1

The Multiple Realizability Thesis: Significance, Scope, and Support

Several decades ago philosophers began to consider the possibility that the mind is multiply realizable: that it is possible for minds to be built in various distinct ways. Multiple realizability is an obvious idea when applied to things other than minds. Watches seem to be multiply realizable—a watch may be either analog or digital, for instance. Similarly, corkscrews appear to be multiply realizable—the double-lever corkscrew relies on a rack and two pinions to do its job, whereas the waiter's corkscrew makes use of a simple lever (figure 1.1).

Applied to minds, the claim of multiple realizability amounts to the thesis that what's true for watches and corkscrews is true as well for minds. For convenience, I shall call the thesis of multiple realizability when applied just to humanlike minds the *multiple realizability thesis* (MRT). If we suppose that the human mind consists of various cognitive capacities—perception, memory, language comprehension, attention, problem-solving abilities—and that these capacities in human beings are identifiably distinct from the way they might appear in other organisms (surely *human* perception, memory, language comprehension, etc. differ from similar capacities in other primates and, if present at all, in other animals), then we can speak of a uniquely human psychological profile. MRT, as I shall be using the term, is the thesis that a mind with this uniquely human psychological profile can be built in distinct ways. The claim I shall defend is that MRT is perhaps false and is in any case far from the well-supported thesis that philosophers have traditionally taken it to be.

In my effort to address the question of the truth of MRT, I shall be interested in assessing the *likelihood* of MRT. As I noted in the

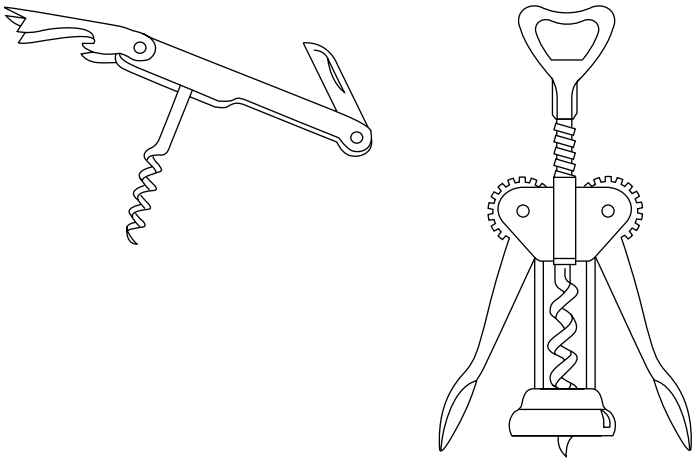

Figure 1.1
Two types of corkscrew: waiter's corkscrew (left); double-lever corkscrew (right).

preface, I intend "likelihood" in its technical sense. Likelihood is one way to measure the support that evidence provides for a hypothesis. In the case at hand, the hypothesis of interest is MRT. I shall be concerned mainly with the kinds of predictions that MRT makes, and I shall be contrasting its predictions with that of a competing hypothesis: the *mental constraint thesis* (MCT). According to MCT, humanlike minds are not multiply realizable, or, at least, many humanlike mental capacities are not multiply realizable. In the work ahead I shall be arguing that MCT and MRT, given certain background assumptions, make different predictions and that on the basis of these distinct predictions it should be possible to determine which of the two has greater likelihood. Crucially, I will *not* be trying to *prove* that MRT is true or false. Rather, I shall be considering evidence and arguments that, I think, make a plausible case that MCT is often the better supported of the two hypotheses, and so MRT is not the unimpeachable claim that so many have assumed it to be. More modestly, I hope to make evident the complexity of issues that are involved in the thesis that minds are multiply realizable.

Of course, one can believe that MRT is true but doubt that the human mind *actually* has multiple realizations (although, as we shall see, some

do believe that human minds are in fact multiply realized to some degree at least). Just as something may be flammable but never actually combust, so a kind of mind may be multiply realizable although it is, in fact, never realized in more than one way. So, it is consistent with the truth of MRT that there currently exists only one kind of realization of humanlike psychology and that there never will exist another. Nevertheless, the bare possibility of multiple realizability establishes MRT as a thesis of philosophical significance.

The remainder of this chapter is dedicated to three tasks. First, I explain in more detail why philosophers have invested so much in the truth of MRT. Second, I draw some distinctions to clarify more exactly how I shall understand the scope of MRT. Finally, I critically discuss the conceptual and empirical arguments that philosophers have made in support of MRT.

1.1 The Significance of the Multiple Realizability Thesis

MRT is important for what it claims about the relationship between the mind and the brain; but more generally, multiple realizability raises issues about how various scientific theories are related to each other. In this latter context, philosophers impressed with the apparent implications of MRT for the relationship between a theory of the mind and a theory of the brain have sought to argue that the multiple realizability of, for instance, various biological kinds has profound implications for the relationship between biology and physics (e.g., Kitcher 1984; Rosenberg 1985). Whereas I will have much more to say about both MRT in particular and the implications of multiple realizability more generally in the pages to follow, a few introductory words at this time are appropriate.

Regarding the first point—the consequences of MRT for our understanding of the mind–brain relationship—it is a simplification but not an egregious one to see prior to the emergence of MRT a division among philosophers over the nature of the mind–brain relationship. On the one hand were the dualists, who conceived of minds and brains as fundamentally different kinds of substances. The brain is a physical substance and thus endowed with physical properties like mass, location,

temperature, and so on. In contrast, the mind, according to the dualist, is without mass or location, and so on. It was this conception of minds and bodies that Ryle (1949) mocked with his ridicule of the "ghost in the machine." To be sure, dualists would acknowledge that goings on in the brain may be reliably correlated with happenings in the mind, but because the mind and brain are on their view so utterly different, their explanations for this correlation tended toward the mystical, the incoherent, the theistic, or the desperate.

On the other hand were the type-identity theorists (Place 1956; Smart 1959) who brushed aside centuries of dualist machinations with a simple and bold proposal: the mind and the brain are one and the same thing. On this view, the difficulty of explaining why and how events in the mind and brain exhibit such precise correlations completely vanishes. According to type-identity theorists, the idea that some mental event, say, pain, is merely correlated with a distinct event in the brain, say, the firing of C-fibers, rests on a confusion. It would be the same confusion that is obvious in the following question: why is it that whenever water is present there also is H_2O? The question is asking for an explanation of the ineluctable correlation between water and H_2O, but of course if this correlation is assumed to hold between distinct things then one has made a mistake. Water *is* H_2O. The terms "water" and "H_2O" are simply two labels that refer to one thing. It is on this model that type-identity theorists wished to think about the connection between the mind and the brain. Speaking properly, pain and C-fiber firings are not simply correlated, they are identical.

Whereas the identification of the mind with the brain provided an easy way out of the difficulties dualists confronted when trying to explain how an entity that exists outside of space could causally interact with a body in space, type-identity theorists faced a hard, and in retrospect, obvious problem of their own. As any fan of science fiction or any advocate of artificial intelligence will tell you, it is possible to build minds out of different kinds of stuff. One need go no further than the movie theater to witness minds composed of silicon chips, flocculent masses of green noodles, or gelatinous blobs of phosphorescent muck. No doubt these varieties of creatures do not in fact exist, but the examples illustrate nonetheless the coherence of the idea of humanlike minds that are not

identical with humanlike brains. Unlike water, which, if it is not H_2O is not water, minds that are not brains are readily conceived.

Thus, MRT is significant for the stance it offers on the relationship between mind and brain. Minds and brains, MRT teaches us, cannot stand in a relation of simple identity. Things are more complicated. Indeed, much of contemporary philosophy of mind has been devoted to developing plausible analyses of the mind–brain relationship in light of MRT. More specifically, discussions of token identity, according to which particular mental properties (e.g., my pain) are identical to particular physical properties (e.g., my C-fibers firing), take the truth of MRT and thus the rejection of type-identity theory as their starting points. At stake in these discussions is an understanding of how minds and brains are related, why minds seem to depend on brains, how mental events can cause physical events, and so on, given that "mind" and "brain" are no longer conceived of as two labels for the same thing.

I mentioned that interest in MRT has extended beyond the important but narrow topic of the mind–brain relation. This is because MRT also has consequences for how one is to understand the relation between theories of mind and theories of those biological, chemical, and physical processes on which mental events seem to depend. Prominent philosophers of science in the mid–twentieth century conceived of science as consisting of theories at various levels, where the kinds at higher levels, like psychology and economics, depend in some way on kinds in lower levels, like biology, chemistry, and, ultimately, physics. In virtue of this dependency, it would be possible, these philosophers believed, to derive laws of, say, psychology, from laws of biology, which in turn could be derived from laws of chemistry and, finally, laws of physics. The derivation of laws of a higher-level theory from laws of a lower-level theory constitutes an *intertheoretic reduction*, and it was the possibility of such reduction that promised to unify the myriad levels of science.

Appealing as this ideal of unification might seem, it effectively relegates higher-level sciences like psychology to rest stops on the road to a final destination. Psychological laws, on the unificationist's conception of science, are in principle eliminable in favor of those physical laws from

which they derive. We continue to value psychology as a way to understand the mind, but only because we don't yet know enough to explain mental phenomena in physical terms. Presumably, the unificationist would hold, the days of higher-level theorizing are numbered; ideally, colleges of letters and science will one day contain only a single department of physics.

It was in response to this austere vision that Fodor (1974) developed the implications of MRT. Just as MRT prevents an identification of mental kinds and neural kinds, and so stands in the way of a reduction of mind to brain, so, more generally, does the multiple realizability of kinds in higher-level sciences challenge the dream of unification. As long as it is possible that kinds in any higher-level science are realizable in different lower-level kinds, it will be false that the higher-level kind is coextensive with a particular lower-level kind, and thus it is hopeless to try to derive higher-level laws from lower-level laws. Or, at any rate, that is the reasoning that led Fodor to see multiple realizability as a kind of salvation for higher-level sciences (but see Sober 1999b for an argument that multiple realizability need not be taken to have these consequences for reduction). So long as the multiple realizability of higher-level kinds remains a possibility, higher-level sciences must retain their autonomy. Whether this verdict is in the end correct, it is uncontroversial that philosophers have by and large *perceived* multiple realizability as securing the autonomy of higher-level sciences, and for this reason the significance of multiple realizability in philosophy of science has been immense.

1.2 The Scope of the Multiple Realizability Thesis

I have said enough, I hope, to motivate the study of multiple realizability pursued in the chapters that follow. I will now turn to clarifying the scope MRT assumes in the following chapters. Philosophers, for whatever reason, have rarely characterized MRT in precise terms. However, I believe that any evaluation of MRT must take care to draw several distinctions. Spending time now to advance these distinctions will help situate and motivate the various arguments and maneuverings in the chapters ahead.

The first distinction concerns the extent of MRT: just how much multiple realizability must one accept before proclaiming MRT to be true? For purposes of answering this question, I shall adopt with slight revisions Polger's (2002, pp. 146–147) useful taxonomy. Polger finds in the literature four conceptions of MRT:

Weak MRT: At least some creatures that are not exactly like us in their physical composition can have minds like ours.

SETI[1] MRT: Some creatures that are significantly different from us in their physical composition can have minds like ours.

Standard MRT: Systems of indefinitely (perhaps infinitely) many physical compositions can have minds like ours.

Radical MRT: Any (every) suitably organized system, regardless of its physical composition, can have minds like ours.

Polger notes that standard and radical MRT seem to receive most discussion and allegiance in contemporary philosophy of mind. In my view, these theses are prima facie incredible; and even if one does not agree with this sentiment, it is surely true that the burden falls on those who endorse them to say why they find them plausible. Weak and SETI MRT, on the other hand, are less easy to dismiss with a simple "prove it." Many might be inclined to see these hypotheses as plausible on their face. But more important, some philosophers believe that there is currently at hand evidence to support them. Accordingly, much of this book is spent defending a view about how one should interpret this evidence. In the end, I hope to have diminished significantly the support for a conception of MRT that falls somewhere between weak and standard MRT.

The label "weak MRT," while clear in its intent, remains vague. Just how "exactly like us" should qualify as *exactly like us*? As we will see in the following chapter, there are clearly some differences in realization that one should never want to claim suffice for a case of multiple realization. A moment's reflection should make obvious that vagueness is likely to permeate any characterization of MRT. After all, MRT is a thesis that

requires one to talk about similarities of various sorts—similarities in kinds of realization and similarities in the kinds that are realized. I shall have much more to say about these matters in the chapters that follow. However, I think it is appropriate at this point to acknowledge that MRT is not an all-or-nothing affair. Whether one takes MRT to be true will depend on how much multiple realizability one requires (i.e., does one require weak, SETI, standard, or radical MRT?); and how much multiple realizability there is will in turn depend on how stringent one's requirements for similarities and differences in realization are. My hope is that by setting my sights on a fairly limited form of MRT—as I said, something between weak and standard MRT—and by employing a conception of similarity that seems to match the grain of description that our best sciences of the mind find useful, I will be able to challenge that statement of MRT about which one should care the most.

A second set of distinctions that an evaluation of MRT makes necessary concerns the nature of possibility. I noted above that MRT is the claim that it is possible to build a humanlike mind in different physical ways. But it is standard philosophical fare that "possible" might refer either to anything that is consistent with laws of nature or anything that is consistent with logical truth. A perpetual motion machine is impossible in the first sense but not in the second. That is, various laws of nature imply that no one will ever build a perpetual motion machine, but there is no logical contradiction in the idea of, say, a wheel that never ceases to spin. On the other hand, because triangles are by definition three-sided figures, a geometer in search of a four-sided triangle is clearly wasting her time. Such a figure is impossible not just in the first sense, but in the second as well, because the existence of a three-sided figure with four sides is logically incoherent. The first kind of possibility, because it is concerned with limits that laws of nature impose, is often called *nomological* or *physical* possibility. The second kind of possibility is known as *logical* possibility.

Given these two kinds of possibility, it should now be evident that MRT is ambiguous. Should the claim be taken to imply just the logical possibility of multiply realizable minds or, more ambitiously, the nomological possibility of such things? Pretty clearly, it must be nomological possibility that MRT assumes. This is apparent from the empirical nature

of the evidence that proponents of MRT muster in its defense. In contrast, no one, as far as I know, has spent much energy trying to defend the *logical* possibility of MRT. Indeed, of the four conceptions of MRT I mentioned above, *all* are logically possible, and so whatever disagreement there is between advocates of the four types, it cannot be over the logical possibility of each type. Accordingly, if the question of the possibility of humanlike minds instantiated in nonhumanlike brains is to have any oomph, it should boil down to empirical considerations rather than claims about how to understand the concept *mind*.

There is yet a third conception of possibility that will figure prominently in my discussion of MRT. Narrower than nomological possibility is circumstantial, or historical, possibility. Circumstantial possibility recognizes the importance of initial conditions in predictions of what can happen. Some things that are nomologically possible in some places or at some times will not be nomologically possible at other places or at other times because of differences in circumstances. Thus, for instance, it may well be that life is possible only on Earth. Were this true, however, it would not be because it is nomologically impossible for life to evolve elsewhere. Presumably, life could evolve elsewhere if other planets had conditions similar to those on Earth. In saying that life is possible only on Earth, one is asserting that the initial conditions necessary for the evolution of life are present only on Earth: were they present elsewhere, life would be possible elsewhere.

In addition to the limitations on possibility that initial conditions impose, circumstantial possibility refers to limitations on what is possible as a result of historical contingences.[2] For instance, it might now be possible for me to arrive in Chicago in time for dinner. It would take about three hours to drive to Chicago from my home in Madison, my car is working fine, and it's early in the afternoon. However, if on the way I become lost, or I run over something that punctures a tire, or I am slowed by traffic, it may become impossible for me to arrive in Chicago in time for dinner. This is an impossibility that results (we can suppose) not from conditions that are present at the start of my trip, but rather from contingencies that develop along the way.

One helpful way to illuminate further the distinction between nomological and circumstantial possibility is to consider a thought experiment

Stephen Jay Gould (1989a) suggested. Imagine, Gould requests, that the tape of life were played over again. Starting with the same initial conditions that were present at the time of the Earth's origin, if life were to evolve all over again, would the life-forms that developed bear any resemblance to the life-forms that have actually populated the Earth? Gould's answer is an emphatic "no." Because of the countless accidents involved in evolutionary processes—accidents ranging from the dramatic, such as collisions with asteroids, to the mundane but just as significant, such as the inadvertent drowning of an organism that, had it the chance, would have given birth to a new species—the odds that anything remotely resembling a human being (or any other actual organism, presumably) would evolve again are infinitesimally small. Here Gould is betting that the initial conditions by themselves leave open the possibility that many different forms of life could evolve.

In the context of MRT, Gould's claim might be put like this. Imagine that the tape of life is played over again and somewhere along its length there evolves a being with a psychological profile just like a human being's. Would the being's mind be realized in anything like the way that a human being's mind is realized? The truth of MRT requires that there be flexibility in the manner in which this new organism's mind is realized. If the tape of life were played over and over again, each time producing a being with a humanlike mind, the truth of MRT predicts that on at least some occasions the humanlike mind would be realized differently than it is realized in human beings. Like Gould's claim about the variation in life-forms that can evolve from identical initial conditions, the version of MRT in which I am interested bets that initial conditions on Earth are consistent with variation in the kinds of things that could realize a humanlike mind. In contrast, if MRT is false, one would expect to see the same kind of realization each and every time a humanlike mind evolves.

Yet, even supposing that the tape of life experiment does disconfirm MRT on Earth, it reveals nothing about the likelihood of MRT elsewhere. Just because the limited number of materials on Earth might make it impossible to build more than only a few kinds of electrical conductors, so too initial conditions on Earth might make it possible to realize humanlike minds in only a few different ways. However, all this is con-

sistent with the possibility that elsewhere in the universe there are materials that would, if put together correctly, realize humanlike minds, just as there may well be substances capable of electrical conductance that are nowhere to be found on Earth.

My main target will be the possibility of MRT here on Earth. This is in some sense inevitable given our ignorance of the kinds of initial conditions that are necessary for the evolution of thinking things like ourselves. However, I believe that questions about the possibility of MRT just on Earth (what I shall sometimes call *terrestrial MRT*) are of sufficient interest to warrant its careful study. Is it just an accident that the human brain has the properties that it does, or could the evolution of humanlike minds result in terrestrial brains very unlike those that human beings possess? Even though limited to what can happen here on Earth, this is a question that must surely attract the interest of anyone seeking to understand why the brain has the structure it does, just as answering questions about the shape of the eye's lens would require one to understand why eye evolution has taken one course rather than another. That is, an evaluation of terrestrial MRT seems as necessary for answering some questions about the relationship between the human mind and the human brain as a study of the multiple realizability of eyes would be for answering some questions about the relationship between the capacities of the eye and anatomy of the eye.

Moreover, those who hope to build an artificial intelligence should be very curious about the prospects of terrestrial MRT. If there turns out to be good reason to believe that terrestrial MRT is true, then clearly the search for alternative ways to build a humanlike mind is not doomed from the start. Of course, if the evidence goes against terrestrial MRT, this would not show that it is impossible to build a realization of a humanlike mind that differs significantly from the brain. After all, the processes of evolution face constraints that the engineer does not. Most basically, an AI engineer approaches the problem of building a mind with the advantage of foresight. The engineer has an idea about how her finished product should behave and sets about looking for the most efficient way of realizing her goal. Evolution, on the other hand, operates without the benefit of a blueprint. The products of evolution are often poorly designed from the perspective of an engineer: they may be unnecessarily complex,

they may be so constrained by earlier stages of their evolution that they can never assume a more optimal design, and so on.

Still, despite these differences between "natural" design and engineered design, AI could benefit much from considering the question "Why has nature chosen to realize a human mind in this way rather than some other?" It seems quite reasonable that the answer to this question should shape the approaches that AI researchers take toward at least some of their goals. The more the answer tends toward a rejection of terrestrial MRT—the more the answer inclines toward a sort of nomological necessity in the mind–brain relationship given the kinds of conditions present on Earth—the more reason for caution in designs that depart radically from that on which nature has settled.

There might appear to be a tension between my decision to focus on terrestrial MRT and my claim to be offering a challenge to something between the weak and standard MRTs that Polger describes. As defined, these theses do not exclude the possibility that humanlike minds can be realized in kinds of material that are not present on Earth. Indeed, SETI MRT, which falls within the range of weak and standard MRT, is quite explicit in its suggestion that minds like ours can evolve from starting conditions that differ from those on Earth. How can a challenge to terrestrial MRT be made to work against these more inclusive statements of MRT? It cannot, unless some of the arguments against terrestrial MRT are arguments about not just the circumstantial possibility of MRT but about its less confining nomological possibility as well. As we will see, it is no easy matter to determine whether the realization of the human mind is as it is because of initial conditions and historical contingencies or because nature simply cannot build a humanlike mind in more ways than one. Indeed, this is just one of many issues we will come across that has been ignored in discussions of MRT despite its obvious importance. Moreover, the muddy empirical flavor of this question suggests that it is one for the scientists to answer. For now, I shall simply claim boldly that at least some of the points I make in chapter 4 about the connection between mental and neural properties might well reflect nomological rather than circumstantial or historical possibilities. Insofar as this is true, my arguments call into question a statement of MRT that falls within the weak–standard range. If I turn out to be wrong about this, I

am content to limit my challenge to terrestrial MRT, which, as I have noted, is a thesis of great interest in its own right.

1.3 Conceptual Arguments for the Multiple Realizability Thesis

In contemporary philosophy of mind, the idea that minds, or mental states, are multiply realizable emerged in the 1960s, primarily in the writings of Hilary Putnam and Jerry Fodor.

The arguments Putnam and Fodor made are now part of every philosopher of mind's basic tool kit, but a close examination of their original formulation provides a useful stepping stone to a thorough examination of the claim that minds are multiply realizable. For the sake of convenience, I will group these arguments into conceptual and empirical versions, leaving discussion of the empirical arguments for the next section. I do not wish to claim that there is a sharp distinction between conceptual and empirical considerations. I use the labels merely because the evaluation of some of these arguments seems to depend less immediately on the weight of empirical evidence than does the evaluation of others. As we will see, the conceptual arguments do not work so well in support of the multiple realizability thesis. The empirical arguments seem on their face more compelling than the conceptual ones. Yet, I shall argue in the next chapter, on a reasonable analysis of how to understand the concept of *multiple realization*, even the best empirical evidence that philosophers cite on behalf of MRT is inconclusive.

1.3.1 Turing Machine Functionalism

Hilary Putnam's arguments against the identity theory (Putnam 1960; 1967) are the modern *locus classicus* of the idea that minds are multiply realizable. As I have already noted, identity theorists argued that minds and brains are strictly identical, in the same way that water is strictly identical with H_2O and lightning is strictly identical with electrical discharges. Motivating the identity theorists was a desire to characterize the mind in a way that avoided nomological danglers (Feigl 1958)—ontologically peculiar entities that fall outside the scope of physical laws. If minds or their parts consisted of nonphysical substances, then the mind's relation to the body, as well as to the rest of the world,

would be irremediably mysterious. If, on the other hand, the mind simply is the brain, then the secrets of the mind, if not immediately apparent, would at least now be in a position where we could get to them. That is, mind–brain identity locates mental properties within an accessibly empirical domain.

Putnam's response to the identity theorists rests on construing the mind as a collection of Turing machine states. Turing machines are theoretical constructs (they do not actually exist because they are endowed with an infinite storage capacity) defined by two functions: one that takes inputs and states to outputs and another that takes inputs and states to other (or the same) states (Block 1978, 1980a,b). A *Turing machine table* is simply a list of instructions that defines these two functions. Turing supposed that numerals would serve as the inputs and outputs to a Turing machine. Thus, given the state of Turing machine *X*, a numerical input to *X* would cause it to produce some numerical output and then either to change state or remain in the same state. Any Turing machine that can be described by the same table that defines the operations of *X* is "equivalent" to *X*. In fact, because inputs and outputs to a Turing machine are defined purely syntactically (i.e., with no regard for what the symbolic inputs and outputs stand for), one can say that *X* and all Turing machines that share *X*'s machine table are realizations of the same abstract type of machine.

Putnam's suggestion was that mental states could be conceived as analogues to Turing machine states.[3] In place of the numerals that serve as inputs and outputs to Turing machines, Putnam substituted sensory stimulations and behavior. Mental states, on this model, are defined by a machine table that specifies functional relations between sensory stimulations, states, and behavior. To be sexually jealous, for instance, is to be in a state that, given the observation of one's mate flirting with someone else, causes one to assume a threatening posture toward the individual and also to *suspect* that the mate has been or is interested in having sex with this individual, and, perhaps, to feel *humiliation* that one's mate would prefer the attention of someone else. Similarly, the mental state of humiliation receives an analysis in terms of the mental states and behavior that it will produce when combined with various sensory stimulations, and so on for all mental states.

Thus, the first step in Putnam's response to the identity theory—in the development of his idea that minds are multiply realizable—is a conception of mental states as Turing machine functional states. The next step presumably follows immediately. Just as a Turing machine table provides the means by which to classify various Turing machines as realizations of the same kind, a given mental machine table suffices to say when various physical organizations are of the same mental kind. Two physical systems are of the same mental kind when they contain states that obey the same instructions—that fall in the domains and ranges of the same functions. And, as Ned Block observes, "[i]f we could formulate a machine table for a human, it would be absurd to identify any of the machine table states with a type of *brain* state, since presumably all manner of brainless machines could be described by that table as well" (1980b, p. 178). Multiple realizability appears to follow quite naturally and immediately from the suggestion that mental states are Turing machine-functional states.

In deciding whether this argument for MRT actually works, it is important to understand its force against the identity theorists. Why not take Putnam's objection to mind–brain identity as simply a friendly amendment to the identity theory? After all, identity theorists were probably less interested in arguing for a strict identity between the mind and the brain than they were in making the mind safe for materialism. As Smart (1959) remarks in the course of considering the possibility that sensations are nonphysical, "for various reasons I just cannot believe that this can be so. That everything should be explicable in terms of physics . . . except the occurrence of sensations seems to me to be frankly unbelievable" (1959, p. 142). If, as I've claimed, the identity theorist's motivating desire was to make the mind explicable in physical terms, then Putnam's claim that minds are multiply realizable in various physical organizations gives identity theorists precisely what they want—though perhaps not for the reasons that they supposed.

Yet, it is in fact a gross misunderstanding of Turing machine functionalism to think that it is materialist in spirit. As Putnam adamantly asserts, "the functional-state hypothesis is *not* incompatible with dualism! Although it goes without saying that the hypothesis is 'mechanistic' in its inspiration, it is a slightly remarkable fact that a system

consisting of a body and a 'soul,' if such things there be, can perfectly well be a Probabilistic Automaton" (1967, p. 228). Putnam's Turing machine conception of mind will not satisfy the identity theorist's desire for a materialistic characterization of mind, not because minds may be souls, but because, when equated with machine tables, there is nothing more to a mind than a complex functional relation. This functional relation may be realized in this or that substance, but mental states, according to Putnam, are not identical with that which realizes a given relation. Rather, they consist simply in the relation itself.

Comparison with David Lewis's approach to functionalism makes Putnam's account clearer. For Lewis (1969, 1978), mental states are the *occupants* of functional roles. To illustrate his position, Lewis notes that there is a particular cat, Bruce, who occupied the role (for Lewis) *my cat* (1978, p. 218). It is a contingent matter that it is Bruce who occupied this role. Had Lewis owned some different cat, it would not be Bruce who occupied the role (for Lewis) *my cat*, but this other cat. Moreover, the role *my cat*, for me, was once occupied by Roxanne. And, again, this is a contingent matter. It is quite possible that my cat, *for me*, might have been occupied by some other feline. What is not contingent is that Bruce is identical with Lewis's cat and that Roxanne is identical with my cat: Bruce is identical with the cat who occupied the role *my cat* for Lewis, and so it is necessary that Bruce is that cat, and similarly, mutatis mutandis, for Roxanne. This example demonstrates the sense in which, on the one hand, what occupies a given role is contingent, but, on the other hand, why it also makes sense to *identify*, in this case, Bruce (the occupant of *my cat*) with Lewis's cat. Bruce *is* Lewis's cat. In a similar vein, Lewis notes that it is perfectly coherent to recognize the contingency in the claim "the winning lottery number is 17" (it might have been something else) while allowing that the winning number is identical with 17 (1969, p. 233).

The lesson Lewis takes from these cases applies in the following way to mental states. Consider Othello's jealousy. Jealousy is a mental state that defines a certain role, just as *my cat* and *the winning lottery number* define certain roles. In Othello, suppose that the role jealousy describes is occupied by a cluster of neurons J_O. J_O, the occupant of the role that jealousy defines, is identical with Othello's jealousy, just as Bruce is iden-

tical with Lewis's cat and 17 is identical with the winning lottery number. It is consistent with Lewis's view that whereas J_O is the kind of brain state that is jealousy in Othello, there might be other kinds of brain states (J_P, J_Q, J_R) that are identical with jealousy in other individuals. "No mystery," Lewis remarks, "that is just like saying that the winning number is 17 in the case of this week's lottery, 137 in the case of last week's" (1969, p. 233). Thus, Lewis concludes, it is possible to be a functionalist and yet not deny that mental states are identical with physical states.

Lewis intends his characterization of functionalism as a defense of some sort of mind–brain identity theory, but it seems clear that Putnam need not endorse Lewis's view. Lewis has not provided a reason to renounce Putnam's style of functionalism; he has merely found a way to provide functionalism with room to accommodate claims of identity between mental states and brain states. Still, Putnam is within his rights to insist that it is the particular role that ought to be identified with a given kind of mental state and not the occupant of the role. After all, one can imagine Putnam arguing, what is it that Kirk and Spock share when both are jealous? They do not, by hypothesis, share a kind of brain state. And, whereas Putnam could acknowledge that whenever Kirk is jealous his brain is in state J_K, and whenever Spock is jealous his brain is in state J_S, Putnam need not agree that jealousy *is* J_K in Kirk and *is* J_S in Spock. That is, Putnam can maintain that what it is to be jealous is to be in a state that describes particular kinds of relations between inputs, other states, and outputs. Putnam can accept that the role jealousy describes is occupied by physical states in both Kirk and Spock, but he need not take the step Lewis does in identifying the jealousy of each with the occupant of the role it describes. In short, Putnam and Lewis differ over whether the relations that mental state terms refer to are, by themselves, mental states. For Putnam they are, but for Lewis they are only a means for picking out some occupant that, Lewis thinks, is the real referent of mental state terms.

Many critics of Putnam's functionalism (including Putnam in later years) have focused on the analogy Putnam draws between human minds and Turing machines (Block and Fodor 1972; Putnam 1988). However, there is a more basic difficulty with the theory. Putnam is quite clear that

in suggesting that the mind is a Turing machine table he takes himself to be "advancing an empirical hypothesis" (1967, p. 226). He claims that his "strategy will be to argue that pain is not a brain state, not on *a priori* grounds, but on the grounds that another hypothesis is more plausible" (ibid.). But this raises several questions. First, in what sense is it an empirical hypothesis that the mind is a collection of relations described by some Turing machine table? This claim seems more like a stipulation on Putnam's part than an empirical hypothesis. There's no doubt that one can describe mental states in terms of their relations to sensory stimuli, other states, and behavior. Indeed, Ramsey (1931) showed that it is possible to define the theoretical terms of any theory in terms of relations between observables. In any event, because Putnam claims to be offering an empirical hypothesis, he should say something about the kind of evidence that might support the hypothesis. Whereas it is possible to imagine evidence that might bear on the question of which relations between stimuli, mental states, and behavior constitute a particular mental state, it seems completely bizarre to claim that evidence can bear on the claim that a particular mental state *is* nothing more than a relation. What experiment, for instance, might settle the dispute between Putnam and Lewis over whether jealousy refers just to a relation or to the physical state that fills the relation? This dispute seems much better suited to philosophical than empirical adjudication.

Perhaps machine functionalism is an empirical hypothesis in the following sense. It predicts that if we build a machine that realizes the mental machine table that describes our mind then this machine would have a mind just like our own. However, if this is the sense in which Putnam's hypothesis is empirical, then it seems to suffer from two shortcomings. First, it will in all likelihood not be testable within our lifetimes and may perhaps never be testable. But second and more significantly, our experience with systems simpler than the mind suggests that the hypothesis is false. Consider the relation that a mind would bear to a device, say, a computer, that has been engineered to respect the same machine table that constitutes our mind. The two systems—mind and computer—would be *functionally isomorphic*. Putnam defines this relation as follows:

> Two systems are functionally isomorphic if *there is a correspondence between the states of one and the states of the other that preserves functional relations*. . . . More generally, if T is a correct theory of the functioning of system 1, at the functional or psychological level, then an isomorphism between system 1 and system 2 must map each property and relation defined in system 2 in such a way that T comes out true when all references to system 1 are replaced by references to system 2, and all property and relation symbols in T are reinterpreted according to the mapping. (1975b, pp. 291–292)

With the concept of a functional isomorphism in hand, Putnam's empirical hypothesis is, on the current suggestion, that all systems functionally isomorphic to a mind are, in turn, minds.

However, we are now in a position to see why existing evidence makes this hypothesis unlikely. Computers are devices that are remarkably successful at simulating the behavior of complex systems. Moreover, a computer simulates a complex system by virtue of establishing a functional isomorphism between its own states and the states of the system it simulates. So, for instance, it is in virtue of maintaining an isomorphism between its own states and the states of a hurricane that a meteorologist can use a computer to predict the hurricane's path and force. Similarly, engineers rely on an isomorphism between a computer's states and the states of an airplane to predict how the airplane will perform in turbulent winds. One can simulate any behavior at all on a computer given a machine table description of the system to be simulated. Accordingly, the question one must now ask Putnam is this: why suppose that if we build a system that is functionally isomorphic to a mind then it would *be* a mind, given that we have built systems that are functionally isomorphic to hurricanes and airplanes but that turn out to be neither hurricanes nor airplanes? What is it about a mind that makes it the kind of thing that can be duplicated by mere functional isomorphism when functional isomorphism typically provides us with nothing more than a simulation? In short, from what evidence we have, it seems that something that satisfies a machine-functional description of a mind is no more likely to be a mind than is something likely to be a hurricane just because it satisfies a machine-functional description of a hurricane. There is simply no more reason to believe that a system functionally isomorphic to a mind can think than there is to be believe that a system functionally

isomorphic to a hurricane can bend palm trees and destroy trailer parks (see also Block 1978; Searle 1980; Sober 1992).

1.3.2 Functional Analysis Functionalism

As we have just seen, "functionalism" in philosophy of mind sometimes refers to a theory of mind that identifies mental states with Turing machine functional states. Certainly Putnam was a functionalist in this sense. However, there are other ways to conceive of functional states and accordingly other kinds of functionalist theories of mind. Fodor (1968) too speaks of mental states as functional states, but for him *function* has a much richer teleological sense than its anemic mathematical cousin. For Fodor, functions are contributions toward a goal, and they become apparent in the course of a functional analysis of a system (see also Cummins 1975; Lycan 1981). The point of a functional analysis of a system is to understand how a system achieves some capacity by way of the activities of its parts. Fodor explains:

> In typical cases of functional analysis . . . one asks about a part of a mechanism *what role it plays* in the activities that are a characteristic of the mechanism as a whole: "What does the camshaft do?" "It opens the valves, permitting the entry into the cylinder of fuel, which will then be detonated to drive the piston." . . . Successful functional analysis . . . requires an appreciation of the sorts of activity that are characteristic of a mechanism and of the contribution made by the functioning of each part of the mechanism to the economy of the whole. (1968, p. 113)

Functional analysis proceeds by assigning functions to the parts of a system, where these functions are better associated with purposes or goals than with mathematical operators that carry arguments from a domain onto values in a range.

If mental states are functional in this sense, the idea that minds are multiply realizable seems quite plausible. Because it is possible that various physical kinds can all exhibit the same characteristic activity, it is possible that a mind, analyzed as if it were a goal-directed system, is multiply realizable. Valve lifters, for instance, might be camshafts (understood as a kind of physical structure), but they might be other kinds of physical structures as well. What matters to whether something is a valve lifter is not that it has a particular kind of physical structure, but that it plays the role of permitting fuel to enter a cylinder in an automobile

engine. As Fodor notes, "If I speak of a device as a 'camshaft,' I am implicitly identifying it by reference to its physical structure, and so I am committed to the view that it exhibits a characteristic and specifiable decomposition into physical parts. But if I speak of the device as a 'valve lifter,' I am identifying it by reference to its function and I therefore undertake no such commitment" (1968, p. 113). Physical structures, in addition to falling under a physical description, can take on a functional description in the course of a functional analysis. But, whereas physical descriptions apply to structures in virtue of their physical type, functional descriptions classify by virtue of functional contributions to a system. Multiple realizability follows as a consequence of the fact that structures that differ in physical description may play the same contributing role in the systems of which they are parts.

Yet, even if we are to accept Fodor's suggestion that the mind is functional in the sense of being defined by a goal or purpose, there remains a gap between this assumption and the conclusion that minds are multiply realizable. Fodor may be right that minds or mental states ought to be identified by their function, but this does not entail that minds or mental states are multiply realizable in physical kinds. The problem is this. Whereas there are many cases in which objects of distinct physical description may bear the same functional description, there are also cases in which a functional description may apply only to a single kind of physical object. For instance, suppose one wishes to build a machine that drills for oil through surfaces composed of extremely hard minerals. *Drill bit* is presumably a functional kind. Like *valve lifter*, it appears that one can speak of drill bits without taking on a commitment to any particular physical description of the kinds of things that can do the job of a drill bit. However, if diamonds are the only substance that in fact are hard enough to drill through very hard surfaces, then *drill bit* picks out a physical kind no less than it refers to a functional kind.

As an alternative example, suppose one wishes to build a solar cell like the kind that powers a hand calculator. To build a solar cell, one needs a substance that turns light into electricity and in which it is possible to control the flow of electrons. Existing solar cells consist of two types of silicon—n-type silicon ("n" for negatively charged) in which some of the silicon atoms have an extra electron, and p-type silicon ("p" for

positively charged) in which some of the silicon atoms lack an electron. By layering n-type silicon on top of p-type silicon and exposing the resulting lattice to light, it is possible to create an electrical current. The light frees electrons, which move from the n-type silicon to the calculator, back into the p-type silicon and then up into the n-type silicon, and so on.[4] The kind *solar cell* appears to be a functional kind—it is defined as that which turns light into a controlled electrical current. But suppose that n- and p-type silicon are the only substances that exhibit the properties necessary for the construction of a solar cell. If such were the case, then "solar cell," which appears to be a functional kind, also picks out a particular type of physical structure, namely, a lattice of n- and p-type silicon atoms.

The point of these examples is not to argue that the mind, construed in the functional sense that Fodor develops, is not multiply realizable, but rather to show that adoption of a functional perspective toward the mind does not *entail* that the mind is multiply realizable. Perhaps it is; perhaps it is not. Whether it is requires empirical investigation of a sort that puts philosophers on the sidelines. Interestingly, Aristotle was sensitive to this point about limits on the multiple realizability of functional kinds, arguing (incorrectly, as it turns out) that, "if one defines the operation of sawing as being a certain kind of dividing, then this cannot come about unless the saw has teeth of a certain kind; and these cannot be unless it is of iron" (McKeon, tr. 1941).

Of course, a functionalist might respond to the above observation with the claim that functional kinds, while perhaps not always nomologically multiply realizable, are always logically multiply realizable. In other words, whereas there may be only one physical kind that fits a functional description, there are certainly many logically possible kinds that fit any given functional description. So what if diamonds are the only physically possible substance that can function in a drill that bores through hard surfaces, or if n- and p-types of silicon are physically necessary for the construction of a solar cell? We can surely *imagine* other substances from which drill bits and solar cells can be built that, although not present in the actual world, are present in logically possible worlds. This shows that minds, regardless of whether they are multiply realizable in the actual world, are multiply realizable in the logical sense.

This response seems correct. I think it would be futile to deny the logical possibility of multiple realizability. However, as I mentioned earlier, there remains the interesting question of whether minds are in *fact* multiply realizable. How could we test whether Fodor is right that minds, like valve lifters, are multiply realizable? As I argued above, the claim that minds are functional kinds does not imply that they are multiply realizable. Some functional kinds are multiply realizable, but some are not. There is nothing in Fodor's conception of functionalism that tells one way or another on the question of mind's multiple realizability.

On reflection, I think it is completely unsurprising that none of the conceptual arguments I have examined provides support for MRT. As advocates of MRT stress time and again, MRT is intended as an empirical thesis. But, this means that if one wants to confirm MRT, one should be looking not at what our concepts mean but at the world. If MRT is an empirical thesis, one should be trying to draw from it predictions that can be measured against observations. A critical discussion of the empirical support that philosophers have offered for MRT will be the business of the remainder of this chapter.

1.4 Empirical Arguments for the Multiple Realizability Thesis

If the conceptual arguments that Putnam and Fodor advance do not force us to accept MRT, perhaps their empirical arguments will do a better job. In this section I examine first a likelihood argument for MRT that Putnam (1967) makes. I then consider three lines of empirical evidence for MRT that Block and Fodor (1972) advance. The issues involved in these arguments become quite complex and untangling them all will require not only a close examination of the arguments, but also a stance on how to interpret claims of multiple realization. Because these topics can be treated separately, I will reserve discussion of the latter until the following chapter.

1.4.1 Putnam's Likelihood Argument for the Multiple Realizability Thesis

In addition to thinking that the Turing machine is the right way to conceive minds, Putnam thinks that there are independent empirical reasons

for doubting the identity theory of mind. A sensible way to understand Putnam's argument is as a likelihood argument (Sober 1993). A likelihood argument is an argument that seeks to adjudicate between two competing hypotheses. It does so by considering a single body of evidence and asking of each of the hypotheses under consideration which makes the evidence more probable. The likelier hypothesis is the one that makes the evidence more probable. So, for instance, in the movie *Rounders* the girlfriend of a poker player alleges that success in poker is merely a matter of luck. The poker player, wishing to rebut this charge, points out to his girlfriend that the same players are present year after year in the World Series of Poker tournament. This is a compelling defense for reasons that a likelihood reconstruction makes clear. In essence, the poker player is asking his girlfriend to consider two hypotheses:

L: Success in poker is merely a matter of luck.

S: Success in poker is due to skill.

To decide between L and S it is necessary to consider the evidence, that is, the fact that the same people are present in the World Series of Poker tournament every year. Suppose L were true. If L were true then we would expect to see different players at the tournament every year because, by definition, luck is a matter of chance. We should no more expect to see the same faces in the tournament every year than we should expect that a single individual would pull an ace from a fair deck a hundred times in a row.

On the other hand, if S were true then the presence of the same people at the tournament every year would not be at all surprising. These players make it to the tournament every year because they are very good poker players. Happy Jack's presence in the poker tournament every year would be no more surprising than Martina Navratilova's appearance at Wimbledon every year.

Notice that likelihood and probability are two different notions. H1 is likelier than H2 when it makes an observation O more probable. This is expressed in probability theory as: $Pr(O|H1) > Pr(O|H2)$. This claim about likelihood should not be confused with a claim about the probabilities of the hypotheses given an observation—that $Pr(H1|O) >$

Pr(H2|O). To see why likelihood and probability differ, consider a third hypothesis:

T: Success in poker is due to telepathic abilities that only a very few people possess.

Hypotheses T and S appear equally likely given the fact that the same people show up in the tournament every year. That is, Pr(O|S) = Pr(O|T). However, there should be no doubt that Pr(S|O) > Pr(T|O). Hypothesis T is improbable to begin with, and O hardly makes it any more probable. The advantage to considering likelihood over probability is that the former focuses purely on the connection between the hypothesis and the evidence at hand. Even with no idea about the initial plausibilities of the competing hypotheses, it is still possible to assess their likelihoods.

For present purposes, we should construe Putnam (1967) as considering the following two hypotheses:

TI: The mind (construed as a collection of mental states) is type-identical with the brain.

MR: The mind (construed as a collection of mental states) is multiply realizable.

Putnam then asks us to evaluate the likelihood of these two hypotheses given observation P:

P: Pain is a mental state present in mammalian brains, reptilian brains, mollusc brains and, conceivably, extraterrestrial brains (if ETs exist).

Putnam's claim then is that Pr(P|MR) > Pr(P|TI). He believes this because if mental states like pain are multiply realizable, it would not be at all surprising to find pain-feeling organisms with different brain structures. On the other hand, TI makes P improbable because, Putnam assumes, the brains of mammals, reptiles, molluscs, and ETs (should they exist) are probably very different in structure. P is probable given TI only if mammals, reptiles, molluscs, and ETs evolved similar brains independently of each other. Putnam concludes, "Thus it is at least possible that parallel evolution, all over the universe, might *always* lead to *one and the same* physical "correlate" of pain. But this is certainly an ambitious hypothesis" (1967, p. 228).

It is worth making two points about Putnam's likelihood argument. First, Putnam's line of reasoning appears to be a good argument against the identity theory, but it is not a strong argument for his Turing machine functionalism. This is because P seems to be equally probable given various kinds of functionalist theories of mind. To see this, consider two further hypotheses:

TM: The mind is a Turing machine table.

FA: The mind is functional in the sense assumed by functional analyses.

Given P, TM and FA seem equally likely because each predicts that brains of different physical organization might realize similar minds. Because P is no more or less surprising given TM or FA, it is of little help to Putnam in arguing for his version of functionalism over other versions.

Second and more important, Putnam's likelihood argument, although compelling, is not as strong as one might first suspect. Consider what makes the poker player's likelihood argument convincing. We know something about the odds of being dealt a winning poker hand. If poker were just a matter of luck, the odds that the same people would often have winning hands would be extraordinarily low. We are more likely to find winning cards in the hands of players who know what to do with the cards they've been dealt than in the hands of players who know nothing about poker. But, we know very little about the properties of the brain that make it a realization of a mind. Perhaps minds are multiply realizable, as Putnam believes. But, as I argued earlier, they may not be. Suppose one were to argue in the following way. We have the following observation:

O: Solar cells are present in all sorts of devices (calculators, satellites, watches, toys, etc.).

Now consider the following two hypotheses:

MR_S: Solar cells are multiply realizable.

TI_S: Solar cells can be built only from silicon.

Can we assert that $Pr(O|MR_S) > Pr(O|TI_S)$? I do not think so. Prior to empirical investigation of semiconductors, we simply do not know whether the myriad devices that contain solar cells offer support for MR_S

or for TI_S. If our investigation shows that there are in fact many ways to build solar cells, then we should believe MR_S. On the other hand, if scientists are unable to design a solar cell without utilizing n- and p-types of silicon, then TI_S gains support. The point is that an observation like O tells us nothing about the relative likelihoods of MR_S and TI_S. O is valuable only when combined with the information that the various devices in which solar cells appear have nothing relevantly similar in common (where "nothing relevantly similar" means something like: have no parts in common that turn light into electricity). But, of course, if one had this information then one would have already confirmed MR_S.

Similarly, why should P offer support for MR over TI unless we already know that the brains of mammals, reptiles, molluscs, and ETs are relevantly different? But if we know this then we know all we need to know to judge the relative merits of MR and TI. The conclusion to draw from this discussion is that whether minds are multiply realizable is as much a matter for empirical investigation as whether solar cells are. We cannot conclude from the observation of nonhuman organisms with mental properties that minds are multiply realizable until we take a peek inside their skulls. But, even then, we have to know what to look for.

1.4.2 Three Lines of Evidence for the Multiple Realizability Thesis

It is in this context that Block and Fodor's (1972) discussion of multiple realizability becomes relevant. Block and Fodor, rather than constructing a likelihood argument for MRT, believe that there currently exists direct evidence of multiple realizability. In particular, Block and Fodor invite us to examine "three kinds of empirical considerations" (1972, p. 238) that, they think, add up to a strong case for the multiple realizability of minds. The first line of evidence comes from studies that reveal the brain to have a very plastic structure. As Block and Fodor say, "it does seem clear that the central nervous system is highly labile and that a given type of psychological process is in fact often associated with a variety of distinct neurological structures" (ibid.). If it is true that the same psychological processes can be realized in different neurological structures, this is living proof that the mind, or at least some mental capacities, are multiply realizable. And, indeed, literature in the neurosciences seems replete with examples of such neural plasticity: people

whose speech centers have developed in the right hemisphere as a result of some sort of insult to the left; the use of auditory cortex to process visual information (von Melchner, Pallas, and Sur 2000), and so on. This evidence is perhaps the best reason to think that MRT is true, but, as we'll see in the next chapter, it is not conclusive.

Second, Block and Fodor point out that convergent evolution—the phenomenon in which species evolve similar traits independently of each other—is just as likely to occur in cases of psychological traits as it is in cases of morphological and behavioral traits (1972, p. 238). They note that "if there are organisms whose psychology is homologous[5] to our own but whose physiology is quite different, such organisms may provide counterexamples to the psychophysical correlations physicalism requires" (ibid.). It is worth spending some time on this point because the issue of convergence will become significant in subsequent chapters as I develop a case for MRT's competitor—the mental constraint thesis (MCT).

Convergence involves two ideas. The first concerns the similarity of a trait (more technically, of a character state) in members of distinct species. The second idea involves an explanation of this similarity. The claim that a similar trait in two distinct species is the product of convergence means that the trait has evolved twice—once in each lineage—as a result of (presumably) similar selection pressures. The alternative hypothesis is that the trait is homologous, that is, the similarity of the trait is the result of inheritance of the trait from a common ancestor. The wings of birds and bats, for instance, are an example of convergence. They are similar traits that have evolved independently in the two lineages. In contrast to such independently derived traits—"homoplasious" traits—the wings of robins and sparrows are homologous. The similarity of these wings is a consequence of the fact that robins and sparrows are descended from an ancestor that had wings and that the lineages from which robins and sparrows descended retained this ancestral state.

Block and Fodor claim that when independent evolution results in analogous traits, it may well turn out that the traits are distinct kinds of realizations. Of course, this by itself is no *argument* that the independent evolution of analogous traits will always or even usually lead to multiple realization. It is an interesting question whether the independent

evolution of analogous traits should lead us to expect that these traits will be realized in different ways. Suppose that two organisms from different species have a similar psychological trait. Further suppose that the similarity is a result of independent evolution. Does this raise the probability that the trait will be realized differently in each organism?

I think the answer to this question must be "no." There is simply no reason to suppose that homoplasious traits are probably different kinds of realizations. This depends on whether the kind of trait is multiply realiz*able* in the first place. The fact that a number of factories might produce a particular kind of drill bit is not evidence that the drill bit is realized in a number of different kinds of ways. It might turn out that diamonds are the only suitable realizer for the kind of bit in demand. If so, the number of factories that produce the bit makes no difference to the probability that the bit is multiply realized.

To illustrate the point further, suppose the psychological trait in question is vision. Block and Fodor's appeal to convergence suggests the following kind of argument. If various species have evolved vision independently then it is more probable that vision is multiply realized than if vision is a result of common ancestry. That is, one might hope convergence justifies the following:

$$\Pr(\text{VMR}|\text{independent evolution}) > \Pr(\text{VMR}|\text{common ancestry}),$$

where "VMR" is the claim that vision is multiply realized, "independent evolution" stands for the hypothesis that the most recent common ancestor of the two lineages lacked the observed trait, and "common ancestry" is the hypothesis that the trait shared in the two lineages is one that is present because it has been inherited from the most recent common ancestor. My claim is that evolutionary theory provides no reason to suppose that this inequality holds. It is true that vision has evolved independently in a number of lineages, and it is true that vision is realized differently in various lineages. But the latter fact is not entailed by the former. Indeed, it is also true that vision has been realized in the *same* way in a number of distinct lineages. Thus, for instance, whereas vision in some lineages is realized in a compound eye and in other lineages it is realized in a camera eye, compound eyes and camera eyes have themselves evolved independently in a number of cases. Whether the

above inequality holds depends on how stingy nature is with the solutions it allows. But the stinginess of nature is an assumption that must itself be put to the test prior to making a judgment about the above inequality. In short, the two hypotheses—common ancestry and independent evolution—make no predictions about the multiple realizability of a psychological capacity without further assumptions about the extent to which nature might allow such multiple realizability.

On the other hand, the following inequality is plausible:

Pr (~VMR|common ancestry) > Pr (VMR|common ancestry).

Given that members of two distinct species have a similar psychological capacity in virtue of having inherited it from their most recent common ancestor, it is reasonable to suppose that the similarity in their psychological capacities is a result of a similarity in the realizers of the capacities. The argument for this claim depends on a parsimony consideration. If two species share the same visual capacities as a result of common ancestry, then to suppose that their visual capacities are differently realized is to suppose that at least one of these species evolved a new way to realize its visual capacities while losing the ancestral manner of realization. Thus, more evolution would be required to have the situation described by the right side of the inequality than that described on the left (see figure 1.2).[6]

So far I have been interpreting Block and Fodor as offering convergence as a hypothesis that, in their view, raises the probability of multiple realization. I think this is the correct way to understand them. However, there are other ways to understand the relation between convergence and multiple realization. In particular, and more important for my purposes, it is possible to think of convergence in a way that allows one to test MRT against its competitor MCT. Putnam must have had a similar idea, when, as we saw, he mentions parallel evolution (which is often equated with convergent evolution) as a reason that might bring one to *doubt* MRT (1967, p. 228). Putnam mentions the similarity between the octopus's eye and the mammalian eye and takes this to mark a challenge, though an extremely improbable one, to the possibility that minds are multiply realizable. I believe a likelihood argument is again the best way to understand the nature of this challenge.

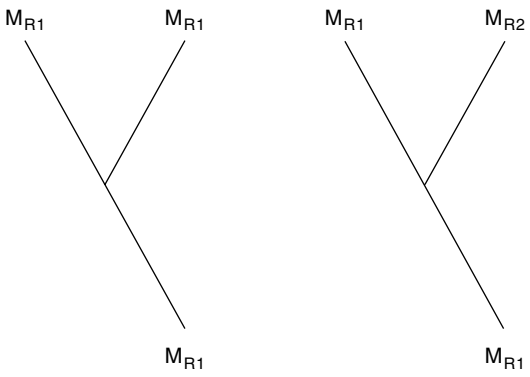

Figure 1.2
In the tree on the left, the ancestral condition consists of a mental capacity *M* realized by *R1*. The descendants also have *M* realized by *R1*. On the right, the ancestral condition is *M* realized by *R1*, but one of the descendants now has *M* realized differently, by *R2*. Thus, the tree on the right involves more evolution—the descendant on the right branch lost *R1* but gained *R2*.

As I mentioned, convergence involves two ideas: a similarity between traits and an explanation of this similarity in terms of independent evolution. If we take independent evolution to be a given, it becomes possible to make the observation of similarity in traits across species work as evidence that can test MRT against MCT. So, suppose that we are considering a trait that is similar in two species, A and B, and we know that the similarity is a result of independent evolution. The probabilities to consider are these:

(1) Pr_{IE} (A and B have similarly realized trait T | MRT); and

(2) Pr_{IE} (A and B have similarly realized trait T | MCT).

Here the "IE" subscript indicates that the probabilities are to be evaluated on the assumption that trait T has evolved independently in A and B. I claim that the first probability is less than the second. The argument is this. Suppose MRT is true of trait T. This means that there are many ways to realize T. But then it should come as a surprise that A and B, which evolved trait T independently, should realize T in the same way. On the other hand, suppose MCT is true. This means that there are few ways to realize T. If trait T has arisen independently in species A and B,

and if MCT is true, there is a higher probability that T will be realized the same way in the two species.

If we think of MRT and MCT as staking out positions at opposite ends of a spectrum, a view I tend to favor, then the use of similarity observations to test MRT against MCT suggests the following. Suppose that MRT makes the claim that, for a given trait T, there are hundreds or thousands of different possible realizers. Assume further that each realization of T is as probable as another. In contrast, suppose MCT makes the claim that there are but a handful or a couple of possible realizers of T. Letting *n* be the number of realizers of T, MRT says that *n* is very large; MCT says that it is quite small. But notice that as *n* increases, the probability that two species will independently evolve the same trait decreases. Similarly, as *n* decreases, the probability that two species will independently evolve the same trait increases. Thus, the inequality between (1) and (2) will vary in size according to the number of possible realizers of a given trait, but in any event (1) will be less than (2).

If the argument above is on the right track, it is possible to understand why Putnam is impressed with the similarity of the octopus and mammalian eye. The similarity is surprising if there are many ways to build an eye, but it is just what one should expect if there are few ways to build an eye. In the same way, if minds are multiply realizable, then a similarity in the ways minds are realized (given that they have evolved independently) should cast doubt on MRT but should support MCT.

Of course, there remain many questions to answer before one can hope to apply the reasoning above in a way that will provide a meaningful evaluation of MRT. For one, more needs to be said about what, exactly, it means for something to be realized and, moreover, what it means to talk about sameness and difference in realization. Additionally, one might wonder how, given that a humanlike mind has evolved only once, it can be possible to gather the observations necessary to compare MRT to MCT. These are important issues, and ones to which I shall soon attend in the next chapter. However, before doing so, it is worth mentioning briefly the third bit of empirical evidence that Block and Fodor take to support MRT. Block and Fodor claim that "it seems very likely that given any psychophysical correlation which holds for an organism, it is possible to build a machine which is similar to the organism psy-

chologically, but physiologically sufficiently different from the organism that the psychophysical correlation does not hold for the machine" (1972, p. 238).

It is hard to know on what Block and Fodor's faith rests regarding this point. Surely they cannot believe that AI has, even thirty years after their initial claim, come anywhere near to creating a mind that is similar in its capacities to a human mind. More bewildering still is their introduction of this claim as a "conceptual possibility" when they are presumably interested in defending the nomological possibility of MRT. In any event, whether AI offers support to MRT depends again on an assumption about how to judge sameness and difference in realizations, for smart computers confirm MRT only if their smarts are realized differently from our own. Accordingly, it is now time to consider more generally how to understand the notion of realization, as well as what makes two realizations of a type *different* types of realization.