodes.
It cannot be but thou hast murder'd him;
So should a murderer look; so dead, so grim.

Very nice. And also very *plausible*; a convincing (though informal) piece of implicit, nondemonstrative, theoretical inference.

Here, leaving out a lot of lemmas, is how the inference must have gone: Hermia has reason to believe herself beloved of Lysander. (Lysander has told her that he loves her—repeatedly and in elegant iambics—and inferences from how people say they feel to how they do feel are reliable, ceteris paribus.) But if Lysander does indeed love Hermia, then, a fortiori, Lysander wishes Hermia well. But if Lysander wishes Hermia well, then Lysander does not voluntarily desert Hermia at night in a darkling wood. (There may be lions. "There is not a more fearful wild-fowl than your lion living.") But Hermia was, in fact, so deserted by Lysander. Therefore not voluntarily. Therefore *in*voluntarily. Therefore it is plausible that Lysander has come to harm. At whose hands? Plausibly at Demetrius's hands. For Demet-

rius is Lysander's rival for the love of Hermia, and the presumption is that rivals in love do *not* wish one another well. Specifically, Hermia believes that Demetrius believes that a live Lysander is an impediment to the success of his (Demetrius's) wooing of her (Hermia). Moreover, Hermia believes (correctly) that if x wants that P, and x believes that not-P unless Q, and x believes that x can bring it about that Q, then (ceteris paribus) x tries to bring it about that Q. Moreover, Hermia believes (again correctly) that, by and large, people succeed in bringing about what they try to bring about. *So*: Knowing and believing all this, Hermia infers that perhaps Demetrius has killed Lysander. And we, the audience, who know what Hermia knows and believes and who share, more or less, her views about the psychology of lovers and rivals, understand how she has come to draw this inference. We sympathize.

In fact, Hermia has it all wrong. Demetrius is innocent and Lysander lives. The intricate theory that connects beliefs, desires, and actions—the implicit theory that Hermia relies on to make sense of what Lysander did and what Demetrius may have done; and that *we* rely on to make sense of Hermia's inferring what she does; and that Shakespeare relies on to predict and manipulate our sympathies (*'deconstruction' my foot*, by the way)—this theory makes no provision for nocturnal interventions by mischievous fairies. Unbeknownst to Hermia, a peripatetic sprite has sprung the ceteris paribus clause and made her plausible inference go awry. "Reason and love keep little company together now-a-days: the more the pity that some honest neighbours will not make them friends."

Granting, however, that the theory fails from time to time—and not just when fairies intervene—I nevertheless want to emphasize *(1) how often it goes right, (2) how deep it is, and (3) how much we do depend upon it.* Commonsense belief/desire psychology has recently come under a lot of philosophical pressure, and it's possible to doubt whether it can be saved in face of the sorts of problems that its critics have raised. There is, however, a prior question: whether it's worth the effort of trying to save it. That's the issue I propose to start with.

1. How Often It Works

Hermia got it wrong; her lover was less constant than she had supposed. Applications of commonsense psychology mediate our relations with one another, and when its predictions fail these relations break down. The resulting disarray is likely to happen in public and to be highly noticeable.

Herm. Since night you lov'd me; yet since night you left me;
Why, then, you left me,—O, the gods forbid!—
In earnest, shall I say?

Lys. Ay, by my life;
And never did desire to see thee more.
Therefore be out of hope. . . .

This sort of thing makes excellent theater; the *successes* of commonsense psychology, by contrast, are ubiquitous and—for that very reason—practically invisible.

Commonsense psychology works so well it disappears. It's like those mythical Rolls Royce cars whose engines are sealed when they leave the factory; only it's better because it isn't mythical. Someone I don't know phones me at my office in New York from—as it might be—Arizona. 'Would you like to lecture here next Tuesday?' are the words that he utters. 'Yes, thank you. I'll be at your airport on the 3 p.m. flight' are the words that I reply. That's *all* that happens, but it's more than enough; the rest of the burden of predicting behavior—of bridging the gap between utterances and actions—is routinely taken up by theory. And the theory works so well that several days later (or weeks later, or months later, or years later; you can vary the example to taste) and several thousand miles away, there I am at the airport, and there he is to meet me. Or if I *don't* turn up, it's less likely that the theory has failed than that something went wrong with the airline. It's not possible to say, in quantitative terms, just how successfully commonsense psychology allows us to coordinate our behaviors. But I have the impression that we manage pretty well with one another; often rather better than we cope with less complex machines.

The point—to repeat—is that the theory from which we get this extraordinary predictive power is just good old commonsense belief/desire psychology. That's what tells us, for example, how to infer people's intentions from the sounds they make (if someone utters the form of words ('I'll be at your airport on the 3 p.m. flight,' then, ceteris paribus, he intends to be at your airport on the 3 p.m. flight) and how to infer people's behavior from their intentions (if someone intends to be at your airport on the 3 p.m. flight, then, ceteris paribus, he will produce behavior of a sort which will eventuate in his arriving at that place at that time, barring mechanical failures and acts of God). And all this works not just with people whose psychology you know intimately: your closest friends, say, or the spouse of your bosom. It works with *absolute strangers;* people you wouldn't know if you bumped into them. And it works not just in laboratory conditions—where you can control the interacting variables—but also, in-

deed preeminently, in field conditions where all you know about the sources of variance is what commonsense psychology tells you about them. Remarkable. If we could do that well with predicting the weather, no one would ever get his feet wet; and yet the etiology of the weather must surely be child's play compared with the causes of behavior.

Yes, but what about all those ceteris paribuses? I commence to digress:

Philosophers sometimes argue that the appearance of predictive adequacy that accrues to the generalizations of commonsense psychology is spurious. For, they say, as soon as you try to make these generalizations explicit, you see that they have to be hedged about with ceteris paribus clauses; hedged about in ways that make them *trivially* incapable of disconfirmation. "False or vacuous" is the charge.

Consider the defeasibility of 'if someone utters the form of words "I'll be at your airport on the 3 p.m. flight," then he intends to be at your airport on the 3 p.m. flight.' This generalization does *not* hold if, for example, the speaker is lying; or if the speaker is using the utterance as an example (of a false sentence, say); or if he is a monolingual speaker of Urdu who happens to have uttered the sentence by accident; or if the speaker is talking in his sleep; or . . . whatever. You can, of course, defend the generalization in the usual way; you can say that *'all else being equal,* if someone utters the form of words "I'll be at your airport on the 3 p.m. flight," then he intends to be at your airport on the 3 p.m. flight.' But perhaps this last means nothing more than: 'if someone says that he intends to be there, then he does intend to be there—unless he doesn't.' That, of course, is predictively adequate for sure; nothing that happens will disconfirm it; nothing that happens could.

A lot of philosophers seem to be moved by this sort of argument; yet, even at first blush, it would be surprising if it were any good. After all, we do use commonsense psychological generalizations to predict one another's behavior; and the predictions do—very often—come out true. But how could that be so if the generalizations that we base the predictions on are *empty?*

I'm inclined to think that what is alleged about the implicit reliance of commonsense psychology on uncashed ceteris paribus clauses is in fact a perfectly general property of the *explicit* generalizations in *all* the special sciences; in all empirical explanatory schemes, that is to say, other than basic physics. Consider the following modest truth of geology: A meandering river erodes its outside bank. "False or vacuous"; so a philosopher might argue. "Take it straight—as a strictly

universal generalization—and it is surely false. Think of the case where the weather changes and the river freezes; or the world comes to an end; or somebody builds a dam; or somebody builds a concrete wall on the outside bank; or the rains stop and the river dries up . . . or whatever. You can, of course, defend the generalization in the usual way—by appending a ceteris paribus clause: '*All else being equal,* a meandering river erodes its outside bank.' But perhaps this last means nothing more than: 'A meandering river erodes its outside bank—unless it doesn't.' That, of course, is predictively adequate for sure. Nothing that happens will disconfirm it; nothing that happens could."

Patently, something has gone wrong. For 'All else being equal, a meandering river erodes its outside bank' is neither false nor vacuous, and it doesn't mean 'A meandering river erodes its outside bank—unless it doesn't.' It is, I expect, a long story how the generalizations of the special sciences manage to be both hedged and informative (or, if you like, how they manage to support counterfactuals even though they have exceptions). Telling that story is part of making clear why we have special sciences at all; why we don't just have basic physics (see Fodor, *SS*). It is also part of making clear how idealization works in science. For surely 'Ceteris paribus, a meandering river erodes its outside bank' means something like 'A meandering river erodes its outside bank in any nomologically possible world where the operative idealizations of geology are satisfied.' That this is, in general, stronger than '*P* in any world where not not-*P*' is certain. So if, as it would appear, commonsense psychology relies upon its ceteris paribus clauses, so too does geology.

There is, then, a face similarity between the way implicit generalizations work in commonsense psychology and the way explicit generalizations work in the special sciences. But maybe this similarity is *merely* superficial. Donald Davidson is famous for having argued that the generalizations of real science, unlike those that underlie commonsense belief/desire explanations, are "perfectible." In the real, but not the intentional, sciences we can (in principle, anyhow) get rid of the ceteris paribus clauses by actually enumerating the conditions under which the generalizations are supposed to hold.

By this criterion, however, the only real science is basic physics. For it simply isn't true that we can, even in principle, specify the conditions under which—say—geological generalizations hold *so long as we stick to the vocabulary of geology.* Or, to put it less in the formal mode, the causes of exceptions to geological generalizations are, quite typically, not themselves *geological* events. Try it and see: 'A meandering river erodes its outer banks unless, for example, the weather changes

and the river dries up.' But 'weather' isn't a term in *geology;* nor are 'the world comes to an end,' 'somebody builds a dam,' and indefinitely many other descriptors required to specify the sorts of things that can go wrong. All you can say that's any use is: If the generalization failed to hold, then the operative idealizations must somehow have failed to be satisfied. But so, too, in commonsense psychology: If he didn't turn up when he intended to, then something must have gone wrong.

Exceptions to the generalizations of a special science are typically *inexplicable* from the point of view of (that is, in the vocabulary of) that science. That's one of the things that makes it a *special* science. But, of course, it may nevertheless be perfectly possible to explain the exceptions *in the vocabulary of some other science.* In the most familiar case, you go 'down' one or more levels and use the vocabulary of a more 'basic' science. (The current failed to run through the circuit because the terminals were oxidized; he no longer recognizes familiar objects because of a cerebral accident. And so forth.) The availability of this strategy is one of the things that the hierarchical arrangement of our sciences buys for us. Anyhow, to put the point succinctly, the same pattern that holds for the special sciences seems to hold for commonsense psychology as well. On the one hand, its ceteris paribus clauses are ineliminable from the point of view of its proprietary conceptual resources. But, on the other hand, we have—so far at least—no reason to doubt that they can be discharged in the vocabulary of some lower-level science (neurology, say, or biochemistry; at worst, physics).

If the world is describable as a closed causal system at all, it is so only in the vocabulary of our most basic science. From this nothing follows that a psychologist (or a geologist) needs to worry about.

I cease to digress. The moral so far is that the predictive adequacy of commonsense psychology is beyond rational dispute; nor is there any reason to suppose that it's obtained by cheating. If you want to know where my physical body will be next Thursday, mechanics—our best science of middle-sized objects after all, and reputed to be pretty good in its field—is *no use to you at all.* Far the best way to find out (usually, in practice, the *only* way to find out) is: *ask me!*

2. *The Depth of the Theory*

It's tempting to think of commonsense psychology as merely a budget of such truisms as one learns at Granny's knee: that the burnt child fears the fire, that all the world loves a lover, that money can't buy happiness, that reinforcement affects response rate, and that the

way to a man's heart is through his stomach. None of these, I agree, is worth saving. However, as even the simple example sketched above serves to make clear, subsumption under platitudes is *not* the typical form of commonsense psychological explanation. Rather, when such explanations are made explicit, they are frequently seen to exhibit the 'deductive structure' that is so characteristic of explanation in real science. There are two parts to this: the theory's underlying generalizations are defined over unobservables, and they lead to its predictions by iterating and interacting rather than by being directly instantiated.

Hermia, for example, is no fool and no behaviorist; she is perfectly aware both that Demetrius's behavior is caused by his mental states and that the pattern of such causation is typically intricate. There are, in particular, no plausible and counterfactual-supporting generalizations of the form *(x) (y) (x is a rival of y)* → *(x kills y)*. Nothing like that is remotely true; not even ceteris paribus. Rather, the generalization Hermia takes to be operative—the one that *is* true and counterfactual-supporting—must be something like *If x is y's rival, then x prefers y's discomfiture, all else being equal.* This principle, however, doesn't so much as mention behavior; it leads to behavioral predictions, but only via a lot of further assumptions about how people's preferences may affect their actions in given situations. Or rather, since there probably are no generalizations which connect preferences to actions irrespective of beliefs, what Hermia must be relying on is an implicit theory of how beliefs, preferences, and behaviors interact; an implicit decision theory, no less.

It is a deep fact about the world that the most powerful etiological generalizations hold of unobservable causes. Such facts shape our science (they'd better!). It is thus a test of the depth of a theory that many of its generalizations subsume interactions among unobservables. By this test, our implicit, commonsense *meteorology* is presumably *not* a deep theory, since it consists largely of rule-of-thumb generalizations of the "red at night, sailor's delight" variety. Correspondingly, the reasoning that mediates applications of commonsense meteorology probably involves not a lot more than instantiation and modus ponens. (All this being so, it is perhaps not surprising that commonsense meteorology doesn't work very well.) Commonsense psychology, by contrast, passes the test. It takes for granted that overt behavior comes at the end of a causal chain whose links are mental events—hence unobservable—and which may be arbitrarily long (and arbitratily kinky). Like Hermia, we are all—quite literally, I expect—born mentalists and Realists; and we stay that way until common sense is driven out by bad philosophy.

3. *Its Indispensability*

We have, in practice, no alternative to the vocabulary of commonsense psychological explanation; we have no other way of describing our behaviors and their causes if we want our behaviors and their causes to be subsumed by any counterfactual-supporting generalizations that we know about. This is, again, hard to see because it's so close.

For example, a few paragraphs back, I spoke of the commonsense psychological generalization *people generally do what they say that they will do* as bridging the gap between an exchange of utterances ("Will you come and lecture . . . ," "I'll be at your airport on Thursday . . .") and the consequent behaviors of the speakers (my arriving at the airport, his being there to meet me). But this understates the case for the indispensability of commonsense psychology, since without it we can't even describe the utterances as forms of words (to say nothing of describing the ensuing behaviors as kinds of acts). *Word* is a *psychological* category. (It is, indeed, *irreducibly* psychological, so far as anybody knows; there are, for example, no acoustic properties that all and only tokens of the same word type must share. In fact, surprisingly, there are no acoustic properties that all and only *fully intelligible* tokens of the same word type must share. Which is why our best technology is currently unable to build a typewriter that you can dictate to.)

As things now stand—to spell it out—we have *no* vocabulary for specifying event types that meets the following four conditions:

1. My behavior in uttering 'I'll be there on Thursday . . .' counts as an event of type T_i.
2. My arriving there on Thursday counts as an event of Type T_j.
3. 'Events of type T_j are consequent upon events of type T_i' is even roughly true and counterfactual supporting.
4. Categories T_i and T_j are other than irreducibly psychological.

For the only known taxonomies that meet conditions 1–3 acknowledge such event types as uttering the *form of words* 'I'll be there on Thursday', or *saying that* one will be there on Thursday, or *performing the act* of meeting someone at the airport; so they fail condition 4.

Philosophers and psychologists used to dream of an alternative conceptual apparatus, one in which the commonsense inventory of types of *behavior* is replaced by an inventory of types of *movements;* the counterfactual-supporting generalizations of psychology would then exhibit the contingency of these movements upon environmental and/or organic variables. That behavior is indeed contingent upon environmental and organic variables is, I suppose, not to be denied;

yet the generalizations were not forthcoming. Why? There's a standard answer: It's because behavior consists of actions, and actions cross-classify movements. The generalization is that the burnt child avoids the fire; but what movement constitutes avoidance depends on where the child is, where the fire is . . . and so, drearily, forth. If you want to know what generalizations subsume a behavioral event, you have to know what *action type* it belongs to; knowing what *motion type* it belongs to usually doesn't buy anything. I take all that to be Gospel.

Yet it is generally assumed that this situation *must* be remediable, at least in principle. After all, the generalizations of a completed physics would presumably subsume every motion of every thing, hence the motions of organisms *inter alia*. So, if we wait long enough, we will after all have counterfactual-supporting generalizations that subsume the motions of organisms *under that description*. Presumably, God has them already.

This is, however, a little misleading. For, the (putative) generalizations of the (putative) completed physics would apply to the motions of organisms qua motions, but not qua organismic. Physics presumably has as little use for the categories of macrobiology as it does for the categories of commonsense psychology; it dissolves the behav*er* as well as the behav*ior*. What's left is atoms in the void. The subsumption of the motions of organisms—and of everything else—by the counterfactual-supporting generalizations of physics does not therefore guarantee that there is any science whose ontology recognizes organisms and their motions. That is: The subsumption of the motions of organisms—and of everything else—by the laws of physics does not guarantee that there are any laws about the motions of organisms qua motions of organisms. So far as anybody knows—barring, perhaps, a little bit of the psychology of classical reflexes—there are no such laws; and there is no metaphysical reason to expect any.[1]

Anyhow, this is all poppycock. Even if psychology were dispensable *in principle*, that would be no argument for dispensing with it. (Perhaps geology is dispensable in principle; every river is a physical object after all. Would that be a reason for supposing that rivers aren't a natural kind? Or that 'meandering rivers erode their outside banks' is untrue?) What's relevant to whether commonsense psychology is worth defending is its dispensability *in fact*. And here the situation is absolutely clear. We have no idea of how to explain ourselves to ourselves except in a vocabulary which is *saturated* with belief/desire psychology. One is tempted to transcendental argument: What Kant said to Hume about physical objects holds, mutatis mutandis, for the

propositional attitudes; we can't give them up *because we don't know how to.*[2]

So maybe we had better try to hold onto them. Holding onto the attitudes—vindicating commonsense psychology—means showing how you could have (or, at a minimum, showing *that* you could have) a respectable science whose ontology explicitly acknowledges states that exhibit the sorts of properties that common sense attributes to the attitudes. That is what the rest of this book is about. This undertaking presupposes, however, some consensus about what sorts of properties common sense does attribute to the attitudes. That is what the next bit of this chapter is about.

The Essence of the Attitudes

How do we tell whether a psychology *is* a belief/desire psychology? How, in general, do we know if propositional attitudes are among the entities that the ontology of a theory acknowledges? These sorts of questions raise familiar and perplexing issues of intertheoretic identification. How do you distinguish elimination from reduction and reconstruction? Is the right story that there's no such thing as dephlogistinated matter, or is 'dephlogistinizing' just a word for oxidizing? Even behaviorists had trouble deciding whether they wanted to deny the existence of the mental or to assert its identity with the behavioral. (Sometimes they did both, in successive sentences. Ah, they really knew about insouciance in those days.)

I propose to stipulate. I will view a psychology as being commonsensical about the attitudes—in fact, as endorsing them—just in case it postulates states (entities, events, whatever) satisfying the following conditions:

(*i*) They are semantically evaluable.
(*ii*) They have causal powers.
(*iii*) The implicit generalizations of commonsense belief/desire psychology are largely true of them.

In effect, I'm assuming that (*i*)–(*iii*) are the essential properties of the attitudes. This seems to me intuitively plausible; if it doesn't seem intuitively plausible to you, so be it. Squabbling about intuitions strikes me as vulgar.

A word about each of these conditions.

(i) Semantic Evaluation

Beliefs are the kinds of things that are true or false; desires are the kinds of things that get frustrated or fulfilled; hunches are the kinds

of things that turn out to be right or wrong; so it goes. I will assume that what makes a belief true (/false) is something about its relation to the nonpsychological world (and not—e.g.—something about its relation to other beliefs; unless it happens to be a belief about beliefs). Hence, to say of a belief that it is true (/false) is to evaluate that belief in terms of its relation to the world. I will call such evaluations 'semantic.' Similarly, mutatis mutandis, with desires, hunches, and so forth.

It is, as I remarked in the preface, a puzzle about beliefs, desires, and the like that they are semantically evaluable; almost nothing else is. (Trees aren't; numbers aren't; people aren't. Propositions *are* [assuming that there are such things], but that's hardly surprising; propositions exist to be what beliefs and desires are attitudes *toward*.) We will see, later in this book, that it is primarily the semantic evaluability of beliefs and desires that gets them into philosophical trouble—and that a defense of belief/desire psychology needs to be a defense of.

Sometimes I'll talk of the *content* of a psychological state rather than its semantic evaluability. These two ideas are intimately interconnected. Consider—for a change of plays—Hamlet's belief that his uncle killed his father. That belief has a certain semantic value; in particular, it's a *true* belief. Why true? Well, because it corresponds to a certain fact. Which fact? Well, the fact that Hamlet's uncle killed Hamlet's father. But why is it *that* fact that determines the semantic evaluation of Hamlet's belief? Why not the fact that two is a prime number, or the fact that Demetrius didn't kill Lysander? Well, because the *content* of Hamlet's belief is *that* his uncle killed his father. (If you like, the belief 'expresses the proposition' that Hamlet's uncle killed his father.) *If you know what the content of a belief is, then you know what it is about the world that determines the semantic evaluation of the belief;* that, at a minimum, is how the notions of content and semantic evaluation connect.

I propose to say almost nothing more about content at this stage; its time will come. Suffice it just to add that propositional attitudes have their contents essentially: the canonical way of picking out an attitude is to say (a) what sort of attitude it is (a belief, a desire, a hunch, or whatever); and (b) what the content of the attitude is (that Hamlet's uncle killed his father; that 2 is a prime number; that Hermia believes that Demetrius dislikes Lysander; or whatever). In what follows, nothing will count as a propositional-attitude psychology—as a reduction or reconstruction or vindication of commonsense belief/desire explanation—that does not acknowledge states that can be individuated in this sort of way.

(ii) Causal Powers

Commonsense psychological explanation is deeply committed to mental causation of at least three sorts: the causation of behavior by mental states; the causation of mental states by impinging environmental events (by 'proximal stimulation,' as psychologists sometimes say); and—in some ways the most interesting commonsense psychological etiologies—the causation of mental states by one another. As an example of the last sort, common sense acknowledges *chains of thought* as species of complex mental events. A chain of thought is presumably a *causal* chain in which one semantically evaluable mental state gives rise to another; a process that often terminates in the fixation of belief. (That, as you will remember, was the sort of thing Sherlock Holmes was supposed to be very good at.)

Every psychology that is Realist about the mental ipso facto acknowledges its causal powers.[3] Philosophers of 'functionalist' persuasion even hold that the causal powers of a mental state determine its identity (that for a mental state to be, as it might be, the state of believing that Demetrius killed Lysander is just for it to have a characteristic galaxy of potential and actual causal relations). This is a position of some interest to us, since if it is true—and if it is also true that propositional attitudes have their contents essentially—it follows that the causal powers of a mental state somehow determine its content. I do not, however, believe that it is true. More of this later.

What's important for now is this: It is characteristic of commonsense belief/desire psychology—and hence of any explicit theory that I'm prepared to view as vindicating commonsense belief/desire psychology—that it attributes contents and causal powers *to the very same mental things that it takes to be semantically evaluable.* It is Hamlet's belief that Claudius killed his father—the very same belief which is true or false in virtue of the facts about his father's death—that causes him to behave in such a beastly way to Gertrude.[4]

In fact, there's a deeper point to make. It's not just that, in a psychology of propositional attitudes, content and causal powers are attributed to the same things. It's also that causal relations among propositional attitudes somehow typically contrive to respect their relations of content, and belief/desire explanations often turn on this. Hamlet believed that somebody had killed his father because he believed that Claudius had killed his father. His having the second belief explains his having the first. How? Well, presumably via some such causal generalization as 'if someone believes Fa, then ceteris paribus he believes $\exists x(Fx)$.' This generalization specifies a causal relation between two kinds of mental states picked out by reference to (the logical form of) the propositions they express; so we have the

usual pattern of a simultaneous attribution of content and causal powers. The present point, however, is that the contents of the mental states that the causal generalization subsumes are themselves semantically related; *Fa entails* $\exists x(Fx)$, so, of course, the semantic value of the latter belief is not independent of the semantic value of the former.

Or, compare the pattern of implicit reasoning attributed to Hermia at the beginning of this chapter. I suggested that she must be relying crucially on some such causal generalization as: 'If x wants that P, and x believes that $-P$ unless Q, and x believes that it is within his power to bring it about that Q, then ceteris paribus x tries to bring it about that Q.' Common sense seems pretty clearly to hold that something like that is true and counterfactual supporting; hence that one has explained x's attempt to bring it about that Q if one shows that x had beliefs and desires of the sort that the generalization specifies. What is absolutely typical is (a) the appeal to causal relations among semantically evaluable mental states as part and parcel of the explanation; and (b) the existence of content relations among the mental states thus appealed to.

Witness the recurrent schematic letters; they function precisely to constrain the content relations among the mental states that the generalization subsumes. Thus, unless, in a given case, what x wants is the same as what x believes that he can't have without Q, and unless what x believes to be required for P is the same as what he tries to bring about, the generalization isn't satisfied and the explanation fails. It is self-evident that the explanatory principles of commonsense psychology achieve generality by quantifying over agents (the 'practical syllogism' purports to apply, ceteris paribus, to all the x's). But it bears emphasis that they also achieve generality by abstracting over *contents* ('If you want P and you believe not-P unless Q . . . you try to bring it about that Q,' whatever the P and Q may be). The latter strategy works only because, very often, the same P's and Q's—the same contents—recur in causally related mental states; viz., only because causal relations very often respect semantic ones.

This parallelism between causal powers and contents engenders what is, surely, one of the most striking facts about the cognitive mind as commonsense belief/desire psychology conceives it: the frequent similarity between trains of thought and *arguments*. Here, for example, is Sherlock Holmes doing his thing at the end of "The Speckled Band":

> I instantly reconsidered my position when . . . it became clear to me that whatever danger threatened an occupant of the room couldn't come either from the window or the door. My attention

> was speedily drawn, as I have already remarked to you, to this ventilator, and to the bell-rope which hung down to the bed. The discovery that this was a dummy, and that the bed was clamped to the floor, instantly gave rise to the suspicion that the rope was there as a bridge for something passing through the hole, and coming to the bed. The idea of a snake instantly occurred to me, and when I coupled it with my knowledge that the Doctor was furnished with a supply of the creatures from India I felt that I was probably on the right track.

The passage purports to be a bit of reconstructive psychology: a capsule history of the sequence of mental states which brought Holmes first to suspect, then to believe, that the doctor did it with his pet snake. What is therefore interesting, for our purposes, is that Holmes's story isn't *just* reconstructive psychology. It does double duty, since it also serves to assemble *premises* for a plausible inference to the *conclusion* that the doctor did it with the snake. Because his train of thought is like an argument, Holmes expects Watson to be *convinced* by the considerations which, when they occurred to Holmes, caused his own conviction. What connects the causal-history aspect of Holmes's story with its plausible-inference aspect is the fact that the thoughts that fix the belief that *P* provide, often enough, reasonable *grounds* for believing that *P*. Were this not the case—were there not this general harmony between the semantical and the causal properties of thoughts, so that, as Holmes puts it in another story, "one true inference invariably suggests others"—there wouldn't, after all, be much profit in thinking.

All this raises a budget of philosophical issues; just *what sorts* of content relations are preserved in the generalizations that subsume typical cases of belief/desire causation? And—in many ways a harder question—how could the mind be so constructed that such generalizations are true of it? What sort of mechanism could have states that are both semantically and causally connected, and such that the causal connections respect the semantic ones? It is the intractability of such questions that causes many philosophers to despair of commonsense psychology. But, of course, the argument cuts both ways: if the parallelism between content and causal relations is, as it seems to be, a deep fact about the cognitive mind, then unless we can save the notion of content, there is a deep fact about the cognitive mind that our psychology is going to miss.

(iii) Generalizations Preserved

What I've said so far amounts largely to this: An explicit psychology that vindicates commonsense belief/desire explanations must permit

the assignment of content to causally efficacious mental states and must recognize behavioral explanations in which covering generalizations refer to (or quantify over) the contents of the mental states that they subsume. I now add that the generalizations that are recognized by the vindicating theory mustn't be *crazy* from the point of view of common sense; the causal powers of the attitudes must be, more or less, what common sense supposes that they are. After all, commonsense psychology won't be vindicated unless it turns out to be at least approximately true.

I don't, however, have a shopping list of commonsense generalizations that must be honored by a theory if it wants to be ontologically committed to bona fide propositional attitudes. A lot of what common sense believes about the attitudes must surely be false (a lot of what common sense believes about *anything* must surely be false). Indeed, one rather hopes that there will prove to be many more—and much odder—things in the mind than common sense had dreamed of; or else what's the fun of doing psychology? The indications are, and have been since Freud, that this hope will be abundantly gratified. For example, contrary to common sense, it looks as though much of what's in the mind is unconscious; and, contrary to common sense, it looks as though much of what's in the mind is unlearned. I retain my countenance, I remain self-possessed.

On the other hand, there is a lot of commonsense psychology that we have—so far at least—no reason to doubt, and that friends of the attitudes would hate to abandon. So, it's hard to imagine a psychology of action that is committed to the attitudes but doesn't acknowledge some such causal relations among beliefs, desires, and behavioral intentions (the 'maxims' of acts) as decision theories explicate. Similarly, it's hard to imagine a psycholinguistics (for English) which attributes beliefs, desires, communicative intentions, and such to speaker/hearers but fails to entail an infinity of theorems recognizably similar to these:

- 'Demetrius killed Lysander' is the form of words standardly used to communicate the belief that Demetrius killed Lysander.
- 'The cat is on the mat' is the form of words standardly used to communicate the belief that the cat is on the mat.
- 'Demetrius killed Lysander or the cat is on the mat' is the form of words standardly used to communicate the belief that Demetrius killed Lysander or the cat is on the mat.

And so on indefinitely. Indeed, it's hard to imagine a psycholinguistics that appeals to the propositional attitudes of speaker/hearers of English to explain their verbal behavior but that doesn't entail that

they *know* at least one such theorem for each sentence of their language. So there's an infinite amount of common sense for psychology to vindicate already.

Self-confident essentialism is philosophically fashionable this week. There are people around who have Very Strong Views ('modal intuitions,' these views are called) about whether there could be cats in a world in which all the domestic felines are Martian robots, and whether there could be Homer in a world where nobody wrote the *Odyssey* or the *Iliad*. Ducky for them; their epistemic condition is enviable, but I don't myself aspire to it. I just don't know how much commonsense psychology would have to be true for there to be beliefs and desires. Let's say, some of it at a minimum; lots of it by preference. Since I have no doubt at all but that lots of it *is* true, this is an issue about which I do not stay up nights worrying.

RTM

The main thesis of this book can now be put as follows: *We have no reason to doubt—indeed, we have substantial reason to believe—that it is possible to have a scientific psychology that vindicates commonsense belief/desire explanation.* But though that is my thesis, I don't propose to argue the case in quite so abstract a form. For there is already in the field a (more or less) empirical theory that is, in my view, reasonably construed as ontologically committed to the attitudes and that—again, in my view—is quite probably approximately true. If I'm right about this theory, it *is* a vindication of the attitudes. Since, moreover, it's the only thing of its kind around (it's the *only* proposal for a scientific belief/desire psychology that's in the field), defending the commonsense assumptions about the attitudes and defending this theory turn out to be much the same enterprise; extensionally, as one might say.

That, in any event, is the strategy that I'll pursue: I'll argue that the sorts of objections philosophers have recently raised against belief/desire explanation are (to put it mildly) not conclusive against the best vindicating theory currently available. The rest of this chapter is therefore devoted to a sketch of how this theory treats the attitudes and why its treatment of the attitudes seems so promising. Since this story is now pretty well known in both philosophical and psychological circles, I propose to be quick.

What I'm selling is the Representational Theory of Mind (hence RTM; for discussion see, among other sources, Fodor, *PA;* Fodor, *LOT;* Field, *MR*). At the heart of the theory is the postulation of a language of thought: an infinite set of 'mental representations' which

function both as the immediate objects of propositional attitudes and as the domains of mental processes. More precisely, RTM is the conjunction of the following two claims:

> *Claim 1* (the nature of propositional attitudes):
>
> For any organism *O*, and any attitude *A* toward the proposition *P*, there is a ('computational'/'functional') relation *R* and a mental representation *MP* such that
>
> *MP* means that *P*, and
> *O* has *A* iff *O* bears *R* to *MP*.

(We'll see presently that the biconditional needs to be watered down a little; but not in a way that much affects the spirit of the proposal.)

It's a thin line between clarity and pomposity. A cruder but more intelligible way of putting claim 1 would be this: To believe that such and such is to have a mental symbol that means that such and such tokened in your head in a certain way; it's to have such a token 'in your belief box,' as I'll sometimes say. Correspondingly, to hope that such and such is to have a token of that same mental symbol tokened in your head, but in a rather different way; it's to have it tokened 'in your hope box.' (The difference between having the token in one box or the other corresponds to the difference between the causal roles of beliefs and desires. Talking about belief boxes and such as a shorthand for representing the attitudes as *functional* states is an idea due to Steve Schiffer. For more on this, see the Appendix.) And so on for every attitude that you can bear toward a proposition; and so on for every proposition toward which you can bear an attitude.

> *Claim 2* (the nature of mental processes):
>
> Mental processes are causal sequences of tokenings of mental representations.

A train of thoughts, for example, is a causal sequence of tokenings of mental representations which express the propositions that are the objects of the thoughts. To a first approximation, to think 'It's going to rain; so I'll go indoors' is to have a tokening of a mental representation that means *I'll go indoors* caused, in a certain way, by a tokening of a mental representation that means *It's going to rain.*

So much for formulating RTM.

There are, I think, a number of reasons for believing that RTM may be more or less true. The best reason is that some version or other of RTM underlies practically all current psychological research on mentation, and our best science is ipso facto our best estimate of what there is and what it's made of. There are those of my colleagues in

philosophy who do not find this sort of argument persuasive. I blush for them. (For a lengthy discussion of how RTM shapes current work on cognition, see Fodor, *LOT*, especially chapter 1. For a discussion of the connection between RTM and commonsense Intentional Realism—and some arguments that, given the latter, the former is practically mandatory—see the Appendix.)

But we have a reason for suspecting that RTM may be true even aside from the details of its empirical success. I remarked above that there is a striking parallelism between the causal relations among mental states, on the one hand, and the semantic relations that hold among their propositional objects, on the other; and that very deep properties of the mental—as, for example, that trains of thought are largely truth preserving—turn on this symmetry. RTM suggests a plausible mechanism for this relation, and that is something that no previous account of mentation has been able to do. I propose to spell this out a bit; it helps make clear just *why* RTM has such a central place in the way that psychologists now think about the mind.

The trick is to combine the postulation of mental representations with the 'computer metaphor.' Computers show us how to connect semantical with causal properties for *symbols*. So, if having a propositional attitude involves tokening a symbol, then we can get some leverage on connecting semantical properties with causal ones for *thoughts*. In this respect, I think there really has been something like an intellectual breakthrough. Technical details to one side, this is—in my view—the only aspect of contemporary cognitive science that represents a major advance over the versions of mentalism that were its eighteenth- and nineteenth-century predecessors. Exactly what was wrong with Associationism, for example, was that there proved to be no way to get a *rational* mental life to emerge from the sorts of causal relations among thoughts that the 'laws of association' recognized. (See the concluding pages of Joyce's *Ulysses* for a—presumably inadvertent—parody of the contrary view.)

Here, in barest outline, is how the new story is supposed to go: You connect the causal properties of a symbol with its semantic properties *via its syntax*. The syntax of a symbol is one of its higher-order physical properties. To a metaphorical first approximation, we can think of the syntactic structure of a symbol as an abstract feature of its shape.[5] Because, to all intents and purposes, syntax reduces to shape, and because the shape of a symbol is a potential determinant of its causal role, it is fairly easy to see how there could be environments in which the causal role of a symbol correlates with its syntax. It's easy, that is to say, to imagine symbol tokens interacting causally *in virtue of* their syntactic structures. The syntax of a symbol might determine the

causes and effects of its tokenings in much the way that the geometry of a key determines which locks it will open.

But, now, we know from modern logic that certain of the semantic relations among symbols can be, as it were, 'mimicked' by their syntactic relations; that, when seen from a very great distance, is what proof-theory is about. So, within certain famous limits, the semantic relation that holds between two symbols when the proposition expressed by the one is entailed by the proposition expressed by the other can be mimicked by syntactic relations in virtue of which one of the symbols is derivable from the other. We can therefore build machines which have, again within famous limits, the following property:

> The operations of the machine consist entirely of transformations of symbols;
>
> in the course of performing these operations, the machine is sensitive solely to syntactic properties of the symbols;
>
> and the operations that the machine performs on the symbols are entirely confined to altering their shapes.

Yet the machine is so devised that it will transform one symbol into another if and only if the propositions expressed by the symbols that are so transformed stand in certain *semantic* relations—e.g., the relation that the premises bear to the conclusion in a valid argument. Such machines—computers, of course—just *are* environments in which the syntax of a symbol determines its causal role in a way that respects its content. This is, I think, a perfectly terrific idea; not least because it works.

I expect it's clear how this is supposed to connect with RTM and ontological commitment to mental representations. Computers are a solution to the problem of mediating between the causal properties of symbols and their semantic properties. So *if* the mind is a sort of computer, we begin to see how you can have a theory of mental processes that succeeds where—literally—all previous attempts had abjectly failed; a theory which explains how there could be nonarbitrary content relations among causally related thoughts. But, patently, there are going to have to be mental representations if this proposal is going to work. In computer design, causal role is brought into phase with content by exploiting parallelisms between the syntax of a symbol and its semantics. But that idea won't do the theory of *mind* any good unless there are *mental* symbols: mental particulars possessed of both semantical and syntactic properties. There must be mental symbols because, in a nutshell, only symbols have syntax, and our best

available theory of mental processes—indeed, the *only* available theory of mental processes that isn't *known* to be false—needs the picture of the mind as a syntax-driven machine.

It is sometimes alleged against commonsense belief/desire psychology, by those who admire it less than I do (see especially Churchland, *EMPA;* Stich, *FFPCS*), that it is a "sterile" theory; one that arguably hasn't progressed much since Homer and hasn't progressed at all since Jane Austen. There is, no doubt, a sense in which this charge is warranted; commonsense psychology may be implicit science, but it isn't, on anybody's story, implicit *research* science. (What novelists and poets do doesn't count as research by the present austere criteria.) If, in short, you want to evaluate progress, you need to look not at the implicit commonsense theory but at the best candidate for its explicit vindication. And here the progress has been enormous. It's not just that we now know a little about memory and perception (qua means to the fixation of belief), and a little about language (qua means to the communication of belief); see any standard psychology text. The real achievement is that we are (maybe) on the verge of solving a great mystery about the mind: *How could its causal processes be semantically coherent?* Or, if you like yours with drums and trumpets: *How is rationality mechanically possible?*[6] Notice that this sort of problem can't even be stated, let alone solved, unless we suppose—just as commonsense belief/desire psychology wants us to—that there are mental states with both semantic contents and causal roles. A good theory is one that leads you to ask questions that have answers. And vice versa, ceteris paribus.

Still, RTM won't do in quite the raw form set forth above. I propose to end this chapter with a little polishing.

According to claim 1, RTM requires both of the following:

> For each tokening of a propositional attitude, there is a tokening of a corresponding relation between an organism and a mental representation;

and

> For each tokening of that relation, there is a corresponding tokening of a propositional attitude.[7]

This is, however, much too strong; the equivalence fails in both directions.

As, indeed, we should expect it to, given our experience in other cases where explicit science co-opts the conceptual apparatus of common sense. For example, as everybody points out, it is simply not true that chemistry identifies each sample of water with a sample of

H_2O; not, at least, if the operative notion of water is the commonsense one according to which what we drink, sail on, and fill our bathtubs with all qualifies. What chemistry does is reconstruct the commonsense categories *in what the theory itself identifies as core cases: chemically pure* water is H_2O. The ecological infrequency of such core cases is, of course, no argument against the claim that chemical science vindicates the commonsense taxonomy: Common sense was right about there being such stuff as water, right about there being water in the Charles River, and right again that it's the water in what we drink that quenches our thirst. It never said that the water in the Charles is chemically pure; 'chemically pure' isn't a phrase in the commonsense vocabulary.

Exactly similarly, RTM vindicates commonsense psychology for what RTM identifies as the core cases; in those cases, what common sense takes to be tokenings of propositional attitudes are indeed tokenings of a relation between an organism and a mental representation. The other cases—where you get either attitude tokenings without the relation or relation tokenings without the attitudes—the theory treats as derivative. This is all, I repeat, *exactly* what you'd expect from scientific precedent. Nevertheless, philosophers have made an awful fuss about it in discussing the vindication of the attitudes (see the controversy over the 'explicit representation'—or otherwise—of grammars recently conducted by, among others, Stabler [*HAGR*] and Demopoulos and Matthews [*HGMR*]). So let's consider the details awhile. Doing so will lead to a sharpening of claim 1, which is all to the good.

Case 1. Attitudes without Mental Representations
Here's a case from Dennett:

> In a recent conversation with the designer of a chess-playing program I heard the following criticism of a rival program: "It thinks it should get its queen out early." This ascribes a propositional attitude to the program in a very useful and predictive way, for as the designer went on to say, one can usually count on chasing that queen around the board. But for all the many levels of explicit representation to be found in that program, nowhere is anything roughly synonymous with "I should get my queen out early" explicitly tokened. The level of analysis to which the designer's remark belongs describes features of the program that are, in an entirely innocent way, emergent properties of the computational processes that have "engineering reality." I see no reason to believe that the relation

> between belief-talk and psychological-process talk will be any more direct (*CCC*, 107; see also Matthews, *TWR*)

Notice that the problem Dennett raises isn't just that some of what common sense takes to be one's propositional attitudes are *dispositional*. It's not like the worry that I might now be said to believe some abstruse consequence of number theory—one that I have, commonsensically speaking, never even thought of—because I *would* accept the proof of the theorem *if* I were shown it. It's true, of course, that merely dispositional beliefs couldn't correspond to *occurrent* tokenings of relations to mental representations, and claim 1 must therefore be reformulated. But the problem is superficial, since the relevant revision of claim 1 would be pretty obvious; viz., that for each *occurrent* belief there is a corresponding *occurrent* tokening of a mental representation; and for each *dispositional* belief there is a corresponding *disposition* to token a mental representation.

This would leave open a question that arises independent of one's views about RTM: viz., when are attributions of dispositional beliefs *true?* I suppose that one's dispositional beliefs could reasonably be identified with the closure of one's occurrent beliefs under principles of inference that one explicitly accepts. And, if it's a little vague just what beliefs belong to such a closure, RTM could live with that. *Qua dispositional,* attitudes play no causal role in *actual* mental processes; only occurrent attitudes—for that matter, only occurrent *anythings*—are actual causes. So RTM can afford to be a little operationalist about merely dispositional beliefs (see Lycan, *TB*) so long as it takes a hard line about occurrent ones.

However, to repeat, the problem raised in Dennett's text is not of this sort. It's not that the program believes 'get your queen out early' *potentially*. Dennett's point is that the program actually operates on this principle; but not in virtue of any tokening of any symbol that expresses it. And chess isn't, of course, the only sort of case. Behavioral commitment to modus ponens, or to the syntactic rule of 'wh'-movement, *might* betoken that these are inscribed in brain writing. But it needn't, since these rules might be—as philosophers sometimes say—complied with but not literally followed.

In Dennett's example, you have an attitude being, as it were, an emergent out of its own implementation. This way of putting it might seem to suggest a way of saving claim 1: The machine doesn't explicitly represent 'get your queen out early,' but at least we may suppose that it *does* represent, explicitly, some more detailed rules of play (the ones that Dennett says have "engineering reality"). For these rules, at least, a strong form of claim 1 would thus be satisfied. But that suggestion won't work either. *None* of the principles in accordance with

which a computational system operates need be explicitly represented by a formula tokened in the device; there is no guarantee that the program of a machine will be explicitly represented in the machine whose program it is. (See Cummins, *IMM;* roughly, the point is that for any machine that computes a function by executing an explicit algorithm, there exists another machine—one that's 'hardwired'—that computes the same function but *not* by executing an explicit algorithm.) So what, you might wonder, does the 'computer metaphor' buy for RTM after all?

There is even a point of principle here—one that is sometimes read in (or into) Lewis Carroll's dialogue between Achilles and the Tortoise: Not all the rules of inference that a computational system runs on *can* be represented *just* explicitly in the system; some of them have to be, as one says, 'realized in the hardware.' Otherwise the machine won't run at all. A computer in which the principles of operation are *only* explicitly represented is just like a blackboard on which the principles have been written down. It has Hamlet's problem: When you turn the thing on, nothing happens.

Since this is all clearly correct and arguably important, the question arises how to state RTM so that these cases where programs are hardwired don't count as disconfirmations of claim 1. We'll return to this momentarily; first let's consider:

Case 2. Mental Representations without Attitudes

What RTM borrows from computers is, in the first instance, the recipe for mechanizing rationality: Use a syntactically driven machine to exploit parallelisms between the syntactic and semantic properties of symbols. Some—but not all—versions of RTM borrow more than this; not just a theory of rationality but a theory of intelligence too. According to this story, intelligent behavior typically exploits a 'cognitive architecture' constituted of *hierarchies* of symbol processors. At the top of such a hierarchy might be a quite complex capacity: solving a problem, making a plan, uttering a sentence. At the bottom, however, are only the sorts of unintelligent operations that Turing machines can perform: deleting symbols, storing symbols, copying symbols, and the rest. Filling in the middle levels is tantamount to reducing—analyzing—an intelligent capacity into a complex of dumb ones; hence to a kind of explanation of the former.

Here's a typical example of a kind of representational theory that runs along these lines:

> This is the way we tie our shoes: There is a little man who lives in one's head. The little man keeps a library. When one acts upon the intention to tie one's shoes, the little man fetches down a

> volume entitled *Tying One's Shoes*. The volume says such things as: "Take the left free end of the shoelace in the left hand. Cross the left free end of the shoelace over the right free end of the shoelace . . . ," etc. . . . When the little man reads "take the left free end of the shoelace in the left hand," we imagine him ringing up the shop foreman in charge of grasping shoelaces. The shop foreman goes about supervising that activity in a way that is, in essence, a microcosm of tying one's shoe. Indeed, the shop foreman might be imagined to superintend a detail of wage slaves, whose functions include: searching representations of visual inputs for traces of shoelace, dispatching orders to flex and contract fingers on the left hand, etc. (Fodor, *ATK*, 63–65, slightly revised)

At the very top are states which may well correspond to propositional attitudes that common sense is prepared to acknowledge (knowing how to tie one's shoes, thinking about shoe tying). But at the bottom and middle levels there are bound to be lots of symbol-processing operations that correspond to nothing that *people*—as opposed to their nervous systems—ever do. These are the operations of what Dennett has called "sub-personal" computational systems; and though they satisfy the present formulation of claim 1 (in that they involve causally efficacious tokenings of mental representations), yet it's unclear that they correspond to anything that common sense would count as the tokening of an attitude. But then how are we to formulate claim 1 so as to avoid disconfirmation by subpersonal information processes?

Vindication Vindicated

There is a sense in which these sorts of objections to claim 1 strike me as not very serious. As I remarked above, the vindication of belief/desire explanation by RTM does *not* require that every case common sense counts as the tokening of an attitude should correspond to the tokening of a mental representation, or vice versa. All that's required is that such correspondences should obtain in what the vindicating theory itself takes to be the core cases. On the other hand, RTM had better be able to say which cases it does count as core. Chemistry is allowed to hold the Charles River largely irrelevant to the confirmation of 'water is H_2O,' but only because it provides independent grounds for denying that what's in the Charles is a chemically pure sample. Of anything!

So, what are the core cases for RTM? The answer should be clear from claim 2. According to claim 2, mental processes are causal sequences of transformations of mental representations. It follows that

tokenings of attitudes *must* correspond to tokenings of mental representations when they—the attitude tokenings—are episodes in mental processes. If the intentional objects of such causally efficacious attitude tokenings are *not* explicitly represented, then RTM is simply false. I repeat for emphasis: If the occurrence of a thought is an episode in a mental process, then RTM is committed to the explicit representation of its content. The motto is therefore No Intentional Causation without Explicit Representation.

Notice that this way of choosing core cases squares us with the alleged counterexamples. RTM says that the contents of a sequence of attitudes that constitutes a mental process must be expressed by explicit tokenings of mental representations. But the rules that determine the course of the transformation of these representations—modus ponens, 'wh'-movement, 'get the queen out early,' or whatever—need not themselves ever be explicit. They can be emergents out of explicitly represented procedures of implementation, or out of hardware structures, or both. Roughly: According to RTM, programs—corresponding to the 'laws of thought'—*may* be explicitly represented; but 'data structures'—corresponding to the contents of thoughts—*have to be.*

Thus, in Dennett's chess case, the rule 'get it out early' may or may not be expressed by a 'mental' (/program language) symbol. That depends on just how the machine works; specifically, on whether *consulting* the rule is a step in the machine's operations. I take it that in the machine that Dennett has in mind, it isn't; *entertaining the thought 'Better get the queen out early' never constitutes an episode in the mental life of that machine.*[8] But then, the intentional content of this thought need *not* be explicitly represented consonant with 'no intentional causation without explicit representation' being true. By contrast, the representations of the board—of actual or possible states of play—over which the machine's computations are defined *must* be explicit, precisely *because* the machine's computations *are* defined over them. These computations constitute the machine's 'mental processes,' so either they are causal sequences of explicit representations, or the representational theory of chess playing is simply false of the machine. To put the matter in a nutshell: Restricting one's attention to the status of rules and programs can make it seem that the computer metaphor is neutral with respect to RTM. But when one thinks about the constitution of mental processes, the connection between the idea that they are computational and the idea that there is a language of thought becomes immediately apparent.[9]

What about the subpersonal examples, where you have mental representation tokenings without attitude tokenings? Commonsense

belief/desire explanations are vindicated if scientific psychology is ontologically committed to beliefs and desires. But it's *not* also required that the folk-psychological inventory of propositional attitudes should turn out to exhaust a natural kind. It would be astounding if it did; how could common sense know all that? What's important about RTM—what makes RTM a vindication of intuitive belief/desire psychology—isn't that it picks out a kind that is precisely coextensive with the propositional attitudes. It's that RTM shows how intentional states could have causal powers; precisely the aspect of commonsense intentional realism that seemed most perplexing from a metaphysical point of view.

Molecular physics vindicates the intuitive taxonomy of middle-sized objects into liquids and solids. But the nearest kind to the liquids that molecular physics acknowledges includes some of what common sense would not; glass, for example. So what?

So much for RTM; so much for this chapter, too. There is a strong prima facie case for commonsense belief/desire explanation. Common sense would be vindicated if some good theory of the mind proved to be committed to entities which—like the attitudes—are both semantically evaluable and etiologically involved. RTM looks like being a good theory of the mind that is so committed; so if RTM is true, common sense is vindicated. It goes without saying that RTM needs to make an empirical case; we need good accounts, independently confirmed, of mental processes as causal sequences of transformations of mental representations. Modern cognitive psychology is devoted, practically in its entirety, to devising and confirming such accounts. For present purposes, I shall take all that as read. What the rest of this book is about is doubts about RTM that turn on its *semantic* assumptions. This is home ground for philosophers, and increasingly the natives are restless.

Notes

Chapter 1

1. Perhaps there are laws that relate the *brain states* of organisms to their motions. But then again, perhaps there aren't, since it seems entirely possible that the lawful connections should hold between brain states and *actions* where, as usual, actions cross-classify movements. This is, perhaps, what you would predict upon reflection. Would you really expect the same brain state that causes the utterance of 'dog' in tokens of 'dog' to be the one that causes it in tokens of 'dogmatic'? How about utterances of (the phonetic sequence) [empedokliz lipt] when you're talking English and when you're talking German?
2. The trouble with transcendental arguments being, however, that it's not obvious why a theory couldn't be both indispensable and *false*. I wouldn't want to buy a transcendental deduction of the attitudes if operationalism were the price I had to pay for it.
3. Denying the etiological involvement of mental states was really what behaviorism was about; it's what 'logical' behaviorists and 'eliminativists' had in common. Thus, for example, to hold—as Ryle did, more or less—that mental states are species of dispositions is to refuse to certify as literally causal such psychological explanations as "He did it with the intention of pleasing her," or, for that matter, "His headache made him groan," to say nothing of "The mere thought of giving a lecture makes him ill." (For discussion, see Fodor, *SSA*.)
4. Some philosophers feel very strongly about enforcing an object/state (or maybe object/event) distinction here, so that what have *causal powers* are tokenings of mental state types (e.g., Hamlet's *believing* that Claudius killed his father), but what have *semantic values* are *propositions* (e.g., the proposition that Claudius killed Hamlet's father). The point is that it sounds odd to say that Hamlet's *believing* that *P* is true but all right to say that Hamlet's *belief* that *P* is.

   I'm not convinced that this distinction is one that I will care about in the long run, since sounding odd is the least of my problems and in the long run I expect I want to do without propositions altogether. However, if you are squeamish about ontology, that's all right with me. In that case, the point in the text should be: Belief/desire psychology attributes causal properties to the very same things (viz., tokenings of certain mental state types) to which it attributes propositional objects. It is thus true of Hamlet's believing that Claudius killed his father both that it is implicated in the etiology of his behavior Gertrudeward and that it has as its object a certain belief, viz., the proposition that Claudius killed his father. If we then speak of Hamlet's *state* of believing that Claudius killed his father (or of the event which consists of the tokening of that state) as semantically evaluable, we can take that as an abbreviation for a more precise way of talking: The state *S* has the semantic value *V* iff *S* has as its object a proposition whose value is *V*.

It goes without saying that none of this ontological fooling around makes the slightest progress toward removing the puzzles about intentionality. If (on my way of talking) it's metaphysically worrying that beliefs and desires are semantically evaluable though trees, rocks, and prime numbers aren't, it's equally metaphysically worrying (on the orthodox way of talking) that believings have propositional objects though trees, rocks, and prime numbers don't.

5. *Any* nomic property of symbol tokens, however—any property in virtue of the posession of which they satisfy causal laws—would, in principle, do just as well. (So, for example, syntactic structure could be realized by relations among electromagnetic states rather than relations among shapes; as, indeed, it is in real computers.) This is the point of the Functionalist doctrine that, in principle, you can make a mind out of almost anything.
6. Which is not to deny that there are (ahem!) certain residual technical difficulties. (See, for example, part 4 of Fodor, *MOM*.) A theory of rationality (i.e., a theory of *our* rationality) has to account not merely for the 'semantic coherence' of thought processes in the abstract but for our ability to pull off the very sorts of rational inferences that we do. (It has to account for our ability to make science, for example.) No such theory will be available by this time next week.
7. Because I don't want to worry about the ontology of mind, I've avoided stating RTM as an identity thesis. But you could do if you were so inclined.
8. Like Dennett, I'm assuming for purposes of argument that the machine *has* thoughts and mental processes; nothing hangs on this, since we could, of course, have had the same discussion about people.
9. We can now see what to say about the philosophical chestnut about Kepler's Law. The allegation is that intentionalist methodology permits the inference from 'x's behavior complies with rule R' to 'R is a rule that x explicitly represents.' The embarrassment is supposed to be that this allows the inference from 'The movements of the planets comply with Kepler's Law' to some astronomical version of LOT.

   But in fact no such principle of inference is assumed. What warrants the hypothesis that R is explicitly represented is not mere behavior in compliance with R; it's an etiology according to which R figures as the content of one of the intentional states whose tokenings are causally responsible for x's behavior. And, of course, it's *not* part of the etiological story about the motions of the planets that Kepler's Law occurs to them as they proceed upon their occasions.

Chapter 2

1. If, however, Loar (*SCPC*) is right, then the commonsense taxonomy actually fits pattern B; i.e., common sense and psychology both individuate the attitudes narrowly and both respect supervenience.

   So far as I know, nobody has explicitly endorsed the fourth logically possible option—viz., that commonsense taxonomy is narrow and psychological taxonomy relational—though I suppose Skinner and his followers may implicitly hold some such view.
2. Notice that taking this line wouldn't commit Burge to a violation of *physicalism;* the differences between the attitudes of Twins and Oscars supervene on the (inter alia, physical) differences between their worlds. Or rather, they do assuming that the relevant differences between the linguistic practices in Oscar's speech community and Oscar2's are physicalistically specifiable. (I owe this caveat to James Higginbotham.)