1

Setting the stage

The relationship between mind and matter, mental states and physical states, has occupied the attention and imagination of the intellectually curious for thousands of years. In most cultures many people are officially committed to religious views that allow for the possibility of our surviving the total annihilation of our bodies. While the answers to questions of diachronic identity (identity through time) are not straightforwardly dependent on questions concerning synchronic identity (the identity of things at a time), it is tempting to think that there might be an intimate connection between the two. In particular, it is tempting to think that if we want to make plans for life after the destruction of the body, we better not tie the existence of our selves at a time too closely to the existence of our bodies at a time.

This book is yet another attempt to shed light on the nature of mental states. Despite the fact that volumes have been written on the subject, I believe that there are still contributions to be made. The issues are truly complicated. The primary reason that philosophical debate over the nature of the mental rages on is that, like so many other debates in philosophy, resolution of the issues involves a host of fundamental philosophical controversies on a wide range of philosophical topics. One can't coherently address problems in the philosophy of mind without working through issues in epistemology, philosophy of language, and broader metaphysical issues concerning the existence and nature of truth, states of affairs, facts, events,

properties, substances, and identity, both at a time and through time. Lurking in the background are critical metaphilosophical questions concerning the nature and appropriate methods of philosophical inquiry.

Consider, for example, the epistemological overtones of historical and contemporary debate concerning the nature of mental states. Certainly from Descartes on, philosophers have been trying to draw conclusions about the nature of the self and its properties by contrasting the ways in which one can know truths about oneself and the ways in which one can know truths about one's body. Descartes famously thought that the way to identify a secure starting point for knowledge was to include in its foundations only that which cannot be doubted. This methodological suggestion itself can be interpreted in importantly different ways. Descartes *could* have been talking about the psychological possibility of wondering whether or not a given assertion is true. More plausibly, however, he was concerned with what our justification allows. On this interpretation, S's belief that P allows for the possibility of doubt when the epistemic justification S possesses for believing P is at least consistent with P's being false. In Descartes's search for infallibly justified belief, he hit upon our belief that we exist. The justification *we* have for believing that we exist is not consistent with that proposition's being false. The proposition that I exist, therefore, belongs in my foundations.

While I have foundational knowledge of my own existence, I have no foundational knowledge of the existence of anything physical. The reconstruction of Descartes's view is a bit tricky since he eventually retracts many of the arguments he gives (including his famous dream argument). But *initially*, Descartes appeared to argue as follows. The justification we have for believing any propositions asserting the existence and describing the properties of physical objects is consistent with those propositions being false. Consider, for example, dreams.

In a vivid dream I could have precisely the same justification for thinking that there is a table before me as I would have had were I veridically perceiving a table. But the justification I have in a dream is obviously consistent with the table's not being there, and so the justification I have in veridical perception leaves open the skeptical possibility as well. Or consider Descartes's evil demon possibility – the inspiration for many a plot of films and TV series.[1] The qualitatively same sensations that come to us in veridical perception, Descartes argued, instead could have been "planted" in us directly by a very powerful being bent on deceiving us with respect to our surroundings. If deceived in this way, we would have the same justification for believing what we do about our surroundings as we would have had were our experiences veridical. The justification we have, therefore, does not preclude the possibility of error.[2]

The argument does presuppose a version of the view that we might call epistemological internalism.[3] On that view, the

[1] *The Matrix*, *Total Recall*, *What Dreams May Come*, and seemingly every other episode of the *Star Trek* series (old and new).

[2] In a popular piece discussing *Matrix*-like skeptical scenarios, Chalmers (2003) argues that if we had always been living "in a matrix" we shouldn't take our beliefs about the external world to be false. Rather we would have been (successfully) talking about a reality different in nature from what we took it to be. The argument rests on what I have called a causal theory of objects – we'll be discussing a view like this later in the book.

[3] There are actually many different versions of internalism now. The one I sketch here might usefully be called internal state internalism after its key idea that the epistemic status of a belief supervenes solely on the internal states of the believer. Even here, however, we need a much more careful account of what makes a state internal. Most self-proclaimed internalists of this sort will still want to allow that a given state is internal if it involves a relation between the subject and some abstract entity like a universal. And the kind of epistemologist who thinks that one can be directly acquainted with the surfaces of physical objects might also want to count that relational state of affairs as internal. Internal state internalism must be distinguished from various versions of access internalisms. The access internalist insists on the idea that for something X to justify S in believing P, S must have actual or potential access to X (a relatively weak requirement) or actual or potential access to the fact that X is a justifier for the belief that P (a much stronger requirement). Access requirements on

justification S possesses for believing a given proposition P at t is solely dependent on S's internal states at t. The internal state internalist's slogan is that necessarily, if S and R are in the same internal states, then whatever justification there is for S to believe some proposition P there is for R to believe P. If one embraces the internalist's doctrine and searches for the justification we have for believing sundry propositions describing our physical world, it will seem obvious that that justification is consistent with our physical object beliefs being false.

The conclusion is so difficult to resist because of what we take to be the causal story surrounding experience. Ultimately, we must rely on sense experience to reach any conclusion about the existence of physical objects and their properties. But the immediate causes of sense experiences, we believe, are events occurring in the brain. We believe that there are physical objects, and that when we are before those objects under certain conditions, a long and complex causal chain begins, starting with changes that occur in and around the surface of the object and terminating in a change in the brain which either is (according to the mind/brain identity theorist) or immediately causes (according to the dualist) sensation. In the case of visual experience, we think that light reflects off the surface of an object eventually hitting the retina of the eye and effecting changes there. Of course, changes in the eye aren't going to result immediately in visual experience. The "information" must be carried all of the way to the relevant region of the brain before the sensation occurs. Given that sensation is the immediate effect of the brain state, the natural thought, then, is that we could, in principle, "break into" the causal chain anywhere and if the causal sequence terminates in the same place,

justification are tricky – they almost always lead one to flirt with vicious regress. See Fumerton (1995, Chapter 3 and 4) for attempts to define more carefully different versions of internalism. And see Bergmann (2006, Part I) for a sustained attack on all forms of epistemic internalism.

the same sensation will occur. More often than not, contemporary epistemologists discussing skeptical scenarios replace evil demons with mad neurologists who have stolen brains and who are manipulating those brains so as to produce massive vivid hallucination.

Again, the upshot of all this for Descartes was the critical observation that we can know the truth that we exist without knowing any truths about the physical world. One might worry that the arguments for this thesis seemed to presuppose that we have knowledge that allows us to describe the character of both veridical and non-veridical experience. So in his more incautious moments you might find Descartes making a claim about the similarity between the experiences of a dream and the experiences of waking life. And the alert philosopher might pounce on that claim. To know that the experiences of dreams are like the experiences of waking life we would need to know what the experiences of waking life are like. But that would presuppose that we *are* able to know when we are awake. Similarly, any skeptical argument that proceeds from premises describing the causal history of perception will explicitly invoke information about the physical world, when it is the possibility of getting such information that is coming under skeptical attack. But it would be a mistake to suppose that there is anything problematic in the appeals to possibility that the Cartesian will invoke in attempts to cast doubt on the strength of the justification we have for believing propositions about the physical world. Descartes argues from the *possibility* of dreaming and hallucinating. One can certainly take advantage of the fact that one's audience has beliefs that certain possibilities are actual in reaching the conclusion that they are possible – the surest proof that something is possible is that that it is actual. But the conclusion that vivid dreams and hallucination are possible would be true whether or not they ever occurred.

If one grants Descartes's claim that one can know that one exists in a way quite different from the way in which one can know any truths describing the physical world, the question then becomes whether or not one can draw any *metaphysical* conclusions about the nature of the self and its relation to the physical from these epistemological premises. Descartes appeared to conclude that since it is epistemically possible that he exists without his body, then he must be something distinct from his body. But critics, even in Descartes own time, immediately questioned the inference.

1.1 LEIBNIZ'S LAW AND THE CARTESIAN ARGUMENT FOR DUALISM

If there is any principle that is, or should be, unproblematic concerning identity it is one half of Leibniz's law, the indiscernibility of identicals. In its metaphysical form the principle states that for any x and any y, if x is identical with y at t then there is only one thing picked out by "x" and "y" and whatever is true of that thing at t is true of that thing at t.[4] Leibniz went on to claim much more controversially the converse: that for any x and any y if whatever is true of x at t is true of y at t and vice versa, then x is identical with y at t. In what follows I'm concerned only with the indiscernibility of identicals. And as I said the principle should be uncontroversial. If x and y are

[4] The principle gets much more controversial if we don't restrict it in this way to identity at a time. One might suppose that if one argues that x at t1 is identical with y at t2 only if there is nothing true of x at t1 that isn't true of y at t2, one will have, implausibly, or at least controversially, eliminated the possibility of a single thing changing over time. One can attempt to avoid that conclusion by insisting that property ascription always be itself time relativized. So I can be identical with the boy who was only 4 feet tall in 1955, because I still have the property of being 4 feet tall in 1955. Temporally relativized properties can't be lost. If one insists that it is ad hoc to insist on time-relativized property ascriptions, one can simply assert that it is only a time-relativized version of Leibniz's law that is true.

identical at t then we have only one thing x/y that has properties and stands in relations to other things at that time. The principle amounts to the claim that if a thing has a property it has the property! Indeed, the only reason that anyone might wonder about the principle is if it gets confused with another superficially similar linguistic principle concerning the substitutivity of co-referential expressions. One might find initially attractive the suggestion that if "x" and "y" pick out precisely the same entity, then we ought to be able to substitute "y" for "x" in any assertion without changing the truth value of the assertion. And we might suppose that this is true because if "x" and "y" are co-referential then the substitution won't make a difference to what is asserted. So, for example, if "my cat" and "Duke" pick out the same cat and it is true that Duke is gray then it is true that my cat is gray. But the principle that one can substitute co-referential expressions *salva veritate* in all contexts seems to be false. Consider all of the following:

(1) "Duke" begins with the letter "D."
(2) Sue believes that Duke is gray.
(3) It is a necessary truth that if Duke exists and is gray then Duke is gray.

It we substitute for "Duke" "my cat" in (1) through (3) (its first occurrence in (3)), we will change the truth value of (1) and (3) and might change it in (2) – we need only imagine a situation in which Sue doesn't believe that Duke is my cat and further believes that my cat is orange.

Now there are all sorts of complications that arise as we think more carefully through the above crude discussion. It is not clear that all, or even any, of (1) through (3) involve genuine instances of failure of *referential* substitutivity. A great deal depends on how we understand the technical philosophical concept of reference and how we locate the occurrence of a referring expression. So although it might sound initially strange, some philosophers

(e.g. Quine, 1966, Chapter 15), will deny that the expression "Duke" occurs in (1). Admittedly, the expression "'Duke'" occurs in (1) but neither "my cat" nor "'my cat'" refers to the same thing as does "'Duke'." It makes no more sense to try to substitute "my cat" for "Duke" in (1) than it would make sense to substitute "automobile" for "car" in "My carpet is dirty." All we would get is the gibberish, "My automobilepet is dirty." Whether one takes (2) and (3) to be clear cases exhibiting referential opacity (contexts in which substitution of referring expressions can change truth value) might depend on the verbal question of what one takes "referring" to be. So as we will discuss later, Russell and his followers will almost always take ordinary names to be disguised definite descriptions. "Duke" might mean something like "the cat I bought when I was 15." "The cat I own" is already an undisguised definite description. At least some of the time Russellians will balk at claiming that definite descriptions refer – indeed they will claim only that the definite description makes a contribution to the meaning of the entire sentence in which it occurs, a sentence which *on analysis* won't contain anything that looks very much like the definite description. So "The F is G" becomes "There is one and only one thing F which is G." Still, the definite description in English clearly picks out some object when it is used successfully, and we need an expression to capture the relation between a definite description and the object it singles out. We can say that a definite description can successfully *denote* some object *a* and that it does so just when that object takes the value of the variable that appears in the logically perspicuous "analysis" of the statement. A hard-core Russellian might reserve the term "reference" for the relation that holds directly between a pure name and an object for which it stands (a name whose meaning really is its referent). No one thinks (or should think) that the *meaning* of a definite description is its denotation. We'll have much more to say about all this in Chapter 5.

To avoid these terminological matters, we can distinguish the metaphysical principle of the indiscernibility of identicals from controversial principles concerning the substitutivity *salva veritate* of either co-referential or co-denotational expressions. Again, there are no counterexamples to the principle of the indiscernibility of identicals. Apparent counterexamples are always counterexamples to various principles concerning substitutivity of linguistic terms. So, to illustrate with a much-used example, one might suppose that one can construct counterexamples to Leibniz's law by thinking about Superman, Clark Kent, and Lois Lane. Superman has the property of being believed by Lois Lane to be really strong, while Clark Kent lacks that property. But if we suppose that there is a property of being believed, what has it? Lois Lane does believe the proposition that Superman is really strong and it is not true that she believes the proposition that Clark Kent is really strong. But all that shows is that there is a property of the one *proposition* that is not a property of the other (as we'll see in a moment, some of this gets decidedly more complicated). We haven't discovered a property of Superman that isn't a property of Clark Kent.

Of course, there is a so-called *de re* reading of descriptions of belief. They are sometimes marked in ordinary language (or at least ordinary philosophical language) as follows: Lois Lane believes *of* Superman that he is really strong, but doesn't believe *of* Clark Kent that he is really strong. The classic indicator that a description of belief is a *de re* description is that the relevant description is referentially transparent – one can substitute co-referential (or co-denotational) descriptions without altering the truth of the statement. The following is a statement that in most contexts wouldn't strike you as all that odd but could only be a plausible claim if it gets a *de re* reading: Henry Hudson believed that Hudson's Bay was a passage to the Orient. Upon finding the huge body of water we call

Hudson's Bay, Hudson did conclude that he was in the much sought after Northwest Passage. To his great disappointment (followed relatively soon by his death) he was wrong. But however mistaken Hudson might have been, he presumably wasn't both clairvoyant and crazy. He didn't believe that the body of water he was in was a bay, and he didn't believe that it was going to be named after him.

Many philosophers think that they can handle the distinction between *de dicto* and *de re* reports of belief using scope distinction, at least when the relevant denoting expressions that occur in the belief report are definite descriptions. So when I say that S believes that the F is G, the *de dicto* reading is:

(1) S believes that the one thing that is F is G.

The *de re* reading is:

(2) The one thing that is F is such that S believes that *it* is G.

I don't think that this is the way to handle the distinction for the simple reason that (2) really doesn't identify the content of S's belief. It is more plausible to suppose that the *de re* report leaves open the proposition S believes other than to stipulate that it is a proposition whose subject content denotes a thing that is in fact F and whose predicate is G. The rough translation then becomes something like the following:

(3) There is just one thing that is F and there is some proposition P such that S believes that P and the subject concept of P denotes a thing that is the one thing that is F and the predicate concept of P picks out G.

Notice that to acknowledge the obvious truth that there are *de re/de dicto* ambiguities in *descriptions* of belief does not have any obvious implications for the ontological question of whether there are two quite different sorts of beliefs – beliefs that have as their content propositions and beliefs that