Others writing on the topic of "consciousness" have been primarily concerned with self-consciousness or introspective consciousness. Van Gulick (1988), in suggesting that consciousness should be analyzed as the possession of "reflexive metapsychological information," is at best providing an analysis of these psychological notions, and indeed concedes that the phenomenal aspects may be left out by such an analysis. Similarly, Jaynes's (1976) elaborate theory of consciousness is concerned only with our awareness of our own thoughts. It says nothing about phenomena associated with perception and therefore could not hope to be a theory of awareness in general, let alone a theory of phenomenal consciousness. Hofstadter (1979) has some

interesting things to say about consciousness, but he is more concerned with introspection, free will, and the sense of self than with experience *per se.*

Insofar as consciousness has been a topic for discussion among psychologists, the phenomenal and psychological notions have not often been carefully distinguished. Usually it is some aspect of awareness, such as introspection, attention, or self-consciousness, that psychological studies address. Even the psychological aspects of consciousness have had something of a bad name in psychology, at least until recently. Perhaps this is because of some unclearness in those notions, and the difficulties associated with high-level phenomena such as introspection. One might speculate that to a larger extent this bad reputation is due to their sharing a name with phenomenal consciousness, giving the appearance of partnership in crime.

One sometimes hears that psychological research has been "returning to consciousness" in recent years. The reality seems to be that the psychological aspects of consciousness have been an active subject of research, and that researchers have not been afraid to use the term "consciousness" for the phenomena. For the most part, however, phenomenal consciousness remains off to the side. Perhaps this is understandable. While one can see how the methods of experimental psychology might lead to an understanding of the various kinds of awareness, it is not easy to see how they could explain phenomenal experience.[17]

Cognitive models are well suited to explaining psychological aspects of consciousness. There is no vast metaphysical problem in the idea that a physical system should be able to introspect its internal states, or that it should be able to deal rationally with information from its environment, or that it should be able to focus its attention first in one place and then in the next. It is clear enough that an appropriate functional account should be able to explain these abilities, even if discovering the correct account takes decades or centuries. But the really difficult problem is that of phenomenal consciousness, and this is left untouched by the explanations of psychological consciousness that have been put forward so far.

In what follows, I revert to using "consciousness" to refer to phenomenal consciousness alone. When I wish to use the psychological notions, I will speak of "psychological consciousness" or "awareness." It is phenomenal consciousness with which I will mostly be concerned.

2—
Supervenience and Explanation

What is the place of consciousness in the natural order? Is consciousness physical? Can consciousness be explained in physical terms? To come to grips with these issues, we need to build a framework; in this chapter, I build one. The centerpiece of this framework is the concept of *supervenience: I* give an account of this concept and apply it to clarify the idea of reductive explanation. Using this account, I sketch a picture of the relationship between most high-level phenomena and physical facts, one that seems to cover everything except, perhaps, for conscious experience.

1—
Supervenience

It is widely believed that the most fundamental facts about our universe are physical facts, and that all other facts are dependent on these. In a weak enough sense of "dependent," this may be almost trivially true; in a strong sense, it is controversial. There is a complex variety of dependence relations between high-level facts and low-level facts in general, and the kind of dependence relation that holds in one domain, such as biology, may not hold in another, such as that of conscious experience. The philosophical notion of supervenience provides a unifying framework within which these dependence relations can be discussed.

The notion of supervenience formalizes the intuitive idea that one set of facts can fully determine another set of facts. [1] The physical facts about the world seem to determine the biological facts, for instance, in that once all the physical facts about the world are fixed, there is no room for the biological facts to vary. (Fixing all the physical facts will simultaneously fix which objects are alive.) This provides a rough characterization of the sense in

which biological properties supervene on physical properties. In general, supervenience is a relation between two sets of properties: *B*-properties—intuitively, the *high-level* properties—and *A*-properties, which are the more basic *low-level* properties.

For our purposes, the relevant A-properties are usually the physical properties: more precisely, the fundamental properties that are invoked by a completed theory of physics. Perhaps these will include mass, charge, spatiotemporal position; properties characterizing the distribution of various spatiotemporal fields, the exertion of various forces, and the form of various waves; and so on. The precise nature of these properties is not important. If physics changes radically, the relevant class of properties may be quite different from those I mention, but the arguments will go through all the same. Such high-level properties as juiciness, lumpiness, giraffehood, and the like are excluded, even though there is a sense in which these properties are physical. In what follows, talk of physical properties is implicitly restricted to the class of fundamental properties unless otherwise indicated. I will sometimes speak of "microphysical" or "low-level physical" properties to be explicit.

The *A-facts* and *B-facts* about the world are the facts concerning the instantiation and distribution of A-properties and B-properties. [2] So the physical facts about the world encompass all facts about the instantiation of physical properties within the spatiotemporal manifold. It is also useful to stipulate that the world's physical facts include its basic physical laws. On some accounts, these laws are already determined by the totality of particular physical facts, but we cannot take this for granted.

The template for the definition of supervenience is the following:

> B-properties *supervene* on A-properties if no two possible situations are identical with respect to their A-properties while differing in their B-properties.

For instance, biological properties supervene on physical properties insofar as any two possible situations that are physically identical are biologically identical. (I use "identical" in the sense of indiscernibility rather than numerical identity here. In this sense, two separate tables might be physically identical.) More precise notions of supervenience can be obtained by filling in this template. Depending on whether we take the "situations" in question to be individuals or entire worlds, we arrive at notions of *local* and *global* supervenience, respectively. And depending on how we construe the notion of possibility, we obtain notions of *logical* supervenience, *natural* supervenience, and perhaps others. I will flesh out these distinctions in what follows.

Local and Global Supervenience

B-properties supervene *locally* on A-properties if the A-properties of an *individual* determine the B-properties of that individual—if, that is, any two

possible individuals that instantiate the same A-properties instantiate the same B-properties. For example, shape supervenes locally on physical properties: any two objects with the same physical properties will necessarily have the same shape. Value does not supervene locally on physical properties, however: an exact physical replica of the Mona Lisa is not worth as much as the Mona Lisa. In general, local supervenience of a property on the physical fails if that property is somehow context-dependent—that is, if an object's possession of that property depends not only on the object's physical constitution but also on its environment and its history. The Mona Lisa is more valuable than its replica because of a difference in their historical context: the Mona Lisa was painted by Leonardo, whereas the replica was not. [3]

B-properties supervene *globally* on A-properties, by contrast, if the A-facts about the entire *world* determine the B-facts: that is, if there are no two possible worlds identical with respect to their A-properties, but differing with respect to their B-properties.[4] A world here is to be thought of as an entire universe; different possible worlds correspond to different ways a universe might be.

Local supervenience implies global supervenience, but not vice versa. For example, it is plausible that biological properties supervene globally on physical properties, in that any world physically identical to ours would also be biologically identical. (There is a small caveat here, which I discuss shortly.) But they probably do not supervene locally. Two physically identical organisms can arguably differ in certain biological characteristics. One might *befitter* than the other, for example, due to differences in their environmental contexts. It is even conceivable that physically identical organisms could be members of different species, if they had different evolutionary histories.

The distinction between global and local supervenience does not matter too much when it comes to conscious experience, because it is likely that insofar as consciousness supervenes on the physical at all, it supervenes locally. If two creatures are physically identical, then differences in environmental and historical contexts will not prevent them from having identical experiences. Of course, context can affect experience indirectly, but only by virtue of affecting internal structure, as in the case of perception. Phenomena such as hallucination and illusion illustrate the fact that it is internal structure rather than context that is *directly* responsible for experience.

Logical and Natural Supervenience

A more important distinction for our purposes is between *logical* (or conceptual) supervenience, and mere *natural* (or nomic, or empirical) supervenience.

B-properties supervene *logically* on A-properties if no two *logically possible* situations are identical with respect to their A-properties but distinct with respect to their B-properties. I will say more about logical possibility later in this chapter. For now, one can think of it loosely as possibility in the broadest sense, corresponding roughly to conceivability, quite unconstrained by the laws of our world. It is useful to think of a logically possible world as a world that it would have been in God's power (hypothetically!) to create, had he so chosen. [5] God could not have created a world with male vixens, but he could have created a world with flying telephones. In determining whether it is logically possible that some statement is true, the constraints are largely *conceptual.* The notion of a male vixen is contradictory, so a male vixen is logically impossible; the notion of a flying telephone is conceptually coherent, if a little out of the ordinary, so a flying telephone is logically possible.

It should be stressed that the logical supervenience is not defined in terms of deducibility in any system of formal logic. Rather, logical supervenience is defined in terms of logically possible *worlds* (and individuals), where the notion of a logically possible world is independent of these formal considerations. This sort of possibility is often called "broadly logical" possibility in the philosophical literature, as opposed to the "strictly logical" possibility that depends on formal systems.[6]

At the global level, biological properties supervene logically on physical properties. Even God could not have created a world that was physically identical to ours but biologically distinct. There is simply no logical space for the biological facts to independently vary. When we fix all the physical facts about the world—including the facts about the distribution of every last particle across space and time—we will in effect also fix the macroscopic shape of all the objects in the world, the way they move and function, the way they physically interact. If there is a living kangaroo in this world, then *any* world that is physically identical to this world will contain a physically identical kangaroo, and that kangaroo will automatically be alive.

We can imagine that a hypothetical superbeing—Laplace's demon, say, who knows the location of every particle in the universe—would be able to straightforwardly "read off" all the biological facts, once given all the microphysical facts. The microphysical facts are enough for such a being to construct a model of the microscopic structure and dynamics of the world throughout space and time, from which it can straightforwardly deduce the macroscopic structure and dynamics. Given all that information, it has all the information it needs to determine which systems are alive, which systems belong to the same species, and so on. As long as it possesses the biological concepts and has a full specification of the microphysical facts, no other information is relevant.

In general, when B-properties supervene logically on A-properties, we can say that the A-facts *entail* the B-facts, where one fact entails another if it is logically impossible for the first to hold without the second. In such cases, Laplace's demon could read off the B-facts from a specification of the A-facts, as long as it possesses the B-concepts in question. (I will say much more about the connections between these different ways of understanding logical supervenience later in the chapter; the present discussion is largely for illustration.) In a sense, when logical supervenience holds, *all there is* to the B-facts being as they are is that the A-facts are as they are.

There can be supervenience without logical supervenience, however. The weaker variety of supervenience arises when two sets of properties are systematically and perfectly *correlated* in the natural world. For example, the pressure exerted by one mole of a gas systematically depends on its temperature and volume according to the law $pV = KT$, where K is a constant (I pretend for the purposes of illustration that all gases are ideal gases). In the actual world, whenever there is a mole of gas at a given temperature and volume, its pressure will be determined: it is empirically impossible that two distinct moles of gas could have the same temperature and volume, but different pressure. It follows that the pressure of a mole of gas supervenes on its temperature and volume in a certain sense. (In this example, I am taking the class of A-properties to be much narrower than the class of physical properties, for reasons that will become clear.) But this supervenience is weaker than logical supervenience. It is *logically* possible that a mole of gas with a given temperature and volume might have a different pressure; imagine a world in which the gas constant K is larger or smaller, for example. Rather, it is just a fact about *nature* that there is this correlation.

This is an example of *natural* supervenience of one property on others: in this instance, pressure supervenes naturally on temperature, volume, and the property of being a mole of gas. In general, B-properties supervene naturally on A-properties if any two *naturally possible* situations with the same A-properties have the same B-properties.

A naturally possible situation is one that could actually occur in nature, without violating any natural laws. This is a much stronger constraint than mere logical possibility. The scenario with a different gas constant is logically possible, for example, but it could never occur in the real world, so it is not naturally possible. Among naturally possible situations, any two moles of gas with the same temperature and volume will have the same pressure.

Intuitively, natural possibility corresponds to what we think of as real *empirical* possibility—a naturally possible situation is one that could come up in the real world, if the conditions were right. These include not just actual situations but counterfactual situations that might have come up in the world's history, if boundary conditions had been different, or that might come up in the future, depending on how things go. A mile-high skyscraper

is almost certainly naturally possible, for example, even though none has actually been constructed. It is even naturally possible (although wildly improbable) that a monkey could type *Hamlet.* We can also think of a naturally possible situation as one that conforms to the laws of nature of our world. [7] For this reason, natural possibility is sometimes called *nomic* or *nomological* possibility,[8] from the Greek term *nomos* for "law."

There are a vast number of logically possible situations that are not naturally possible. Any situation that violates the laws of nature of our world falls into this class: a universe without gravity, for example, or with different values of fundamental constants. Science fiction provides many situations of this sort, such as antigravity devices and perpetual-motion machines. These are easy to imagine, but almost certainly could never come to exist in our world.

In the reverse direction, any situation that is naturally possible will be logically possible. The class of natural possibilities is therefore a subset of the class of logical possiblities. To illustrate this distinction: both a cubic mile of gold and a cubic mile of uranium-235 seem to be logically possible, but as far as we know, only the first is naturally possible—a (stable) cubic mile of uranium-235 could not exist in our world.

Natural supervenience holds when, among all naturally possible situations, those with the same distribution of A-properties have the same distributon of B-properties: that is, when the A-facts about a situation *naturally* *necessitate* the B-facts. This happens when the same clusters of A-properties in our world are always accompanied by the same B-properties, and when this correlation is not just coincidental but *lawful:* that is, when instantiating the A-properties will always bring about the B-properties, wherever and whenever this happens. (In philosophical terms, the dependence must support counterfactuals.) This co-occurrence need not hold in every logically possible situation, but it must hold in every naturally possible situation.

It is clear that logical supervenience implies natural supervenience. If any two logically possible situations with the same A-properties have the same B-properties, then any two naturally possible situations will also. The reverse does not hold, however, as the gas law illustrates. The temperature and volume of a mole of gas determine pressure across naturally but not logically possible situations, so pressure depends naturally but not logically on temperature and volume. Where we have natural supervenience without logical supervenience, I will say that we have *mere* natural supervenience.

For reasons that will become clear, it is hard to find cases of natural supervenience on the set of *physical* properties without logical supervenience, but consciousness itself can provide a useful illustration. It seems very likely that consciousness is naturally supervenient on physical properties, locally or globally, insofar as in the natural world, any two physically identical creatures will have qualitatively identical experiences. It is not at all clear that consciousness is logically supervenient on physical properties, however.

It seems *logically* possible, at least to many, that a creature physically identical to a conscious creature might have no conscious experiences at all, or that it might have conscious experiences of a different kind. (Some dispute this, but I use it for now only as an illustration.) If this is so, then conscious experience supervenes naturally but not logically on the physical. The necessary connection between physical structure and experience is ensured only by the laws of nature, and not by any logical or conceptual force.

The distinction between logical and natural supervenience is vital for our purposes. [9] We can intuitively understand the distinction as follows. If B-properties supervene logically on A-properties, then once God (hypothetically) creates a world with certain A-facts, the B-facts come along for free as an automatic consequence. If B-properties merely supervene naturally on A-properties, however, then after making sure of the A-facts, God has to do more work in order to make sure of the B-facts: he has to make sure there is a law relating the A-facts and the B-facts. (I borrow this image from Kripke 1972.) Once the law is in place, the relevant A-facts will automatically bring along the B-facts; but one could, in principle, have had a situation where they did not.

One also sometimes hears talk of *metaphysical* supervenience, which is based on neither logical nor natural necessity, but on "necessity *tout court*," or "metaphysical necessity" as it is sometimes known (drawing inspiration from Kripke's [1972] discussion of *a posteriori* necessity). I will argue later that the metaphysically possible worlds are just the logically possible worlds (and that metaphysical possibility of statements is logical possibility with an *a posteriori* semantic twist), but for now it is harmless to assume there is a notion of metaphysical supervenience, to be spelled out by analogy with the notions of logical and natural supervenience above. A notion of "weak" supervenience is also mentioned occasionally, but seems too weak to express an interesting dependence relation between properties.[10]

The logical-natural distinction and the global-local distinction cut across each other. It is reasonable to speak of both global logical supervenience and local logical supervenience, although I will more often be concerned with the former. When I speak of logical supervenience without a further modifier, global logical supervenience is intended. It is also coherent to speak of global and local natural supervenience, but the natural supervenience relations with which we are concerned are generally local or at least localizable, for the simple reason that evidence for a natural supervenience relation generally consists in local regularities between clusters of properties.[11]

*A Problem with Logical Supervenience**

A technical problem with the notion of logical supervenience needs to be dealt with. This problem arises from the logical possibility of a world physically identical to ours, but with additional nonphysical stuff that is not present in our own world: angels, ectoplasm, and ghosts, for example. There is a *conceivable* world just like ours except that it has some extra angels hovering in a non-physical realm, made of ectoplasm. These angels might have biological properties of their own, if they reproduced and evolved. Presumably the angels could have all sorts of beliefs, and their communities might have complex social structure.

The problem these examples pose is clear. The angel world is physically identical to ours, but it is biologically distinct. If the angel world is logically possible, then according to our definition biological properties are not supervenient on physical properties. But we certainly *want* to say that biological properties are supervenient on physical properties, at least in *this* world if not in the angel world (assuming there are no angels in the actual world!). Intuitively, it seems undesirable for the mere logical possibility of the angel world to stand in the way of the determination of biological properties by physical properties in our own world.

This sort of problem has caused some (e.g., Haugeland 1982; Petrie 1987) to suggest that logical possibility and necessity are too strong to serve as the relevant sort of possibility and necessity in supervenience relations, and that a weaker variety such as natural possibility and necessity should be used instead. But this would render useless the very useful distinction between logical and natural supervenience outlined above, and would also ignore the fact that there is a very real sense in which the biological facts about our world are logically determined by the physical facts. Others (e.g., Teller 1989) have bitten the bullet by stipulating that worlds with extra nonphysical stuff are not logically or metaphysically possible, despite appearances, but this makes logical and metaphysical possibility seem quite arbitrary. Fortunately, such moves are not required. It turns out that it is possible to retain a useful notion of logical supervenience compatible with the possibility of these worlds, as long as we fix the definition appropriately. [12]

The key to the solution is to turn supervenience into a thesis about *our* world (or more generally, about particular worlds). This accords with the intuition that biological facts are logically determined by the physical facts in our world, despite the existence of bizarre worlds where they are not so determined. According to a revised definition, B-properties are logically supervenient on A-properties if the B-properties in our world are logically determined by the A-properties in the following sense: in any possible world with the same A-facts, the same B-facts will hold.[13] The existence of possible worlds with *extra* B-facts will thus not count against logical supervenience in our world, as long as *at least* the B-facts true in our world are true in all physically identical worlds. And this they generally will be (with an exception discussed below). If there is a koala eating in a gum tree in this world, there will be an identical koala eating in a gum tree in any physically identical world, whether or not that world has any angels hanging around.

There is a minor complication. There is a certain sort of biological fact about our world that does not hold in the angel world: the fact that our world has no living ectoplasm, for example, and the fact that all living things are based on DNA. Perhaps the angel world might even be set up with ectoplasm causally dependent on physical processes, so that wombat copulation on the physical plane sometimes gives rise to baby ectoplasmic wombats on the nonphysical plane. If so, then there might be a wombat that is childless (in a certain sense) in our world, with a counterpart that is not childless in the physically identical angel world. It follows that the property of being childless does not supervene according to our definition, and nor do the world-level properties such as that of having no living ectoplasm. Not all the facts about our world follow from the physical facts alone.

To analyze the problem, note that these facts all involve negative existence claims, and so depend not only on what is going on in our world but on what is not. We cannot expect these facts to be determined by any sort of localized facts, as they depend not just on local goings-on in the world but on the world's limits. Supervenience theses should apply only to *positive* facts and properties, those that cannot be negated simply by enlarging a world. We can define a positive fact in *W* as one that holds in every world that contains *W* as a proper part; [14] a positive property is one that if instantiated in a world *W*, is also instantiated by the corresponding individual in all worlds that contain *W* as a proper part.[15] Most everyday facts and properties are positive—think of the property of being a kangaroo, or of being six feet tall, or of having a child. Negative facts and properties will always involve negative existence claims in one form or another. These include explicitly negative existential facts such as the nonexistence of ectoplasm, universally quantified facts such as the fact that all living things are made of DNA, negative relational properties such as childlessness, and superlatives such as the property of being the most fecund organism in existence.

In future, the supervenience relations with which we are concerned should be understood to be restricted to positive facts and properties. When claiming that biological properties supervene on physical properties, it is only the positive biological properties that are at issue. All the properties with which we are concerned are positive—local physical and phenomenal properties, for instance—so this is not much of a restriction.

The definition of global logical supervenience of B-properties on A-properties therefore comes to this: for any logically possible world *W* that is A-indiscernible from our world, then the B-facts true of our world are true of *W*. We need not build in a clause about positiveness, but it will usually be understood that the only relevant B-facts and properties are positive facts and properties. Similarly, B-properties supervene locally and logically on A-properties when for every actual individual *x* and every logically possible

individual *y*, if *y* is A-indiscernible from x, then the B-properties instantiated by *x* are instantiated by *y*. More briefly and more generally: B-properties supervene logically on A-properties if the B-facts about actual situations are entailed by the A-facts, where situations are understood as worlds and individuals in the global and local cases respectively. This definition captures the idea that supervenience claims are usually claims about our world, while retaining the key role of logical necessity. [16]

Supervenience and Materialism

Logical and natural supervenience have quite different ramifications for ontology: that is, for the matter of what there is in the world. If B-properties are logically supervenient on A-properties, then there is a sense in which once the A-facts are given, the B-facts are a free lunch. Once God (hypothetically) made sure that all the physical facts in our world held, the biological facts came along for free. The B-facts merely redescribe what is described by the A-facts. They may be *different* facts (a fact about elephants is not a microphysical fact), but they are not *further* facts.

With mere natural supervenience, the ontology is not so straightforward. Contingent lawful connections connect distinct features of the world. In general, if B-properties are merely naturally supervenient on A-properties in our world, then there *could* have been a world in which our A-facts held without the B-facts. As we saw before, once God fixed all the A-facts, in order to fix the B-facts he had more work to do. The B-facts are something over and above the A-facts, and their satisfaction implies that there is something new in the world.

With this in mind we can formulate precisely the widely held doctrine of *materialism* (or *physicalism*) which is generally taken to hold that everything in the world is physical, or that there is nothing over and above the physical, or that the physical facts in a certain sense exhaust all the facts about the world. In our language, materialism is true if all the positive facts about the world are globally logically supervenient on the physical facts. This captures the intuitive notion that if materialism is true, then once God fixed the physical facts about the world, all the facts were fixed.

(Or at least, all the positive facts were fixed. The restriction to positive facts is needed to ensure that worlds with extra ectoplasmic facts do not count against materialism in our world. Negative existential facts such as "There are no angels" are not strictly logically supervenient on the physical, but their nonsupervenience is quite compatible with materialism. In a sense, to fix the negative facts, God had to do more than fix the physical facts; he also had to declare, "That's all." If we wanted, we could add a second-order "That's all" fact to the supervenience base in the definition of materialism, in which case the positive-fact constraint could be removed.)

According to this definition, materialism is true if all the positive facts about our world are entailed by the physical facts. [17] That is, materialism is true if for any logically possible world *W* that is physically indiscernible from our world, all the positive facts true of our world are true of *W*. This is equivalent in turn to the thesis that any world that is physically indiscernible from our world contains a copy of our world as a (proper or improper) part, which seems an intuitively correct definition.[18] (This matches the definition of physicalism given by Jackson [1994], whose criterion is that every minimal physical duplicate of our world is a duplicate *simpliciter* of our world.[19])

I will discuss this matter at much greater length in Chapter 4, where this definition of materialism will be further justified. Some may object to the use of logical possibility rather than possibility *tout court* or "metaphysical possibility." Those people may substitute metaphysical possibility for logical possibility in the definition above. Later, I will argue that it comes to the same thing.

2—
Reductive Explanation

The remarkable progress of science over the last few centuries has given us good reason to believe that there is very little that is utterly mysterious about the world. For almost every natural phenomenon above the level of microscopic physics, there seems in principle to exist a *reductive explanation:* that is, an explanation wholly in terms of simpler entities. In these cases, when we give an appropriate account of lower-level processes, an explanation of the higher-level phenomenon falls out.

Biological phenomena provide a clear illustration. Reproduction can be explained by giving an account of the genetic and cellular mechanisms that allow organisms to produce other organisms. Adaptation can be explained by giving an account of the mechanisms that lead to appropriate changes in external function in response to environmental stimulation. Life itself is explained by explaining the various mechanisms that bring about reproduction, adaptation, and the like. Once we have told the lower-level story in enough detail, any sense of fundamental mystery goes away: the phenomena that needed to be explained have been explained.

One can tell a similar story for most natural phenomena. In physics, we explain heat by telling an appropriate story about the energy and excitation of molecules. In astronomy, we explain the phases of the moon by going into the details of orbital motion and optical reflection. In geophysics, earthquakes are explained via an account of the interaction of subterranean masses. In cognitive science, to explain a phenomenon such as learning, all we have to do is explain various functional mechanisms—the mechanisms that give rise to appropriate changes in behavior in response to environmental stimulation, at a first approximation (any worries about the *experience* of learning aside). Many of the details of these explanations currently evade our grasp, and are likely to prove very complex, but we know that if we find out enough about the low-level story, the high-level story will eventually come along.

I will not precisely define the notion of reductive explanation until later. For now, it remains characterized by example. However, I can issue some caveats about what reductive explanation is not. A reductive explanation of a phenomenon need not require a *reduction* of that phenomenon, at least in some senses of that ambiguous term. In a certain sense, phenomena that can be realized in many different physical substrates—learning, for example—might not be reducible in that we cannot *identify* learning with any specific lower-level phenomenon. But this multiple realizability does not stand in the way of reductively *explaining* any instance of learning in terms of lower-level phenomena. [20] Reductive explanation of a phenomenon should also not be confused with a reduction of a high-level *theory*. Sometimes a reductive explanation of a phenomenon will provide a reduction of a pre-existing high-level theory, but other times it will show such theories to be on the wrong track. Often there might not even be a high-level theory to reduce.

Reductive explanation is not the be-all and end-all of explanation. There are many other sorts of explanation, some of which may shed more light on a phenomenon than a reductive explanation in a given instance. There are *historical* explanations, for instance, explaining the genesis of a phenomenon such as life, where a reductive explanation only gives a synchronic account of how living systems function. There are also all sorts of *high-level* explanations, such as the explanation of aspects of behavior in terms of beliefs and desires. Even though this behavior might in principle be explainable reductively, a high-level explanation is often more comprehensible and enlightening. Reductive explanations should not be seen as displacing these other sorts of explanation. Each has its place.

Reductive Explanation via Functional Analysis

What is it that allows such diverse phenomena as reproduction, learning, and heat to be reductively explained? In all these cases, the nature of the concepts required to characterize the phenomena is crucial. If someone objected to a cellular explanation of reproduction, "This explains how a cellular process can lead to the production of a complex physical entity that is similar to the original entity, but it doesn't explain *reproduction*," we would have little patience—for that is all that "reproduction" *means*. In general, a reductive explanation of a phenomenon is accompanied by some rough-and-ready *analysis* of the phenomenon in question, whether implicit or explicit. The notion of reproduction can be roughly analyzed in terms of

the ability of an organism to produce another organism in a certain sort of way. It follows that once we have explained the processes by which an organism produces another organism, we have explained that instance of reproduction.

The point may seem trivial, but the possibility of this kind of analysis undergirds the possibility of reductive explanation in general. Without such an analysis, there would be no explanatory bridge from the lower-level physical facts to the phenomenon in question. With such an analysis in hand, all we need to do is to show how certain lower-level physical mechanisms allow the analysis to be satisfied, and an explanation will result.

For the most interesting phenomena that require explanation, including phenomena such as reproduction and learning, the relevant notions can usually be analyzed *functionally*. The core of such notions can be characterized in terms of the performance of some function or functions (where "function" is taken causally rather than teleologically), or in terms of the capacity to perform those functions. It follows that once we have explained how those functions are performed, then we have explained the phenomena in question. Once we explain how an organism performs the function of producing another organism, we have explained reproduction, for all it means to reproduce is to perform that function. The same goes for an explanation of learning. All it means for an organism to learn, roughly, is for its behavioral capacities to adapt appropriately in response to environmental stimulation. If we explain how the organism is able to perform the relevant functions, then we have explained learning.

(At most, we may have failed to explain any *phenomenal* aspects of learning, which I leave aside here for obvious reasons. If there is a phenomenal element to the concept of learning, then that part of learning may go unexplained; but I concentrate on the psychological aspects of learning here, which are plausibly the core of the concept.)

Explaining the performance of these functions is quite straightforward, in principle. As long as the results of such functions are themselves characterizable physically, and all physical events have physical causes, then there should be a physical explanation for the performance of any such function. One need only show how certain sorts of states are responsible for the production of appropriate resultant states, by a causal process in accord with the laws of nature. Of course the details of this kind of physical explanation can be nontrivial. Indeed, the details constitute the vast bulk of any reductive explanation, while the analysis component is often trivial. But once the relevant details are in, a story about low-level physical causation will explain how the relevant functions are performed, and will therefore explain the phenomenon in question.

Even a physical notion such as heat can be construed functionally: roughly, heat is the kind of thing that expands metals, is caused by fire, leads to a

particular sort of sensation, and the like. Once we have an account of how these various causal relations are fulfilled, then we have an account of heat. Heat is a *causal-role concept*, characterized in terms of what it is typically caused by and of what it typically causes, under appropriate conditions. Once empirical investigation shows how the relevant causal role is played, the phenomenon is explained.

There are some technical complications here, but they are inessential. For example, Kripke (1980) has pointed out a difference between a term such as "heat" and the associated description of a causal role: given that heat is realized by the motion of molecules, then the motion of molecules might qualify as heat in a counterfactual world, whether or not those molecules play the relevant causal role. It remains the case, however, that *explaining* heat involves explaining the fulfillment of the causal role, rather than explaining the motion of molecules. To see this, note that the equivalence of heat with the motion of molecules is known *a posteriori:* we know this *as a result* of explaining heat. The concept of heat that we had *a priori—before* the phenomenon was explained—was roughly that of "the thing that plays this causal role in the actual world." Once we discover how that causal role is played, we have an explanation of the phenomenon. As a bonus, we know what heat *is*. It is the motion of molecules, as the motion of molecules is what plays the relevant causal role in the actual world.

A second minor complication is that many causal-role concepts are somewhat ambiguous between the state that plays a certain causal role and the actual performance of that role. "Heat" can be taken to denote either the molecules that do the causal work or the causal process (heating) itself. Similarly, "perception" can be used to refer to either the act of perceiving or the internal state that arises as a result. Nothing important turns on this ambiguity, however. An explanation of how the causal role is played will explain heat or perception in either of these senses.

A third complication is that many causal-role concepts are partly characterized in terms of their effect on *experience:* for example, heat is naturally construed as the cause of heat sensations. Does this mean that we have to explain heat sensations before we can explain heat? Of course, we have no good account of heat sensations (or of experience generally), so what happens in practice is that that part of the phenomenon is left unexplained. If we can explain how molecular motion comes about in certain conditions, and causes metals to expand, and stimulates our skin in certain ways, then the observation that this motion is *correlated* with heat sensations is good enough. From the correlation, we infer that there is almost certainly a causal connection. To be sure, no explanation of heat will be complete until we have an account of how that causal connection works, but the incomplete account is good enough for most purposes. It is somewhat paradoxical that we end up explaining almost everything about a *phenomenon* except for the details of

how it affects our phenomenology, but it is not a problem in practice. It would not be a happy state of affairs if we had to put the rest of science on hold until we had a theory of consciousness.

Reductive Explanations in Cognitive Science

The paradigm of reductive explanation via functional analysis works beautifully in most areas of cognitive science, at least in principle. As we saw in the previous chapter, most nonphenomenal mental concepts can be analyzed functionally. Psychological states are characterizable in terms of the causal role they play. To explain these states, we explain how the relevant causation is performed.

In principle, one can do this by giving an account of the underlying neurophysiology. If we explain how certain neurophysiological states are responsible for the performance of the functions in question, then we have explained the psychological state. We need not always descend to the neurophysiological level, however. We can frequently explain some aspect of mentality by exhibiting an appropriate *cognitive model*—that is, by exhibiting the details of the abstract causal organization of a system whose mechanisms are sufficient to perform the relevant functions, without specifying the physiochemical substrate in which this causal organization is implemented. In this way, we give a *how-possibly* explanation of a given aspect of psychology, in that we have shown how the appropriate causal mechanisms *might* support the relevant mental processes. If we are interested in explaining the mental states of an *actual* organism or type of organism (e.g., learning in humans, as opposed to the possibility of learning in general), this sort of explanation must be supplemented with a demonstration that the causal organization of the model mirrors the causal organization of the organism in question.

To explain the possibility of learning, we can exhibit a model whose mechanisms lead to the appropriate changes in behavioral capacity in response to various kinds of environmental stimulation—a connectionist learning model, for example. To explain human learning, we must also demonstrate that such a model reflects the causal organization responsible for the performance of such functions in humans. The second step is usually difficult: we cannot exhibit such a correspondence directly, due to our ignorance of neurophysiology, so we usually have to look for indirect evidence, such as qualitative similarities in patterns of response, measurements of timing, and the like. This is one reason why cognitive science is currently in an undeveloped state. But as usual, the in-principle possibility of such explanation is a straightforward consequence of the functional nature of psychological concepts.

Unfortunately, the kind of functional explanation that works so well for psychological states does not seem to work in explaining phenomenal states.

The reason for this is straightforward. Whatever functional account of human cognition we give, there is a *further question:* Why is this kind of functioning accompanied by consciousness? No such further question arises for psychological states. If one asked about a given functional model of learning, "Why is this functioning accompanied by learning?" the appropriate answer is a semantic answer: "Because all it *means* to learn is to function like this." There is no corresponding analysis of the concept of consciousness. Phenomenal states, unlike psychological states, are not defined by the causal roles that they play. It follows that explaining how some causal role is played is not sufficient to explain consciousness. After we have explained the performance of a given function, the fact that consciousness accompanies the performance of the function (if indeed it does) remains quite unexplained.

One can put the point the following way. Given an appropriate functional account of learning, it is simply *logically impossible* that something could instantiate that account without learning (except perhaps insofar as learning requires consciousness). However, no matter what functional account of cognition one gives, it seems logically possible that that account could be instantiated without any accompanying consciousness. It may be naturally impossible—consciousness may in fact *arise* from that functional organization in the actual world—but the important thing is that the notion is logically coherent.

If this is indeed logically possible, then any functional and indeed any physical account of mental phenomena will be fundamentally incomplete. To use a phrase coined by Levine (1983), there is an *explanatory gap* between such accounts and consciousness itself. Even if the appropriate functional organization always gives rise to consciousness in practice, the question of *why* it gives rise to consciousness remains unanswered. This point will be developed at length later.

If this is so, it follows that there will be a partial explanatory gap for any mental concept that has a phenomenal element. If conscious experience is required for belief or learning, for example, we may not have a fully reductive explanation for belief or learning. But we at least have reason to believe that the *psychological* aspects of these mental features—which are arguably at the core of the relevant concepts—will be susceptible to reductive explanation in principle. If we leave worries about phenomenology aside, cognitive science seems to have the resources to do a good job of explaining the mind.

3— Logical Supervenience and Reductive Explanation

The epistemology of reductive explanation meets the metaphysics of supervenience in a straightforward way. A natural phenomenon is reductively explainable in terms of some low-level properties precisely when it is logically

supervenient on those properties. It is reductively explainable in terms of physical properties—or simply "reductively explainable"—when it is logically supervenient on the physical.

To put things more carefully: A natural phenomenon is reductively explainable in terms of some lower-level properties if the property of instantiating that phenomenon is globally logically supervenient on the low-level properties in question. A phenomenon is reductively explainable *simpliciter* if the property of exemplifying that phenomenon is globally logically supervenient on physical properties.

This can be taken as an *explication* of the notion of reductive explanation, with perhaps an element of stipulation. That our prior notion of reductive explanation implies (global) logical supervenience should be clear from the earlier discussion. If the property of exemplifying a phenomenon fails to supervene logically on some lower-level properties, then given any lower-level account of those properties, there will always be a further unanswered question: Why is this lower-level process accompanied by the phenomenon? Reductive explanation requires some kind of analysis of the phenomenon in question, where the low-level facts imply the realization of the analysis. So reductive explanation requires a logical supervenience relation. For example, it is precisely because reproduction is logically supervenient on lower-level facts that it is reductively explainable in terms of those facts.

That logical supervenience *suffices* for reductive explainability is somewhat less clear. If a phenomenon *P* supervenes logically on some lower-level properties, then given an account of the lower-level facts associated with an instance of *P*, the exemplification of *P* is a logical consequence. An account of the lower-level facts will therefore automatically yield an explanation of *P*. Nevertheless, such an explanation can sometimes seem unsatisfactory, for two reasons. First, the lower-level facts might be a vast hotchpotch of arbitrary-seeming details without any clear explanatory unity. An account of all the molecular motions underlying an instance of learning might be like this, for example. Second, it is possible that different instances of *P* might be accompanied by very different sets of low-level facts, so that explanations of particular instances do not yield an explanation of the phenomenon as a type.

One option is to hold that logical supervenience is merely *necessary* for reductive explanation, rather than sufficient. This is all that is required for my arguments about consciousness in the next chapter. But it is more useful to note that there is *a* useful notion of reductive explanation such that logical supervenience is both necessary and sufficient. Instead of taking the problems above as indicating that the accounts in question are not *explanations*, we can instead take them to indicate that a reductive explanation is not necessarily an *illuminating* explanation. Rather, a reductive explanation is a *mystery-removing* explanation.

As I noted earlier, reductive explanation is not the be-all and end-all of explanation. Its chief role is to remove any deep sense of mystery surrounding a high-level phenomenon. It does this by reducing the bruteness and arbitrariness of the phenomenon in question to the bruteness and arbitrariness of lower-level processes. Insofar as the low-level processes may themselves be quite brute and arbitrary, a reductive explanation may not give us a ***deep*** understanding of a phenomenon, but it at least eliminates any sense that there is something "extra" going on.

The gap between a reductive explanation and an illuminating explanation can generally be closed much further than this, however. This is due to two basic facts about the physics of our world: ***autonomy*** and ***simplicity.*** Microphysical causation and explanation seem to be autonomous, in that every physical event has a physical explanation; the laws of physics are sufficient to explain the events of physics on their own terms. Further, the laws in question are reasonably simple, so that the explanations in question have a certain compactness. Both of these things might have been otherwise. We might have lived in a world in which there were brutely emergent fundamental laws governing the behavior of high-level configurations such as organisms, with an associated downward causation that overrides any relevant microphysical laws. (The British emergentists, such as Alexander [1920] and Broad [1925], believed our world to be something like this.) Alternatively, our world might have been a world in which the behavior of microphysical entities is governed only by a vast array of baroque laws, or perhaps a world in which microphysical behavior is lawless and chaotic. In worlds like these, there would be little hope of achieving an illuminating reductive explanation, as the bruteness of low-level accounts might never be simplified.

But the actual world, with its low-level autonomy and simplicity, seems to allow that sense can generally be made even of complex processes. The low-level facts underlying a high-level phenomenon often have a basic unity that allows for a comprehensible explanation. Given an instance of high-level causation, such as a released trigger causing a gun to fire, we can not only isolate a bundle of lower-level facts that fix this causation; we can also tell a fairly simple story about how the causation is enabled, by encapsulating those facts under certain simple principles. This may not always work. It may be the case that some domains, such as those of sociology and economics, are so far removed from the simplicity of low-level processes that illuminating reductive explanation is impossible, even if the phenomena are logically supervenient. If so, then so be it: we can content ourselves with high-level explanations of those domains, while noting that logical supervenience implies that there is a reductive explanation in principle, although perhaps one that only a superbeing could understand.

Note also that on this account reductive explanation is fundamentally ***particular,*** accounting for particular instances of a phenomenon, without necessarily accounting for all instances together. This is what we should expect. If a property can be instantiated in many different ways, we cannot expect a single explanation to cover all the instances. Temperature is instantiated quite differently in different media, for example, and there are different explanations for each. At a much higher level, it is most unlikely that there should be a single explanation covering all instances of murder. Still, there is frequently a certain unity across the explanation of particulars, in that a good explanation of one is often an explanation of many. This is again a consequence of the underlying simplicity of our world, rather than a necessary property of explanation. In our world, the simple unifying stories that one can tell about lower-level processes often apply across the board, or at least across a wide range of particulars. It is also frequently the case, especially in the biological sciences, that the particulars have a common ancestry that leads to a similarity in the low-level processes involved. So the second problem mentioned, that of unifying the explanations of specific instances of a phenomenon, is not as much of a problem as it might be. In any case, it is the explanation of particulars that is central.

There is much more that could be said about closing the gap between reductive explanation and illuminating explanation, but the matter deserves a lengthy treatment in its own right and is not too important for my purposes. What is most important is that if logical supervenience fails (as I will argue it does for consciousness), then *any* kind of reductive explanation fails, even if we are generous about what counts as explanation. Also important is that logical supervenience removes any residual *metaphysical* mystery about a high-level phenomenon, by reducing any brutality in that phenomenon to brutality in lower-level facts. Of secondary importance is that if logical supervenience holds, then some sort of reductive explanation is possible. Although such explanations can fail to be illuminating or useful, this failure is not nearly as fundamental as the failure of explanation in domains where logical supervenience does not hold.

Further Notes on Reductive Explanation

A few further notes: First, a practical reductive explanation of a phenomenon does not usually go all the way to the microphysical level. To do this would be enormously difficult, giving rise to all the brutality problems just discussed. Instead, high-level phenomena are explained in terms of some properties at a slightly more basic level, as when reproduction is explained in terms of cellular mechanisms, or the phases of the moon are explained in terms of orbital motion. In turn, one hopes that the more basic phenomena

will themselves be reductively explainable in terms of something more basic still. If all goes well, biological phenomena may be explainable in terms of cellular phenomena, which are explainable in terms of biochemical phenomena, which are explainable in terms of chemical phenomena, which are explainable in terms of physical phenomena. As for the physical phenomena, one tries to unify these as far as possible, but at some level physics has to be taken as brute: there may be no explanation of why the fundamental laws or boundary conditions are the way they are. This ladder of explanation is little more than a pipe dream at the moment, but significant progress has been made. Given logical supervenience, along with the simplicity and autonomy of the lowest level, this sort of explanatory connection between the sciences ought to be possible in principle. Whether the complexities of reality will make it practically infeasible is an open question.

Second, it is at least conceivable that a phenomenon might be reductively explainable in terms of lower-level properties without being reductively explainable *simpliciter.* This might happen in a situation where C-properties are logically supervenient on B-properties, and are therefore explainable in terms of B-properties, but where B-properties themselves are not logically supervenient on the physical. There is clearly one sense in which such an explanation is reductive and another sense in which it is not. For the most part, I will be concerned with reductive explanation in terms of the physical, or in terms of properties that are themselves explainable in terms of the physical, and so on. Even if the C-properties here are reductively explainable in a relative sense, their very existence implies the failure of reductive explanation in general.

Third, *local* logical supervenience is too stringent a requirement for reductive explanation. One can reductively explain even context-dependent properties of an individual by giving an account of how relevant environmental relations come to be satisfied. As long as a phenomenon is globally supervenient, it will be reductively explainable in terms of some lower-level facts, even if these are spread widely in space and time.

Fourth, in principle there are two projects in reductive explanation of a phenomenon such as life, learning, or heat. There is first a project of *explication,* where we clarify just what it is that needs to be explained, by means of analysis. Learning might be analyzed as a certain kind of adaptational process, for example. Second, there is a project of *explanation,* where we see how that analysis comes to be satisfied by the low-level facts. The first project is conceptual, and the second is empirical. For many or most phenomena, the conceptual stage will be quite trivial. For some phenomena, however, such as belief, explication can be a major hurdle in itself. In practice, of course, there is never a clean separation between the projects, as explication and explanation take place in parallel.

4—
Conceptual Truth and Necessary Truth*

In my account of supervenience and explanation, I have relied heavily on the notions of logical possibility and necessity. It is now time to say something more about this. The basic way to understand the logical necessity of a statement is in terms of its truth across all logically possible worlds. This requires some care in making sense of both the relevant class of worlds and the way that statements are evaluated in worlds; I will discuss this at some length later in this section. It is also possible to explicate the logical necessity of a statement as truth in virtue of meaning: a statement is logically necessary if its truth is ensured by the meaning of the concepts involved. But again, this requires care in understanding just how the "meanings" should be taken. I will discuss both of these ways of looking at things, and their relation, later in this section.

(As before, the notion of logical necessity is not to be identified with a narrow notion involving derivability in first-order logic, or some other syntactic formalism. Indeed, it is arguable that the justification of the axioms and rules in these formalisms depends precisely on their logical necessity in the broader, more primitive sense.)

All this requires taking seriously, at least to some extent, the notion of *conceptual truth*—that is, the notion that some statements are true or false simply by virtue of the meanings of the terms involved. Key elements of my discussion so far have depended on characterizations of various concepts. I have accounted for the reductive explanation of reproduction, for example, by arguing that low-level details entail that certain functions are performed, and that performance of these functions is all there is to the concept of reproduction.

The notion of conceptual truth has had a bad name in some circles since the critique by Quine (1951), who argued that there is no useful distinction between conceptual truths and empirical truths. The objections to these notions usually cluster around the following points:

1. Most concepts do not have definitions giving necessary and sufficient conditions (this observation has been made many times but is often associated with Wittgenstein 1953).

2. Most apparent conceptual truths are in fact revisable, and could be withdrawn in the face of sufficient empirical evidence (a point raised by Quine).

3. Considerations about *a posteriori* necessity, outlined by Kripke (1972), show that application-conditions of many terms across possible worlds cannot be known *a priori*.

These considerations count against an overly simplistic view of conceptual truth, but not against the way I am using these notions. In particular, it turns

out that the class of *supervenience conditionals*—"*If* the A-facts about a situation are *X*, then the B-facts are *Y*," where the A-facts fully specify a situation at a fundamental level—are unaffected by these considerations. These are the only conceptual truths that my arguments need, and we will see that none of the considerations above count against them. I will also analyze the relationship between conceptual truth and necessary truth in more detail, and spell out the role these play in understanding logical supervenience.

Definitions

The absence of cut-and-dried definitions is the least serious of the difficulties with conceptual truth. None of my arguments depend on the existence of such definitions. I occasionally rely on analyses of various notions, but these analyses need only be rough and ready, without any pretense at providing precise necessary and sufficient conditions. Most concepts (e.g., "life") are somewhat vague in their application, and there is little point trying to remove that vagueness by arbitrary precision. Instead of saying "A system is alive if and only if it reproduces, adapts with utility 800 or greater, and metabolizes with efficiency 75 percent, or exhibits these in a weighted combination with such-and-such properties," we can simply note that if a system exhibits these phenomena to a sufficient degree then it will be alive, by virtue of the meaning of the term. If an account of relevant low-level facts fixes the facts about a system's reproduction, utility, metabolism, and so on, then it also fixes the facts about whether the system is *alive,* insofar as that matter is factual at all.

We can sum this up with a schematic diagram (Figure 2.1) showing how a high-level property P might depend on two low-level parameters A and B, each of which can take on a range of values. If we had a crisp definition in terms of necessary and sufficient conditions, then we would have something like the picture at left, where the dark rectangle represents the region in which property P is instantiated. Instead, the dependence is invariably something like the picture at right, where the boundaries are vague and there is a large area in which the matter of P-hood is indeterminate, but there is also an area in which the matter is clear. (It may be indeterminate whether bacteria or computer viruses are alive, but there is no doubt that dogs are alive.) Given an example in the determinate area, exemplifying A and B to sufficient degrees that P is exemplified, the conditional "If *x* is A and B to this degree, then *x* is P" is a conceptual truth, despite the lack of a clean definition of P. Any indeterminacy in such conditionals, in the gray areas, will reflect indeterminacy in the facts of the matter, which is as it should be. The picture can straightforwardly be extended to dependence of a property on an arbitrary number of factors, and to supervenience conditionals in general.

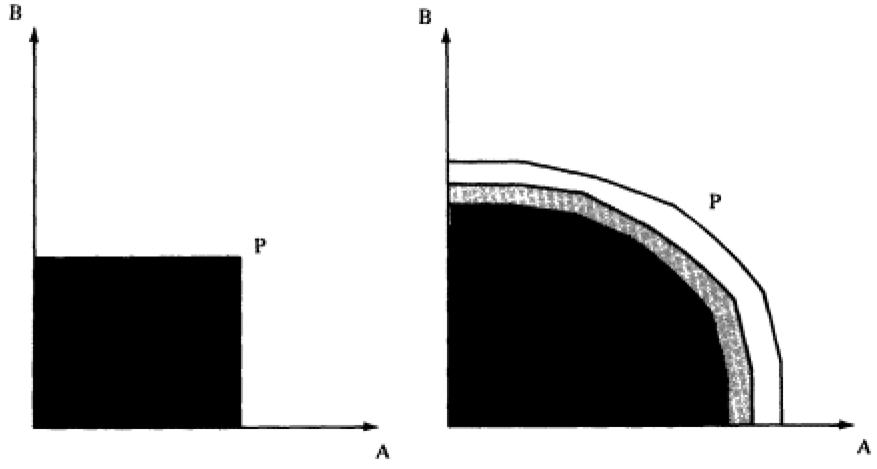

Figure 2.1.
Two ways in which a property P might depend on properties A and B.

Importantly, then, one set of facts can ***entail*** another set without there being a clean definition of the latter notions in terms of the former. The above case provides an example: there is no simple definition of P in terms of A and B, but the facts about A and B in an instance entail the facts about P. For another example, think about the ***roundness*** of closed curves in two-dimensional space (Figure 2.2). There is certainly no perfect definition of roundness in terms of simpler mathematical notions. Nevertheless, take the figure at left, specified by the equation $2x^2 + 3y^2 = 1$. There is a fact of the matter—this figure is round—insofar as there are ever facts about roundness at all (compare to the figure at right, which is certainly not round). Further, this fact is ***entailed*** by the basic description of the figure in mathematical terms—given that description, and the concept of roundness, the fact that the figure is round is determined. Given that A-facts can entail B-facts without a definition of B-facts in terms of A-facts, the notion of logical supervenience is unaffected by the absence of definitions. (In thinking about more complex issues and objections concerning logical supervenience, it may be worthwhile to keep this example in mind.)

We can put the point by saying that the sort of "meaning" of a concept that is relevant in most cases is not a definition, but an *intension*: a function specifying how the concept applies to different situations. Sometimes an intension might be summarizable in a definition, but it need not be, as these cases suggest. But as long as there is a fact of the matter about how concepts apply in various situations, then we have an intension; and as I will discuss shortly, this will generally be all the "meaning" that my arguments will need.

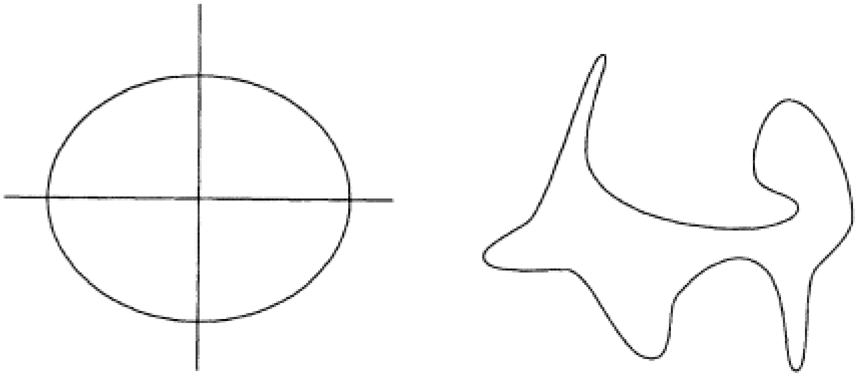

Figure 2.2.
The round curve $2x^2 + 3y^2 = 1$ and nonround friend.

Revisability

The second objection, raised by Quine (1951), is that purported conceptual truths are always subject to revision in the face of sufficient empirical evidence. For instance, if evidence forces us to revise various background statements in a theory, it is possible that a statement that once appeared to be conceptually true might turn out to be false.

This is so for many purported conceptual truths, but it does not apply to the supervenience conditionals that we are considering, which have the form "If the low-level facts turn out like this, then the high-level facts will be like that." The facts specified in the antecedent of this conditional effectively include all relevant empirical factors. Empirical evidence could show us that the antecedent of the conditional is false, but not that the conditional is false. In the extreme case, we can ensure that the antecedent gives a full specification of the low-level facts about the world. The very comprehensiveness of the antecedent ensures that empirical evidence is irrelevant to the conditional's truth-value. (This picture is somewhat complicated by the existence of *a posteriori* necessities, which I discuss shortly. Here, I am only concerned with epistemic conditionals about ways the actual world might turn out.)

While considerations about revisability provide a plausible argument that there are not many *short* conceptual truths, nothing in these considerations counts against the constrained, complex sort of conceptual truth that I have been concerned with. The upshot of these observations is that the truth-conditions of a high-level statement may not be easily *localizable,* as all sorts of factors might have some kind of indirect relevance; but the global truth conditions provided by a supervenience conditional are not threatened. Indeed, if meaning determines a function from possible worlds to reference classes (an intension), and if possible worlds are finitely describable (in

terms of arrangement of basic qualities in those worlds, say), then there will automatically be a vast class of conceptually true conditionals that result.

A Posteriori Necessity

It has traditionally been thought that all conceptual truths are knowable *a priori,* as are all necessary truths, and that the classes of *a priori* truths, necessary truths, and conceptual truths are closely related or even coextensive. Saul Kripke's book *Naming and Necessity* (1972) threw a wrench into this picture by arguing that there is a large class of necessarily true statements whose truth is not knowable *a priori.* An example is the statement "Water is H_2O." We cannot know this to be true *a priori;* for all we know (or for all we knew at the beginning of inquiry), water is made out of something else, perhaps XYZ. Kripke argues that nevertheless, given that water is H_2O in the actual world, then water is H_2O in all possible worlds. It follows that "Water is H_2O" is a necessary truth despite its *a posteriori* nature.

This raises a few difficulties for the framework I have presented. For example, on some accounts these necessary truths are conceptual truths, implying that not all conceptual truths are knowable *a priori.* On alternative accounts, such statements are not conceptual truths, but then the link between conceptual truth and necessity is broken. At various points in this book, I use *a priori* methods to gain insight into necessity; this is the sort of thing that Kripke's account is often taken to challenge.

On analysis, I think it can be seen that these complications do not change anything fundamental to my arguments; but it worth taking the trouble to get clear about what is going on. I will spend some time setting up a systematic framework for dealing with these issues, which will recur. In particular, I will present a natural way of capturing Kripke's insights in a two-dimensional picture of meaning and necessity. This framework is a synthesis of ideas suggested by Kripke, Putnam, Kaplan, Stalnaker, Lewis, Evans, Davies and Humberstone, and others who have addressed these two-dimensional phenomena.

On the traditional view of reference, derived from Frege although cloaked here in modern terminology, a concept determines a function *f: W* ® *R* from possible worlds to referents. Such a function is often called an *intension;* together with a specification of a world w, it determines an *extension* $f(w)$ In Frege's own view, every concept had a *sense,* which was supposed to determine the reference of the concept depending on the state of the world; so these senses correspond closely to intensions. The sense was often thought of as the *meaning* of the concept in question.

More recent work has recognized that no single intension can do all the work that a meaning needs to do. The picture developed by Kripke complicates things by noting that reference in the actual world and in counterfactual

possible worlds is determined by quite different mechanisms. In a way, the Kripkean picture can be seen to split the Fregean picture into two separate levels.

Kripke's insight can be expressed by saying that there are in fact *two* intensions associated with a given concept. That is, there are two quite distinct patterns of dependence of the referent of a concept on the state of the world. First, there is the dependence by which reference is fixed in the *actual* world, depending on how the world turns out: if it turns out one way, a concept will pick out one thing, but if it turns out another way, the concept will pick out something else. Second, there is the dependence by which reference in *counterfactual* worlds is determined, given that reference in the actual world is already fixed. Corresponding to each of these dependencies is an intension, which I will call the *primary* and *secondary* intensions, respectively.

The *primary* intension of a concept is a function from worlds to extensions reflecting the way that actual-world reference is fixed. In a given world, it picks out what the referent of the concept would be if that world turned out to be actual. Take the concept "water." If the actual world turned out to have XYZ in the oceans and lakes, then "water" would refer to XYZ, [21] but given that it turns out to have H_2O in the oceans and lakes, "water" refers to H_2O. So the primary intension of "water" maps the XYZ world to XYZ, and the H_2O world to H_2O. At a rough approximation, we might say that the primary intension picks out the dominant clear, drinkable liquid in the oceans and lakes; or more briefly, that it picks out the *watery stuff* in a world.

However, *given* that "water" turns out to refer to H_2O in the actual world, Kripke notes (as does Putnam [1975]) that it is reasonable to say that water is H_2O in every counterfactual world. The *secondary intension* of "water" picks out the water in every counterfactual world; so if Kripke and Putnam are correct, the secondary intension picks out H_2O in all worlds.[22]

It is the primary intension of a concept that is most central for my purposes: for a concept of a natural phenomenon, it is the primary intension that captures what needs explaining. If someone says, "Explain water," long before we know that water is in fact H_2O, what they are asking for is more or less an explanation of the clear, drinkable liquid in their environment. It is only *after* the explanation is complete that we know that water is H_2O. The primary intension of a concept, unlike the secondary intension, is independent of empirical factors: the intension *specifies* how reference depends on the way the external world turns out, so it does not itself depend on the way the external world turns out.

Of course, any brief characterization of the primary intension of a concept along the lines of "the dominant clear, drinkable liquid in the environment" will be a simplification. The true intension can be determined only from detailed consideration of specific scenarios: What would we say if the world turned out this way? What would we say if it turned out that way? For example, if it had turned out that the liquid in lakes was H_2O and the liquid in oceans XYZ, then we probably would have said that both were water; if the stuff in oceans and lakes was a mixture of 95 percent A and 5 percent B, we would probably have said that A but not B was water; if it turned out that a substance neither clear not drinkable bore an appropriate microphysical relation to the clear, drinkable liquid in our environment, we would probably call that substance "water" too (as we do in the case of ice or of "dirty water"). The full conditions for what it takes to qualify as "water" will be quite vague at the edges and need not be immediately apparent on reflection, but none of this makes much difference to the picture I am describing. I will use "watery stuff" as a term of art to encapsulate the primary intension, whatever it is. [23]

In certain cases, the decision about what a concept refers to in the actual world involves a large amount of reflection about what is the most reasonable thing to say; as, for example, with questions about the reference of "mass" when the actual world turned out to be one in which the theory of general relativity is true,[24] or perhaps with questions about what qualifies as "belief" in the actual world. So consideration of just what the primary intension picks out in various actual-world candidates may involve a corresponding amount of reflection. But this is not to say that the matter is not a *priori:* we have the ability to engage in this reasoning independently of how the world turns out. Perhaps the reports of experiments confirming relativity are disputed, so we are not sure whether the actual world has turned out to be a relativistic world: either way, we have the ability to reason about what "mass" will refer to *if* that state of affairs turns out to be actual.

(Various intricacies arise in analyzing the primary intensions of concepts used by individuals within a linguistic community. These might be handled by noting that an individual's concept may have a primary intension that involves deference to a surrounding community's concept—so my concept "elm" might pick out what those around me call "elms"; but in any case this sort of problem is irrelevant to the issues I will be concerned with, for which we might as well assume that there is just one person in the community, or that all individuals are equally well informed, or even that the community is a giant individual. There are also a few technical problems that might come up in using primary intensions to build a general semantic theory—for example, is the reference of a concept essential to the concept? Might different speakers associate different primary intensions with the same word? But I am not trying to build a full semantic theory here, and we can abstract away from this sort of concern.

Sometimes philosophers are suspicious of entities such as primary intensions because they see them as reminiscent of a "description" theory of reference. But descriptions play no essential part in this framework; I use

them merely to flesh out some of the character of the relevant functions from possible worlds to extensions. It is the function itself, rather than any summarizing description, that is truly central. This picture is quite compatible with the "causal" theory of reference: we need simply note that the primary intension of a concept such as "water" may require an appropriate causal connection between the referent and the subject. Indeed, we are led to believe in a causal theory of reference in the first place precisely by considering various ways the actual world might turn out, and noting what the referent of the concept would turn out to be in those cases; that is, by evaluating the primary intension of a concept at those worlds.)

Given that the actual-world reference of "water" is fixed by picking out the watery stuff, one might think that water is watery stuff in all possible worlds. Kripke and Putnam pointed out that this is not so: if water is H_2O in the actual world, then water is H_2O in all possible worlds. In a world (Putnam's "Twin Earth") in which the dominant clear, drinkable liquid is XYZ rather than H_2O, this liquid is not water; it is merely watery stuff. All this is captured by the *secondary* intension of "water," which picks out the water in all worlds: that is, it picks out H_2O in all worlds.

The secondary intension of a concept such as "water" is not determined *a priori,* as it depends on how things turn out in the actual world. But it still has a close relation to the primary intension above. In this case, the secondary intension is determined by first evaluating the primary intension at the actual world, and then *rigidifying* this evaluation so that the same sort of thing is picked out in all possible worlds. Given that the primary intension ("watery stuff") picks out H_2O in the actual world, it follows from rigidification that the secondary intension picks out H_2O in all possible worlds.

We can sum this up by saying "water" is conceptually equivalent to "*dthat* (watery stuff)," where *dthat* is a version of Kaplan's rigidifying operator, converting an intension into a rigid designator by evaluation at the actual world (Kaplan 1979).The single Fregean intension has fragmented into two: a primary intension ("watery stuff") that fixes reference in the actual world, and a secondary intension ("H_2O ") that picks out reference in counterfactual possible worlds, and which depends on how the actual world turned out.

(There is sometimes a tendency to suppose that *a posteriori* necessity makes *a priori* conceptual analysis irrelevant, but this supposition is ungrounded. Before we even get to the point where rigid designation and the like become relevant, there is a story to tell about what makes an actual-world *X qualify* as the referent of "*X*" in the first place. This story can only be told by an analysis of the primary intension. And this project is an *a priori* enterprise, as it involves questions about what our concept *would* refer to if the actual world turned out in various ways. Given that we have the ability to know what our concepts refer to when we know how the actual world turns out, then we have the ability to know what our concepts would

refer to *if* the actual world turned out in various ways. Whether or not the actual world *does* turn out a certain way makes little difference in answering this question, except in focusing our attention.)

Both the primary and secondary intensions can be seen as functions *f* : *W* ® *R* from possible worlds to extensions, where the possible worlds in question are seen in subtly different ways. We might say that the primary intension picks out the referent of a concept in a world when it is *considered as actual*—that is, when it is considered as a candidate for the actual world of the thinker—whereas the secondary intension picks out the referent of a concept in a world when it is *considered as counterfactual,* given that the actual world of the thinker is already fixed. When the XYZ world is considered as actual, my term "water" picks out XYZ in the world, but when it is considered as counterfactual, "water" picks out H_2O.

The distinction between these two ways of looking at worlds corresponds closely to Kaplan's (1989) distinction between the *context of utterance* of an expression and the *circumstances of evaluation.* When we consider a world *w* as counterfactual, we keep the actual world as the context of utterance, but use *w* as a circumstance of evaluation. For example, if I utter "There is water in the ocean" in this world and *evaluate* it in the XYZ world, "water" refers to H_2O and the statement is false. But when we consider *w a*s actual, we think of it as a potential context of utterance, and wonder how things would be if the context of the expression turned out to be *w.* If the context of my sentence "There is water in the ocean" turned out to be the XYZ world, then the statement would be true when evaluated at that world. The primary intension is therefore closely related to what Kaplan calls the *character* of a term, although there are a few differences, [25] and the secondary intension corresponds to what he calls a term's *content.*

There is a slight asymmetry in that a context of utterance but not the circumstance of evaluation is what Quine (1969) calls a *centered* possible world. This is an ordered pair consisting of a world and a *center* representing the viewpoint within that world of an agent using the term in question: the center consists in (at least) a "marked" individual and time. (This suggestion comes from Lewis 1979; Quine suggests that the center might be a point in space-time.) Such a center is necessary to capture the fact that a term like "water" picks out a different extension for me than for my twin on Twin Earth, despite the fact that we live in the same universe.[26] It is only our position in the universe that differs, and it is this position that makes a relevant difference to the reference-fixing process.

This phenomenon arises in an especially obvious way for indexical terms such as "I", whose reference clearly depends on who is using the term and not just on the overall state of the world: the primary intension of "I" picks out the individual at the center of a centered world. (The secondary intension of my concept "I" picks out David Chalmers in all possible worlds.)

Page 61

There is a less overt indexical element in notions such as "water," however, which can be roughly analyzed as "*dthat*(*the* dominant clear, drinkable liquid *in our environment*)." [27] It is this indexical element that requires primary intensions to depend on centered worlds. Once actual-world reference is fixed, however, no center is needed to evaluate reference in a counterfactual world. The circumstance of evaluation can therefore be represented by a simple possible world without a center.

All this can be formalized by noting that the full story about reference in counterfactual worlds is not determined *a priori* by a singly indexed function *f* : *W* ® *R.* Instead, reference in a counterfactual world depends both on that world and on the way the actual world turns out. That is, a concept determines a doubly indexed function

*F: W** × *W* ® *R*

where *W** is the space of centered possible worlds, and *W* is the space of ordinary possible worlds. The first parameter represents contexts of utterance, or ways the actual world might turn out, whereas the second parameter represents circumstances of evaluation, or counterfactual possible worlds. Equivalently, a concept determines a family of functions

F_v : *W* ® *R*

for each *v* Î *W** representing a way the actual world might turn out, where $F_v(w) = F(v, w)$ For "water," if *a* is a world in which watery stuff is H_2O, then F_a picks out H_2O in any possible world. Given that in our world water *did* turn out to be H_2O, this F_a specifies the correct application conditions for "water" across counterfactual worlds. If our world had turned out to be a different world *b* in which watery stuff was XYZ, then the relevant application conditions would have been specified by F_b a different intension which picks out XYZ in any possible world.

The function F is determined *a priori,* as all *a posteriori* factors are included in its parameters. From F we can recover both of our singly indexed intensions. The primary intension is the function $f : W^*$ ® R determined by the "diagonal" mapping $f : w$ |® $F(w, w')$, where w' is identical to w except that the center is removed. This is the function whereby reference in the actual world is fixed. The secondary intension is the mapping $F_a : w$ |® $F(a, w)$ where a is our actual world. This intension picks out reference in counterfactual worlds. An immediate consequence is that the primary intension and secondary intension coincide in their application to the actual world: $f(a) = F_a(a') = F(a, a')$.

In the reverse direction, the doubly indexed function F and therefore the secondary intension F_a can usually be derived from the primary intension f,

with the aid of a "rule" about how the secondary intension depends on the primary intension and the actual world a. This rule depends on the type of concept. For a concept that is a rigid designator, the rule is that in a world w, the secondary intension picks out in w whatever the primary intension picks out in a (or perhaps, for natural-kind terms, whatever has the same underlying structure as what the primary intension picks out in a). More formally, let $D: R \times W$ ® R be a "projection" operator that goes from a class picked out in some world to members of "that" class in another possible world. Then the secondary intension F_a is just the function $D(f(a),–)$ which we can think of as *dthat* applied to the intension given by f.

For other concepts, derivation of the secondary intension from the primary intension will be easier. With "descriptive" expressions such as "doctor," "square," and "watery stuff," rigid designation plays no special role: they apply to counterfactual worlds independently of how the actual world turns out. In these cases, the secondary intension is a simple copy of the primary intension (except for differences due to centering). The framework I have outlined can handle both sorts of concepts.

Property terms, such as "hot," can be represented in one of two ways in an intensional framework. We can see the intension of a property as a function from a world to a class of individuals (the individuals that irstantiate the property), or from a world to properties themselves. Either way of doing things is compatible with the current framework: we can easily find a primary and a secondary intension in either case, and it is easy to move back and forth between the two frameworks. I will usually do things the first way, however, so that the primary intension of "hot" picks out the entities that qualify as "hot" things in the actual world, depending on how it turns out, and the secondary intension picks out the hot things in a counterfactual world, given that the actual world has turned out as it has.

Both the primary and the secondary intensions can be thought of as candidates for the "meaning" of a concept. I think there is no point choosing one of these to qualify as *the* meaning; the term "meaning" here is largely an honorific. We might as well think of the primary and secondary intensions as the *a priori* and *a posteriori* aspects of meaning, respectively.

If we make this equation, both of these intensions will back a certain kind of conceptual truth, or truth in virtue of meaning. The primary intension backs *a priori* truths, such as "Water is watery stuff." Such a statement will be true no matter how the actual world turns out, although it need not hold in all nonactual possible worlds. The secondary intension does not back *a priori* truths, but backs truths that hold in all counterfactual possible worlds, such as "Water is H_2O." Both varieties qualify as truths in virtue of meaning; they are simply true in virtue of different aspects of meaning.

Page 63

It is also possible to see both as varieties of *necessary* truth. The latter corresponds to the more standard construal of a necessary truth. The former, however, can also be construed as truth across possible worlds, as long as these possible worlds are construed as contexts of utterance, or as ways the actual world might turn out. On this subtly different construal, a statement S is necessarily true if no matter how the actual world turns out, it would turn out that S was true. If the actual world turns out to be a world in which watery stuff is XYZ, then my statement "XYZ is water" will turn out to be true. So, according to this construal on which possible worlds are *considered as actual,* "Water is watery stuff" is a necessary truth.

This kind of necessity is what Evans (1979) calls "deep necessity," as opposed to "superficial" necessities like "Water is H2O." It is analyzed in detail by Davies and Humberstone (1980) by means of a modal operator they call "fixedly actually." Deep necessity, unlike superficial necessity, is unaffected by *a posteriori* considerations. These two varieties of possibility and necessity apply always to *statements.* There is only one relevant kind of possibility of *worlds;* the two approaches differ on how the truth of a statement is evaluated in a world.

We can see this in a different way by noting that there are two sets of *truth conditions* associated with any statement. If we evaluate the terms in a statement according to their primary intensions, we arrive at the *primary* truth conditions of the statement; that is, a set of centered possible worlds in which the statement, evaluated according to the primary intensions of the terms therein, turns out to be true. The primary truth conditions tell us how the actual world has to be for an utterance of the statement to be true in that world; that is, they specify those *contexts* in which the statement would turn out to be true. For instance, the primary truth conditions of "Water is wet" specify roughly that such an utterance will be true in the set of worlds in which watery stuff is wet.

If instead we evaluate the terms involved according to their secondary intensions, we arrive at the more familiar *secondary truth conditions.* These conditions specify the truth-value of a statement in counterfactual worlds, given that the actual world has turned out as it did. For instance, the secondary truth conditions of "Water is wet" (uttered in this world) specifies those worlds in which water is wet: so given that water is H_2O, it specifies those worlds in which H_2O is wet. Note that there is no danger of an ambiguity in actual-world truth: the primary and secondary truth conditions will always specify the same truth-value when evaluated at the actual world.

If we see a proposition as a function from possible worlds to truth-values, then these two sets of truth conditions yield two *propositions* associated with any statement. Composing the primary intensions of the terms involved yields a *primary proposition,* which holds in precisely those contexts of

Page 64

utterance in which the statement would turn out to express a truth. (This is the "diagonal proposition" of Stalnaker 1978. Strictly speaking, it is a centered proposition, or a function from centered worlds to truth-values.) The secondary intensions yield a *secondary* proposition, which holds in those counterfactual circumstances in which the statement, as uttered in the actual world, is true. The secondary proposition is Kaplan's "content" of an utterance and is more commonly seen as the proposition expressed by a statement, but the primary proposition is also central.

The two kinds of necessary truth of a statement correspond precisely to the necessity of the two kinds of associated proposition. A statement is necessarily true in the first (*a priori*) sense if the associated primary proposition holds in all centered possible worlds (that is, if the statement would turn out to express a truth in any context of utterance). A statement is necessarily true in the ***a posteriori*** sense if the associated secondary proposition holds in all possible worlds (that is, if the statement as uttered in the *actual* world is true in all counterfactual worlds). The first corresponds to Evans's deep necessity, and the second to the more familiar superficial necessity.

To illustrate, take the statement "Water is H_2O." The primary intensions of "water" and "H_2O" differ, so that we cannot know *a priori* that water is H_2O; the associated *primary* proposition is not necessary (it holds in those centered worlds in which the watery stuff has a certain molecular structure). Nevertheless, the secondary intensions coincide, so that "Water is H_2O" is true in all possible worlds when evaluated according to the secondary intensions—that is, the associated *secondary* proposition is necessary. Kripkean *a posteriori* necessity arises just when the secondary intensions in a statement back a necessary proposition, but the primary intensions do not.

Consider by contrast the statement "Water is watery stuff." Here the associated primary intensions of "water" and "watery stuff" are the same, so that we can know this statement to be true *a priori,* as long as we possess the concepts. The associated primary proposition is necessary, so that this statement is necessarily true in Evans's "deep" sense. However, the secondary intensions differ, as "water" is rigidified but "watery stuff" is not: in a world where XYZ is the clear, drinkable liquid, the secondary intension of "watery stuff" picks out XYZ but that of "water" does not. The associated *secondary* proposition is therefore not necessary, and the statement is not a necessary truth in the more familiar sense; it is an example of Kripke's "contingent *a priori.*"

In general, many apparent "problems" that arise from these Kripkean considerations are a consequence of trying to squeeze the doubly indexed picture of reference into a single notion of meaning or of necessity. Such problems can usually be dissolved by explicitly noting the two-dimensional

character of reference, and by taking care to explicitly distinguish the notion of meaning or of necessity that is in question. [28]

It is also possible to use this two-dimensional framework to give an account of the semantics of *thought,* as well as of language. I do this at much greater length elsewhere (Chalmers 1994c). This aspect of the framework will not be central here, but it is worth mentioning, as it will come up in one or two minor places. The basic idea is very similar: given an individual's *concept* in thought, we can assign a primary intension corresponding to what it will pick out depending on how the actual world turns out, and a secondary intension corresponding to what it picks out in counterfactual worlds, given that the actual world turns out as it has. Given a *belief,* we can assign a primary proposition and a secondary proposition in a similar way (what I elsewhere call the "notional" and "relational" content of the belief).

For example, concepts such as "Hesperus" and "Phosphorus" will have different primary intensions (one picks out the evening star in a given centered world, the other picks out the morning star), but the same secondary intensions (both pick out Venus in all worlds). The thought "Hesperus is Phosphorus" will have a primary proposition true in all centered worlds in which the evening star is the morning star: the fact that this thought is informative rather than trivial corresponds to the fact that the primary proposition is contingent, as the primary intensions of the two terms differ.

The primary proposition, more than the secondary proposition, captures how things seem from the point of view of the subject: it delivers the set of centered worlds which the subject, in having the belief, is endorsing as potential environments in which he or she might be living (in believing that Hesperus is Phosphorus, I endorse all those centered worlds in which the evening star and the morning star around the center are identical). It is also fairly easy to argue that the primary proposition, rather than the secondary proposition, governs the cognitive and rational relations between thoughts. For this reason it is natural to think of the primary proposition as the *cognitive* content of a thought.[29]

Logical Necessity, Conceptual Truth, and Conceivability

With this framework in hand, we can spell out the relationships among logical necessity, conceptual truth, and conceivability. Starting with logical necessity: this is just necessity as explicated above. A statement is logically necessary if and only if it is true in all logically possible worlds. Of course we have two varieties of logical necessity of statements, depending on whether we evaluate truth in a possible world according to primary and secondary intensions. We might call these varieties *1-necessity* and *2-necessity,* respectively.

This analysis explicates the logical necessity and possibility of a *statement* in terms of (a) the logical possibility of *worlds,* and (b) the intensions determined by the terms involved in the statement. I have already discussed the intensions. As for the notion of a logically possible world, this is something of a primitive: as before, we can intuitively think of a logically possible world as a world that God might have created (questions about God himself aside). I will not engage the vexed question of the ontological status of these worlds, but simply take them for granted as a tool, in the same way one takes mathematics for granted. [30] As for the *extent* of the class, the most important feature is that every conceivable world is logically possible, a matter on which I will say more in a moment.

As for conceptual truth, if we equate meaning with intension (primary or secondary), it is easy to make the link between truth in virtue of meaning and logical necessity. If a statement is logically necessary, its truth will be an automatic byproduct of the intensions of the terms (and the compositional structure of the statement). We do not need to bring in the world in any further role, as the intensions in question will be satisfied in every possible world. Similarly, if a statement is true in virtue of its intensions, it will be true in every possible world.

As before, there are two varieties of conceptual truth, depending on whether we equate the "meanings" with primary or secondary intensions, paralleling the two varieties of necessary truth. As long as one makes parallel decisions in the two cases, a statement is conceptually true if and only if it is necessarily true. "Water is watery stuff" is conceptually true and necessarily true in the first sense; and "Water is H_2O" is conceptually true and necessarily true in the second. Only the first variety of conceptual truth will in general be accessible *a priori*. The second variety will include many *a posteriori* truths, as the secondary intension depends on the way the actual world turns out.

(I do not claim that intensions are *the* correct way to think of meanings. Meaning is a many-faceted notion, and some of its facets may not be perfectly reflected by intensions, so one could resist the equation of the two at least in some cases.[31] Rather, the equation of meaning and intension should here be thought of as stipulative: if one makes the equation, then one can make various useful connections. Not much rests on the use of the word "meaning." In any case, truth in virtue of intension is the only sort of truth in virtue of meaning that I will need.)

We can also make a link between the logical possibility of statements and the *conceivability* of statements, if we are careful. Let us say that a statement is conceivable (or conceivably true) if it is true in all conceivable worlds. This should not be confused with other senses of "conceivable." For example, there is a sense according to which a statement is conceivable if for all we know it is true, or if we do not know that it is impossible. In this sense, both

Goldbach's conjecture and its negation are conceivable. But the false member of the pair will not qualify as conceivable in the sense I am using, as there is no conceivable world in which it is true (it is false in every world) .

On this view of conceivability, the conceivability of a statement involves two things: first, the conceivability of a relevant world, and second, the truth of the statement in that world. [32] It follows that in making conceivability judgments, one has to make sure that one describes the world that one is conceiving correctly, by properly evaluating the truth of a statement in the world. One might at first glance think it is conceivable that Goldbach's conjecture is false, by conceiving of a world where mathematicians announce it to be so; but if in fact Goldbach's conjecture is true, then one is *misdescribing* this world; it is really a world in which the conjecture is true and some mathematicians make a mistake.

In practice, to make a conceivability judgment, one need only consider a conceivable *situation*—a small part of a world—and then make sure that one is describing it correctly. If there is a conceivable situation in which a statement is true, there will obviously be a conceivable world in which the statement is true, so this method will give reasonable results while straining our cognitive resources less than conceiving of an entire world!

Sometimes it is said that examples such as "Water is XYZ" show that conceivability does not imply possibility, but I think the situation is subtler than this. In effect, there are two varieties of conceivability, which we might call *1-conceivability* and *2-conceivability,* depending on whether we evaluate a statement in a conceivable world according to the primary or secondary intensions of the terms involved. "Water is XYZ" is 1-conceivable, as there is a conceivable world in which the statement (evaluated according to primary intensions) is true, but it is not 2-conceivable, as there is no conceivable world in which the statement (evaluated according to secondary intension) is true. These two sorts of conceivability precisely mirror the two sorts of logical possibility mentioned previously.

Often, the conceivability of a statement is equated with 1-conceivability (the sense in which "Water is XYZ" is conceivable), as it is this sort of conceivability that is accessible *a priori*. And most often, the *possibility* of a statement is equated with 2-possibility (the sense in which "Water is XYZ" is impossible). Taken *this* way, conceivability does not imply possibility. But it remains the case that 1-conceivability implies 1-possibility, and 2-conceivability implies 2-possibility. One simply has to be careful not to judge 1-conceivability when 2-possibility is relevant. That is, one has to be careful not to describe the world that one is conceiving (the XYZ world, say) according to primary intensions, when secondary intensions would be more appropriate.[33]

It follows from all this that the oft-cited distinction between "logical" and "metaphysical" possibility stemming from the Kripkean cases—on which it is held to be logically possible but not metaphysically possible that water is XYZ—is not a distinction at the level of *worlds,* but at most a distinction at the level of *statements.* A statement is "logically possible" in this sense if it is true in some world when evaluated according to primary intensions; a statement is "metaphysically possible" if it is true in some world when evaluated according to secondary intensions. The relevant space of worlds is the same in both cases. [34]

Most importantly, none of the cases we have seen give reason to believe that any conceivable *worlds* are impossible. Any worries about the gap between conceivability and possibility apply at the level of statements, not worlds: either we use a statement to misdescribe a conceived world (as in the Kripkean case, and the second Goldbach case), or we claim that a statement is conceivable without conceiving of a world at all (as in the first Goldbach case). So there seems to be no reason to deny that conceivability of a world implies possibility. I will henceforth take this for granted as a claim about logical possibility; any variety of possibility for which conceivability does not imply possibility will then be a narrower class. Someone might hold that there is a narrower variety of "metaphysically possible worlds," but any reason for believing in such a class would have to be quite independent of the standard reasons I have considered here. In any case, it is logical possibility that is central to the issues about explanation. (A stronger "metaphysical" modality might at best be relevant to issues about ontology, materialism, and the like; I will discuss it when those issues become relevant in Chapter 4.)

An implication in the other direction, from logical possibility to conceivability, is trickier in that limits on our cognitive capacity imply that there are some possible situations that we cannot conceive, perhaps due to their great complexity. However, if we understand conceivability as conceivability-in-principle—perhaps conceivability by a superbeing—then it is plausible that logical possibility of a world implies conceivability of the world, and therefore that logical possibility of a statement implies conceivability of the statement (in the relevant sense). In any case, I will be more concerned with the other implication.

If a statement is logically possible or necessary according to its primary intension, the possibility or necessity is knowable *a priori,* at least in principle. Modality is not epistemically inaccessible: the possibility of a statement is a function of the intensions involved and the space of possible worlds, both of which are epistemically accessible in principle, and neither of which is dependent on *a posteriori* facts in this case. So matters of 1-possibility and 1-conceivability are in principle accessible from the armchair. By contrast, matters of 2-possibility and 2-conceivability will in many cases be accessible only *a posteriori,* as facts about the external world may play a role in determining the secondary intensions.

The class of 1-necessary truths corresponds directly to the class of *a priori* truths. If a statement is true *a priori,* then it is true no matter how the actual world turns out; that is, it is true in all worlds considered as actual, so it is 1-necessary. And conversely, if a statement is 1-necessary, then it will be true no matter how the actual world turns out, so it will be true *a priori.* In most such cases, the statement's truth will be knowable by us *a priori;* the exceptions may be certain mathematical statements whose truth we cannot determine, and certain statements that are so complex that we cannot comprehend them. Even in these cases, it seems reasonable to say that they are knowable *a priori* at least *in principle,* although they are beyond our limited cognitive capacity. (I will return to this matter when it becomes relevant later.)

Logical Necessity and Logical Supervenience

We obtain two slightly different notions of logical supervenience depending on whether we use the primary or secondary brands of logical necessity. If "gloop" has both a primary and a secondary intension associated with it, then gloopiness may supervene logically on physical properties according to either the primary or the secondary intension of "gloop". Supervenience according to secondary intension—that is, supervenience with *a posteriori* necessity as the relevant modality—corresponds to what some call "metaphysical supervenience," but we have now seen how this can be regarded as a variety of logical supervenience.

(There is really only one kind of logical supervenience of *properties,* just as there is only one kind of logical necessity of *propositions.* But we have seen that terms or concepts effectively determine two properties, one via a primary intension ["watery stuff"] and the other via a secondary intension ["H_2O"]. So for a given concept ["water"], there are two ways in which properties associated with that concept might supervene. I will sometimes talk loosely of the primary and secondary intensions associated with a property, and of the two ways in which a property might supervene.)

I will discuss both the primary and secondary versions of logical supervenience in specific cases, but the former will be more central. Especially when considering questions about explanation, primary intensions are more important than secondary intensions. As noted before, we have only the primary intension to work with at the start of inquiry, and it is this intension that determines whether or not an explanation is satisfactory. To explain water, for example, we have to explain things like its clarity, liquidity, and so on. The secondary intension ("H_2O") does not emerge until after an explanation is complete, and therefore does not itself determine a criterion for explanatory success. It is logical supervenience according to a primary intension that determines whether reductive explanation is possible. Where I do not specify otherwise, it is logical supervenience according to primary intension that I will generally be discussing.

If we choose one sort of intension—say, the primary intension—and stick with it, then we can see that various ways of formulating logical supervenience are equivalent. According to the definition given at the start of this chapter, B-properties are logically supervenient on A-properties if for any logically possible situation *Y* that is A-indiscernible from an actual situation *X,* then all the B-facts true of *X* are true of *Y*. Or more simply, B-properties are logically supervenient on A-properties if for any actual situation *X*, the A-facts about *Xentail* the B-facts about *X* (where "P entails Q" is understood as "It is logically impossible that P and not Q").

Sticking to global supervenience, this means that B-properties supervene logically on A-facts if the B-facts about the actual world are entailed by the A-facts. Similarly, B-properties supervene logically on A-properties if there is no conceivable world with the same A-properties as our world but different B-properties. We can also say that logical supervenience holds if, given the totality of A-facts A* and any B-fact *B* about our world *W,* "A*(*W*) ® *B*(*W*)" is true in virtue of the meanings of the A-terms and the B-terms (where meanings are understood as intensions).

Finally, if B-properties are logically supervenient on A-properties according to primary intensions, then the implication from A-facts to B-facts will be *a priori.* So in principle, someone who knows all the A-facts about an actual situation will be able to ascertain the B-facts about the situation from those facts alone, given that they possess the B-concepts in question. This sort of inference may be difficult or impossible in practice, due to the complexity of the situations involved, but it is at least possible in principle. For logical supervenience according to *secondary* intensions, B-facts about a situation can also be ascertained from the A-facts in principle, but only *a posteriori*. The A-facts will have to be supplemented with contingent facts about the actual world, as those facts will play a role in determining the B-intensions involved.

There are therefore at least three avenues to establishing claims of logical supervenience: these involve conceivability, epistemology, and analysis. To establish that B-properties logically supervene on A-properties, we can (1) argue that instantiation of A-properties without instantiation of the B-properties is inconceivable; (2) argue that someone in possession of the A-facts could come to know the B-facts (at least in cases of supervenience via primary intension); or (3) analyze the intensions of the B-properties in sufficient detail that it becomes clear that B-statements follow from A-statements in virtue of these intensions alone. The same goes for establishing the failure of logical supervenience. I will use all three methods in arguing for central claims involving logical supervenience.

Not everybody may be convinced that the various formulations of logical supervenience are equivalent, so when arguing for important conclusions involving logical supervenience I will run versions of the arguments using each of the different formulations. In this way it will be seen that the arguments are robust, with nothing depending on a subtle equivocation between different notions of supervenience.

5—
Almost Everything Is Logically Supervenient on the Physical*

In the following chapter I will argue that conscious experience does not supervene logically on the physical, and therefore cannot be reductively explained. A frequent response is that conscious experience is not alone here, and that all sorts of properties fail to supervene logically on the physical. It is suggested that such diverse properties as tablehood, life, and economic prosperity have no *logical* relationship to facts about atoms, electromagnetic fields, and so on. Surely those high-level facts could not be logically entailed by the microphysical facts?

On a careful analysis, I think that it is not hard to see that this is wrong, and that the high-level facts in question are (globally) logically supervenient on the physical insofar as they are facts at all. [35] Conscious experience is almost unique in its failure to supervene logically. The relationship between consciousness and the physical facts is different in kind from the standard relationship between high-level and low-level facts.

There are various ways to make it clear that most properties supervene logically on physical properties. Here I will only be concerned with properties that characterize *natural phenomena—that* is, contingent aspects of the world that need explaining. The property of being an angel might not supervene logically on the physical, but angels are not something that we have reason to believe in, so this failure need not concern us. I will also not concern myself with facts about abstract entities such as mathematical entities and propositions, which need to be treated separately.[36]

It should be noted that in claiming that most high-level properties supervene on the physical, I am not suggesting that high-level facts and laws are entailed by microphysical *laws*, or even by microphysical laws in conjunction with microphysical boundary conditions. That would be a strong claim, and although it might have some plausibility if qualified appropriately, the evidence is not yet in. I am making the much weaker claim that high-level facts are entailed by all the microphysical *facts* (perhaps along with microphysical laws). This enormously comprehensive set includes the facts about the distribution of every last particle and field in every last corner of space-time: from the atoms in Napoleon's hat to the electromagnetic fields in the outer ring of Saturn. Fixing this set of facts leaves very little room for anything else to vary, as we shall see.

Before moving to the arguments I should note some harmless reasons why logical supervenience on the physical sometimes fails. First, some high-level properties fail to supervene logically because of a dependence on conscious experience. Perhaps conscious experience is partly constitutive of a property like love, for example. The primary (although not the secondary) intensions associated with some external properties such as color and heat may also be dependent on phenomenal qualities, as we will see. If so, then love and perhaps heat do not supervene logically on the physical. These should not be seen as providing counterexamples to my thesis, as they introduce no new failure of logical supervenience. Perhaps the best way to phrase the claim is to say that all facts supervene logically on the combination of physical facts and phenomenal facts, or that all facts supervene logically on the physical facts *modulo conscious experience*. Similarly, a dependence on conscious experience may hinder the reductive explainability of some high-level phenomena, but we can still say that they are reductively explainable modulo conscious experience.

Second, an *indexical* element enters into the application of some primary intensions, although not secondary intensions, as we saw earlier. The primary intension of "water," for example, is something like "the clear, drinkable liquid in our environment," so that if there is watery H_2O and watery XYZ in the actual universe, which of them qualifies as "water" depends on which is in the environment of the agent using the term. In principle we therefore need to add a *center* representing the location of an agent to the supervenience base in some cases. This yields logical supervenience and reductive explanation modulo conscious experience and indexicality.

Finally, cases where the high-level facts are indeterminate do not count against logical supervenience. The claim is only that insofar as the high-level facts are determinate, they are determined by the physical facts. If the world itself does not suffice to fix the high-level facts, we cannot expect the physical facts to. Some might suggest that logical supervenience would fail if there were two equally good high-level theories of the world that differed in their description of the high-level facts. One theory might hold that a virus is alive, for instance, whereas another might hold that it is not, so the facts about life are not determined by the physical facts. This is not a counterexample, however, but a case in which the facts about life are indeterminate. Given indeterminacy, we are free to legislate the terms one way or the other where it is convenient. If the facts *are* determinate—for example, if it is true that viruses are alive—then one of the descriptions is simply wrong. Either way,

insofar as the facts about the situation are determinate at all, they are entailed by the physical facts.

I will argue for the ubiquity of logical supervenience using arguments that appeal to conceivability, to epistemological considerations, and to analysis of the concepts involved.

Conceivability. The logical supervenience of most high-level facts is most easily seen by using conceivability as a test for logical possibility. What kind of world could be identical to ours in every last microphysical fact but be biologically distinct? Say a wombat has had two children in our world. The physical facts about our world will include facts about the distribution of every particle in the spatiotemporal hunk corresponding to the wombat, and its children, and their environments, and their evolutionary histories. If a world shared those physical facts with ours, but was not a world in which the wombat had two children, what could that difference consist in? Such a world seems quite inconceivable. Once a possible world is fixed to have all those physical facts the same, then the facts about wombathood and parenthood are automatically fixed. These biological facts are not the sort of thing that can float free of their physical underpinnings even as a conceptual possibility.

The same goes for architectural facts, astronomical facts, behavioral facts, chemical facts, economic facts, meteorological facts, sociological facts, and so on. A world physically identical to ours, but in which these sort of facts differ, is inconceivable. In conceiving of a microphysically identical world, we conceive of a world in which the location of every last particle throughout space and time is the same. It follows that the world will have the same macroscopic structure as ours, and the same macroscopic dynamics. Once all this is fixed there is simply no room for the facts in question to vary (apart, perhaps, from any variation due to variations in conscious experience).

Furthermore, this inconceivability does not seem to be due to any contingent limits in our cognitive capacity. Such a world is inconceivable *in principle*. Even a superbeing, or God, could not imagine such a world. There is simply not anything for them to imagine. Once they imagine a world with all the physical facts, they have automatically imagined a world in which all the high-level facts hold. A physically identical world in which the high-level facts are false is therefore logically impossible, and the high-level properties in question are logically supervenient on the physical.

Epistemology. Moving beyond conceivability intuitions, we can note that if there *were* a possible world physically identical to ours but biologically distinct, then this would raise radical epistemological problems. How would we know that we were not in that world rather than this one? How would we know that the biological facts in our world are as they are? To see this, note that if I were in the alternative world, it would certainly *look* the same as this one. It instantiates the same distribution of particles found in the plants and animals in this world; indistinguishable patterns of photons are reflected from those entities; no difference would be revealed under even the closest examination. It follows that all the external evidence we possess fails to distinguish the possibilities. Insofar as the biological facts about our world are not logically supervenient, there is no way we can know those facts on the basis of external evidence.

In actuality, however, there is no deep epistemological problem about biology. We come to know biological facts about our world on the basis of external evidence all the time, and there is no special skeptical problem that arises. It follows that the biological facts are logically supervenient on the physical. The same goes for facts about architecture, economics, and meteorology. There is no special skeptical problem about knowing these facts on the basis of external evidence, so they must be logically supervenient on the physical.

We can back up this point by noting that in areas where there *are* epistemological problems, there is an accompanying failure of logical supervenience, and that conversely, in areas where logical supervenience fails, there are accompanying epistemological problems.

Most obviously, there is an epistemological problem about consciousness—the problem of other minds. This problem arises because it seems logically compatible with all the external evidence that beings around us are conscious, and it is logically compatible that they are not. We have no way to peek inside a dog's brain, for instance, and observe the presence or absence of conscious experience. The status of this problem is controversial, but the mere *prima facie* existence of the problem is sufficient to defeat an epistemological argument, parallel to those above, for the logical supervenience of consciousness. By contrast, there is not even a *prima facie* problem of other biologies, or other economies. Those facts are straightforwardly publically accessible, precisely because they are fixed by the physical facts.

(Question: Why doesn't a similar argument force us to the conclusion that if conscious experience fails to supervene logically, then we can't know about even our *own* consciousness? Answer: Because conscious experience is at the very center of our epistemic universe. The skeptical problems about nonsupervenient biological facts arise because we only have access to biological facts by external, physically mediated evidence; external nonsupervenient facts would be out of our direct epistemic reach. There is no such problem with our own consciousness.)

Another famous epistemological problem concerns facts about causation. As Hume argued, external evidence only gives us access to regularities of

succession between events; it does not give us access to any further fact of causation. So if causation is construed as something over and above the presence of a regularity (as I will assume it must be), it is not clear that we can know that it exists. Once again, this skeptical problem goes hand in hand with a failure of logical supervenience. In this case, facts about causation fail to supervene logically on matters of particular physical fact. Given all the facts about distribution of physical entities in space-time, it is logically possible that all the regularities therein arose as a giant cosmic coincidence without any real causation. At a smaller scale, given the particular facts about any apparent instance of causation, it is logically possible that it is a mere succession. We infer the existence of causation by a kind of inference to the best explanation—to believe otherwise would be to believe in vast, inexplicable coincidences—but belief in causation is not forced on us in the direct way that belief in biology is forced on us.

I have sidestepped problems about the supervenience of causation by stipulating that the supervenience base for our purposes includes not just particular physical facts but all the physical laws. It is reasonable to suppose that the addition of laws fixes the facts about causation. But of course there is a skeptical problem about laws paralleling the problem about causation: witness Hume's problem of induction, and the logical possibility that any apparent law might be an accidental regularity.

As far as I can tell, these two problems exhaust the epistemological problems that arise from failure of logical supervenience on the physical. There are some other epistemological problems that in a sense precede these, because they concern the existence of the physical facts themselves. First, there is Descartes's problem about the existence of the external world. It is compatible with our experiential evidence that the world we think we are seeing does not exist; perhaps we are hallucinating, or we are brains in vats. This problem can be seen to arise precisely because the facts about the external world do not supervene logically on the facts about our experience. (Idealists, positivists, and others have argued controversially that they do. Note that if these views are accepted the skeptical problem falls away.) There is also an epistemological problem about the theoretical entities postulated by science—electrons, quarks, and such. Their absence would be logically compatible with the directly observable facts about objects in our environment, and some have therefore raised skeptical doubts about them. This problem can be analyzed as arising from the failure of theoretical facts to supervene logically on observational facts. In both these cases, skeptical doubts are perhaps best quelled by a form of inference to the best explanation, just as in the case of causation, but the in-principle possibility that we are wrong remains.

In any case, I am bypassing this sort of skeptical problem by giving myself the physical world for free, and fixing all physical facts about the world in

the supervenience base (thereby assuming that the external world exists, and that there are electrons, and so on). Given that those facts are known, there is no room for skeptical doubts about most high-level facts, precisely because they are logically supervenient. To put the matter the other way around: All our sources of external evidence supervene logically on the microphysical facts, so that insofar as some phenomenon does not supervene on those facts, external evidence can give us no reason to believe in it. One might wonder whether some further phenomena might be posited via inference to the best explanation, as above, to explain the microphysical facts. Indeed, this process takes us from particular facts to simple underlying laws (and hence yields causation), but then the process seems to stop. It is in the nature of fundamental laws that they are the end of the explanatory chain (except, perhaps, for theological speculation). This leaves phenomena that we have *internal* evidence for —namely conscious experience—and that is all. Modulo conscious experience, all phenomena are logically supervenient on the physical.

We can also make an epistemological case for logical supervenience more directly, by arguing that someone in possession of all the physical facts could in principle come to know all the high-level facts, given that they possess the high-level concepts involved. True, one could never *in practice* ascertain the high-level facts from the set of microphysical facts. The vastness of the latter set is enough to rule that out. (Even less am I suggesting that one could perform a formal derivation; formal systems are irrelevant for reasons canvased earlier.) But as an in-principle point, there are various ways to see that someone (a superbeing?) armed with only the microphysical facts and the concepts involved could infer the high-level facts.

The simplest way is to note that in principle one could build a big mental simulation of the world and watch it in one's mind's eye, so to speak. Say that a man is carrying an umbrella. From the associated microphysical facts, one could straightforwardly infer facts about the distribution and chemical composition of mass in the man's vicinity, giving a high-level structural characterization of the area. One could determine the existence of a male fleshy biped straightforwardly enough. For instance, from the structural information one could note that there was an organism atop two longish legs that were responsible for its locomotion, that the creature has male anatomy, and so on. It would be clear that he was carrying some device that was preventing drops of water, otherwise prevalent in the neighborhood, from hitting him. Doubts that this device is really an umbrella could be assuaged by noting from its physical structure that it can fold and unfold; from its history that it was hanging on a stand that morning, and was originally made in a factory with others of a similar kind, and so on. Doubts that the fleshy biped is really a human could be assuaged by noting the composition of his DNA, his evolutionary history, his relation to other beings, and so on. We

need only assume that the being possesses enough of the concept involved to be able to apply it correctly to instances (that is, the being possesses the intension). If so, then the microphysical facts will give it all the evidence it needs to apply the concepts, and to determine that there really is a person carrying an umbrella here.

The same goes for almost any sort of high-level phenomena: tables, life, economic prosperity. By knowing all the low-level facts, a being in principle can infer all the facts necessary to determine whether or not this is an instance of the property involved. Effectively, what is happening is that a possible world compatible with the microphysical facts is constructed, and the high-level facts are simply read off that world using the appropriate intension (as the relevant facts are invariant across physically identical possible worlds). Hence the high-level facts are logically supervenient on the physical.

Analyzability. So far, I have argued that microphysical facts fix high-level facts without saying much explicitly about the high-level concepts involved. In any specific case, however, this entailment relationship relies on a concept's intension. If microphysical facts entail a high-level fact, this is because the microphysical facts suffice to fix those features of the world in virtue of which the high-level intension applies. That is, we should be able to *analyze* what it takes for an entity to satisfy the intension of a high-level concept, at least to a sufficient extent that we can see why those conditions for satisfaction could be satisfied by fixing the physical facts. It is therefore useful to look more closely at the intensions of high-level concepts, and to examine the features of the world in virtue of which they apply.

There are some obstacles to elucidating these intensions and to summarizing them in words. As we saw earlier, application conditions of a concept are often indeterminate in places. Is a cup-shaped object made of tissues a cup? Is a computer virus alive? Is a booklike entity that coagulates randomly into existence a book? Our ordinary concepts do not give straightforward answers to these questions. In a sense, it is a matter for stipulation. Hence there will not be determinate application conditions for use in the entailment process. But as we saw earlier, this indeterminacy precisely mirrors an indeterminacy about the facts themselves. Insofar as the intension of "cup" is a matter for stipulation, the facts about cups are also a matter for stipulation. What counts for our purposes is that the intension together with the microphysical facts determines the high-level facts insofar as they are really factual. Vagueness and indeterminacy can make discussion awkward, but they affect nothing important to the issues.

A related problem is that any short analysis of a concept will invariably fail to do justice to it. As we have seen, concepts do not usually have crisp definitions. At a first approximation, we can say something is a table if it

Page 78

has a flat horizontal surface with legs as support; but this lets in too many things (Frankenstein's monster on stilts?) and omits others (a table with no legs, sticking out from a wall?). One can refine the definition, adding further conditions and clauses, but we quickly hit the problems with indeterminacy, and in any case the product will never be perfect. But there is no need to go into all the details required to handle every special case: after a point the details are just more of the same. As long as we know what *sort* of properties the intension applies in virtue of, we will have enough to make the point.

As we saw before, we do not need a definition of B-properties in terms of A-properties in order for A-facts to entail B-facts. Meanings are fundamentally represented by intensions, not definitions. The role of analysis here is simply to characterize the intensions in sufficient detail that the existence of an entailment becomes clear. For this purpose, a rough-and-ready analysis will suffice. Intensions generally apply to individuals in a possible world in virtue of some of their properties and not others; the point of such an analysis is to see what sort of properties the intension applies in virtue of, and to make the case that properties of this sort are compatible with entailment by physical properties.

A third problem stems from the division between the *a priori* and *a posteriori* application conditions of many concepts. As long as we keep primary and secondary intensions separate, however, this is not much of a problem. The secondary intension associated with "water" is something like "H_2O," which is obviously logically supervenient on the physical. But the primary intension, something like "the clear, drinkable liquid in our environment" is equally logically supervenient, as the clarity, drinkability, and liquidity of water is entailed by the physical facts. [37] We can run things either way. As we have seen, it is the primary intension that enters into reductive explanation, so it is this that we are most concerned with. In general, if a primary intension *I* is logically supervenient on the physical, then so is a rigidified secondary intension *dthat*(*I*) as it will generally consist in a projection of some intrinsic physical structure across worlds.

Considerations about *a posteriori* necessity have led some to suppose that there can be no logical entailment from low-level facts to high-level facts. Typically one hears something like "Water is necessarily H_2O, but that is not a truth of meaning, so there is no conceptual relation." But this is a vast oversimplification. For a start, the secondary intension "H_2O" can be seen as part of the meaning of "water" in some sense, and it certainly supervenes logically. But more importantly, the primary intension ("the clear, drinkable liquid . . .") which fixes reference also supervenes, perhaps modulo experience and indexicality. It is precisely in virtue of its satisfying this intension that we deemed that H_2O was water in the first place. Given the primary intension *I,* the high-level facts are derivable unproblematically

from the microphysical facts (modulo the contribution of experience and indexicality). The Kripkean observation that the concept is better represented as *dthat*(*I*) affects this derivability not at all. The semantic phenomenon of rigidification does not alone make an ontological difference.

With these obstacles out of the way, we can look at the intensions associated with various high-level concepts. In most cases these are characterizable in functional or structural terms, or as a combination of the two. For example, the sorts of things relevant to something's being a table include (1) that it have a flat top and be supported by legs, and (2) that people use it to support various objects. The first of these is a structural condition: that is, a condition on the intrinsic physical structure of the object. The second is a functional condition: that is, it concerns the external causal role of an entity, characterizing the way it interacts with other entities. Structural properties are clearly entailed by microphysical facts. So are functional properties in general, although this is slightly less straightforward. Such properties depend on a much wider supervenience base of microphysical facts, so that facts about an object's environment will often be relevant; and insofar as such properties are characterized dispositionally (something is soluble if it *would* dissolve *if* immersed in water), one needs to appeal to counterfactuals. But the truth-values of those counterfactuals are fixed by the inclusion of physical laws in the antecedent of our supervenience conditionals, so this is not a problem.

To take another example, the conditions on life roughly come down to some combination of the ability to reproduce, to adapt, and to metabolize, among other things (as usual, we need not legislate the weights, or all other relevant factors). These properties are all characterizable functionally, in terms of an entity's relation to other entities, its ability to convert external resources to energy, and its ability to react appropriately to its environment. These functional properties are all derivable, in principle, from the physical facts. As usual, even if there is no perfect definition of life in functional terms, this sort of characterization shows us that life is a functional property, whose instantiation can therefore be entailed by physical facts.

A complication is raised by the fact that functional properties are often characterized in terms of a causal role relative to other high-level entities. It follows that logical supervenience of the properties depends on the logical supervenience of the other high-level notions involved, where these notions may themselves be characterized functionally. This is ultimately not a problem, as long as causal roles are eventually cashed out by nonfunctional properties: typically either by structural or phenomenal properties. There may be some circularity in the interdefinability of various functional properties—perhaps it is partly constitutive of a stapler that it deliver staples, and partly constitutive of staples that they are delivered by staplers. This circularity can be handled by cashing out the causal roles of all the properties simultaneously, [38] as long as the analyses have a noncircular part that is

ultimately grounded in structural or phenomenal properties. (The appeal to phenomenal properties may seem to count against logical supervenience on the physical, but see below. In any case, it is compatible with logical supervenience modulo conscious experience.)

Many properties are characterized relationally, in terms of relations to an entity's environment. Usually such relations are causal, so that the properties in question are functional, but this is not always so: witness the property of being on the same continent as a duck. Similarly, some properties are dependent on history (although these can usually be construed causally); to be a kangaroo, a creature must have appropriate ancestors. In any case these properties pose no problems for logical supervenience, as the relevant historical and environmental facts will themselves be fixed by the global physical facts.

Even a complex social fact such as "There was economic prosperity in the 1950s" [39] is characterizable in mostly functional terms, and so can be seen to be entailed by the physical facts. A full analysis would be very complicated and would be made difficult by the vagueness of the notion of prosperity, but to get an idea how it might go, one can ask why we say that there *was* economic prosperity in the 1950s? At a first approximation, because there was high employment, people were able to purchase unusually large amounts of goods, there was low inflation, much development in housing, and so on. We can in turn give rough-and-ready analyses of the notion of housing (the kind of place people sleep and eat in), of employment (organized labor for reward), and of monetary notions (presumably money will be roughly analyzable in terms of the systematic ability to exchange for other items, and its value will be analyzable in terms of how much one gets in exchange). All these analyses are ridiculously oversimplified, but the point is clear enough. These are generally functional properties that can be entailed by physical facts.

Many have been skeptical of the possibility of conceptual analysis. Often this has been for reasons that do not make any difference to the arguments I am making—because of indeterminacy in our concepts, for example, or because they lack crisp definitions. Sometimes this skepticism may have arisen for deeper reasons. Nevertheless, if what I have said earlier in this chapter is correct, and if the physical facts about a possible world fix the high-level facts, we should *expect* to be able to analyze the intension of the high-level concept in question, at least to a good approximation, in order to see how its application can be determined by physical facts. This is what I have tried to do in the examples given here. Other examples can be treated similarly.[40]

I am not advocating a program of performing such analyses in general. Concepts are too complex and unruly for this to do much good, and any explicit analysis is usually a pale shadow of the real thing. What counts is

the general point that most high-level concepts are not primitive, unanalyzable notions. They are generally analyzable to the extent that their intensions can be seen to specify functional or structural properties. It is in virtue of this analyzability that high-level facts are in principle derivable from microphysical facts and reductively explainable in terms of physical facts.

Some Problem Cases

There are some types of properties that might be thought to provide particular difficulties for logical supervenience, and therefore for reductive explanation. I will examine a number of such candidates, paying particular attention to the question of whether the associated phenomena pose problems for reductive explanation analogous to the problems posed by consciousness. It seems to me that with a couple of possible exceptions, no significant new problems arise here.

Consciousness-dependent properties. As discussed already, some concepts' primary intensions involve a relation to conscious experience. An obvious example is redness, taken as a property of external objects. On at least some accounts, the primary intension associated with redness requires that for something to be red, it must be the kind of thing that tends to cause red experiences under appropriate conditions. [41] So in its primary intension, redness is not logically supervenient on the physical, although it supervenes modulo conscious experience. On the other hand, its secondary intension almost certainly supervenes. If it turns out that in the actual world, the sort of thing that tends to cause red experience is a certain surface reflectance, then objects with that reflectance are red even in worlds in which there is no conscious being to see them. Redness is identified *a posteriori* with that reflectance, which is logically supervenient on the physical alone.

We saw earlier that failure of a primary intension to supervene logically is associated with a failure of reductive explanation. So, does reductive explanation fail for redness? The answer is yes, in a weak sense. If redness is construed as the tendency to cause red experiences, then insofar as experience is not reductively explainable, neither is redness. But one can come close. One can note *that* a certain physical quality causes red experiences; and one can even explain the causal relation between the quality and red-*judgments*. It is just the final step to experience that goes unexplained. In practice, our strictures on explanation are weak enough that this sort of thing counts. To explain a phenomenon to which reference is fixed by some experience, we do not require an explanation of experience. Otherwise we would wait a long time.

The same goes for phenomena such as heat, light, and sound. Although their secondary intensions determine structural properties (molecular motion, the presence of photons, waves in air), their primary intensions involve a relation to conscious experience: heat is the thing that causes heat sensations, light causes visual experiences, and so on. But as Nagel (1974) and Searle (1992) have noted, we do not require an explanation of heat sensations when explaining heat. Explanation modulo experience is good enough.

Other properties depend even more directly on conscious experience, in that experience not only plays a role in reference fixation but is partly constitutive of the *a posteriori* notion as well. The property of standing next to a conscious person is an obvious example. On some accounts, mental properties such as love and belief, although not themselves phenomenal properties, have a conceptual dependence on the existence of conscious experience. If so, then in a world without consciousness, such properties would not be exemplified. Such properties therefore are not logically supervenient even *a posteriori*, and reductive explanation fails even more strongly than in the above cases. But they are logically supervenient and reductively explainable modulo conscious experience, so no *further* failure of reductive explanation arises here.

Intentionality. It is worth separately considering the status of intentionality, as this is sometimes thought to pose problems analogous to those posed by consciousness. It is plausible, however, that any failure of intentional properties to supervene logically is derivative on the nonsupervenience of consciousness. As I noted in Chapter 1, there seems to be no conceivable world that is physically and phenomenally identical to ours, but in which intentional contents differ. [42] If phenomenology is partly constitutive of intentional content, as some philosophers suggest, then intentional properties may fail to supervene logically on the physical, but they will supervene modulo conscious experience. The claim that consciousness is partly constitutive of content is controversial, but in any case there is little reason to believe that intentionality fails to supervene in a separate, nonderivative way.

Leaving any phenomenological aspects aside, intentional properties are best seen as a kind of third-person construct in the explanation of human behavior, and should therefore be analyzable in terms of causal connections to behavior and the environment. If so, then intentional properties are straightforwardly logically supervenient on the physical. Lewis (1974) makes a thorough attempt at explicating the entailment from physical facts to intentional facts by giving an appropriate functional analysis. More recent accounts of intentionality, such as those by Dennett (1987), Dretske (1981), and Fodor (1987) can be seen as contributing to the same project. None of these analyses are entirely compelling, but it may be that a more sophisticated descendant might do the job. There is no argument analogous to the arguments against the supervenience of consciousness showing that intentionality

cannot supervene logically on physical and phenomenal properties. [43] Indeed, conceivability arguments indicate that intentional properties must be logically supervenient on these if such properties are instantiated at all, and epistemological arguments lead us to a similar conclusion. So there is no separate *ontological* problem of intentionality.

Moral and aesthetic properties. It is often held that there is no *conceptual* connection from physical properties to moral and aesthetic properties. According to Moore (1922), nothing about the *meaning* of notions such as "goodness" allows that facts about goodness should be entailed by physical facts. In fact, Moore claimed that there is no conceptual connection from *natural* facts to moral facts, where the natural may include the mental as well as the physical (so supervenience modulo conscious experience does not help here). Does this mean that moral properties are as problematic as conscious experience?

There are two disanalogies, however. First, there does not seem to be a conceivable world that is naturally identical to ours but morally distinct, so it is unlikely that moral facts are further facts in any strong sense. Second, moral facts are not phenomena that force themselves on us. When it comes to the crunch, we can deny that moral facts exist at all. Indeed, this reflects the strategy taken by moral antirealists such as Blackburn (1971) and Hare (1984). These antirealists argue that because moral facts are not entailed by natural facts and are not plausibly "queer" further facts, they have no objective existence and morality should be relativized into a construct or projection of our cognitive apparatus. The same strategy cannot be taken for phenomenal properties, whose existence is forced upon us.

For moral properties, there are at least two reasonable alternatives available. The first is antirealism of some sort, perhaps relativizing "objective moral facts" into "subjective moral facts,"[44] or embracing a view on which moral discourse does not state facts at all. The second is to claim that there is an *a priori* connection from natural facts to moral facts, one that (contra Moore) can be seen to hold in virtue of an analysis and explication of moral concepts. If a concept such as "good" determines a stable nonindexical primary intension, then the second position follows: we will have an *a priori* function from naturally specified worlds to moral facts. If it only determines an indexical primary intension, or if different subjects can equally reasonably associate different primary intensions with the concept, or if it determines no primary intension at all, then a version of the first position will follow.

Some other positions are sometimes taken, but none seem tenable. Moore held that there is a nonconceptual *a priori* connection between natural and moral facts that we obtain through a mysterious faculty of "moral intuition," but this view is widely rejected (it is hard to see what could ground such intuitions' truth or falsity). A position on which moral properties supervene

by a fundamental nomic link seems out of the question, as there is no conceivable world in which the natural facts are the same as ours but in which the moral facts are different. A popular position among contemporary moral realists (see, e.g., Boyd 1988; Brink 1989) is that moral facts supervene on natural facts with *a posteriori* necessity; that is, they supervene according to the secondary but not the primary intensions of moral concepts. This position is difficult to maintain, however, given that even *a posteriori* equivalences must be grounded in *a priori* reference fixation. Even though it is *a posteriori* that water is H_2O, the facts about water follow from the micro-physical facts *a priori.* Similarly, if moral concepts have a primary intension and if naturally identical centered worlds are morally identical, an *a priori* link from natural facts to moral facts would seem to follow. (Horgan and Timmons [1992a; 1992b] provide a critique along these lines.)

Aesthetic properties can be treated in a similar way. If anything, an anti-realist treatment is even more plausible here. In the final analysis, although there are interesting conceptual questions about how the moral and aesthetic domains should be treated, they do not pose metaphysical and explanatory problems comparable to those posed by conscious experience.

Names. On many accounts (e.g., Kaplan 1989), there is no analysis associated with a name such as "Rolf Harris," which simply picks out its referent directly. Does this mean that the property of being Rolf Harris fails to supervene logically on the physical? There is no problem about the supervenience of the secondary intension (e.g., Rolf might be the person conceived from a given sperm and egg in all possible worlds), but the absence of a primary intension might be thought to pose problems for reductive explanation. Still, it is plausible that even though there is no primary intension that is shared across the community, every individual use of the name has a primary intension attached. When I use the name "Rolf Harris," there is *some* systematic way in which its referent depends on the way the world turns out; for me, the primary intension might be something like "the man called 'Rolf Harris' who bangs around on paint cans, and who bears the appropriate causal relation to me."[45] Such an intension will supervene logically. Rather than justifying this in detail, however, it is easier to note that any failure of logical supervenience will not be accompanied by an explanatory mystery. The property of being Rolf Harris does not constitute a phenomenon in need of explanation, as opposed to explication. What needs explaining is the existence of a person named "Rolf Harris" who bangs around on paint cans, and so on. These properties certainly supervene, and are explainable in principle in the usual way.

Indexicals. Reference fixation of many concepts, from "water" to "my dog," involves an indexical element. The reference of these notions is fixed on the basis of both physical facts and an agent-relative "indexical fact" representing the location of an agent using the term in question. Such a fact is determinate for any given agent, so reference fixation is determinate. Supervenience and explanation succeed modulo that indexical fact.

Does indexicality pose a problem for reductive explanation? For arbitrary speakers, perhaps not, as the "fact" in question can be relativized away. But for myself, it is not so easy. The indexical fact expresses something very salient about the world as I find it: that David Chalmers is *me*. How could one explain this seemingly brute fact? Indeed, is there really a fact here to be explained, as opposed to a tautology? The issue is extraordinarily difficult to get a grip on, but it seems to me that even if the indexical is not an objective fact about the world, it is a fact about the world as I find it, and it is the world as I find it that needs explanation. The nature of the brute indexical is quite obscure, though, and it is most unclear how one might explain it. [46] (Of course, we can give a reductive explanation of why David Chalmers's utterance of "I am David Chalmers" is true. But this nonindexical fact seems quite different from the indexical fact that I am David Chalmers.)

It is tempting to look to consciousness. But while an explanation of consciousness might yield an explanation of "points of view" in general, it is hard to see how it could explain why a seemingly arbitrary one of those points of view is *mine*, unless solipsism is true. The indexical fact may have to be taken as primitive. If so, then we have a failure of reductive explanation distinct from and analogous to the failure with consciousness. Still, the failure is less worrying than that with consciousness, as the unexplained fact is so "thin" by comparison to the facts about consciousness in all its glory. Admitting this primitive indexical fact would require far less revision of our materialist worldview than would admitting irreducible facts about conscious experience.

Negative facts. As we saw earlier, certain facts involving negative existentials and universal quantifiers are not logically determined by the physical facts, or indeed by any set of localized facts. Consider the following facts about our world: there are no angels; Don Bradman is the greatest cricketer; everything alive is based on DNA. All these could be falsified, consistently with all the physical facts about our world, simply by the addition of some new nonphysical stuff: cricket-playing angels made of ectoplasm, for instance. Even addition of facts about conscious experience or indexicality cannot help here.[47]

Does this mean that these facts are not reductively explainable? It seems so, insofar as there is no physical explanation of why there is no extra nonphysical stuff in our world. That is indeed a further fact. The best way to deal with this situation is to introduce a second-order fact that says of

the set of basic particular facts, be they microphysical, phenomenal, indexical, or whatever: *That's all.* This fact says that all the particular facts about the world are included in or entailed by the given set of facts. From this second-order fact, in conjunction with all the basic particular facts, all the negative facts will follow.

This does not constitute a very serious failure of reductive explanation. Presumably there will be such a "That's all" fact true of any world, and such a fact will never be entailed by the particular facts. It simply expresses the bounded nature of our world, or of any world. It is a cheap way to bring all the negative existential and universally quantified facts within our grasp.

Physical laws and causation. On the most plausible accounts of physical laws, these are not logically supervenient on the physical facts, taken as a collection of particular facts about a world's spatiotemporal history. One can see this by noting the logical possibility of a world physically indiscernible from ours over its entire spatiotemporal history, but with different laws. For example, it might be a law of that world that whenever two hundred tons of pure gold are assembled in a vacuum, it will transmute into lead. Otherwise its laws are identical, with minor modifications where necessary. As it happens, in the spatiotemporal history of our world, two hundred tons of gold are never assembled in a vacuum. It follows that our world and the other world have identical histories, but their laws differ nevertheless.

Arguments like this suggest that the laws of nature do not supervene logically on the collection of particular physical facts. [48] By similar arguments one can see that a causal connection between two events is something over and above a regularity between the events. Holders of various Humean views dispute these conclusions, but it seems to me that they have the worse of the arguments here.[49] There is something irreducible in the existence of laws and causation.

I have bypassed these problems elsewhere by including physical laws in the supervenience base, but this steps over the metaphysical puzzle rather than answering it. It is true that laws and causation lead to less significant failure of reductive explanation than consciousness. The laws and causal relations are themselves posited to explain existing physical phenomena, namely the manifold regularities present in nature, whereas consciousness is a brute explanandum. Nevertheless the very existence of such irreducible further facts raises deep questions about their metaphysical nature. Apart from conscious experience and perhaps indexicality, these constitute the only such further facts in which we have any reason to believe. It is not unnatural to speculate that these two nonsupervenient kinds, consciousness and causation, may have a close metaphysical relation.

Recap

The position we are left with is that almost all facts supervene logically on the physical facts (including physical laws), with possible exceptions for conscious experience, indexicality, and negative existential facts. To put the matter differently, we can say that the facts about the world are exhausted by (1) particular physical facts, (2) facts about conscious experience, (3) laws of nature, (4) a second-order "That's all" fact, and perhaps (5) an indexical fact about my location. (The last two are minor compared to the others, and the status of the last is dubious, but I include them for completeness.) Modulo conscious experience and indexicality, it seems that all positive facts are logically supervenient on the physical. To establish this conclusively would require a more detailed examination of all kinds of phenomena, but what we have seen suggests that the conclusion is reasonable. We can sum up the ontological and epistemological situations with a couple of fables. Perhaps there is a grain of truth in the shape of these stories, if not in the details.

Creation myth. Creating the world, all God had to do was fix the facts just mentioned. For maximum economy of effort, he first fixed the laws of nature—the laws of physics, and any laws relating physics to conscious experience. Next, he fixed the boundary conditions: perhaps a time-slice of physical facts, and maybe the values in a random-number generator. These combined with the laws to fix the remaining physical and phenomenal facts. Last, he decreed, "That's all."

Epistemological myth. At first, I have only facts about my conscious experience. From here, I infer facts about middle-sized objects in the world, and eventually microphysical facts. From regularities in these facts, I infer physical laws, and therefore further physical facts. From regularities between my conscious experience and physical facts, I infer psychophysical laws, and therefore facts about conscious experience in others. I seem to have taken the abductive process as far as it can go, so I hypothesize: that's all. The world is much larger than it once seemed, so I single out the original conscious experiences as *mine.*

Note the very different order involved from the two perspectives. One could almost say that epistemology recapitulates ontology backward. Note also that it seems beyond God's powers to fix my indexical fact. Perhaps this is another reason to be skeptical about it.

The logical supervenience of most high-level phenomena is a conclusion that has not been as widely accepted as it might have been, even among those who discuss supervenience. Although the matter is often not discussed, many have been wary about invoking the conceptual modality as relevant to super-

venience relations. As far as I can tell there have been a number of separate reasons for this hesitation, none of which are ultimately compelling.

First, the problem with logically possible physically identical worlds with extra nonphysical stuff (angels, ectoplasm) has led some to suppose that supervenience relations cannot be logical (Haugeland 1982; Petrie 1987); but we have seen how to fix this problem. Second, many have supposed that considerations about *a posteriori* necessity demonstrate that supervenience relations cannot be underwritten by meanings (Brink 1989; Teller 1984); but we have seen that supervenience relations based on *a posteriori* necessity can be seen as a variety of logical supervenience. Third, there is a general skepticism about the notion of conceptual truth, deriving from Quine; but we have seen that this is a red herring here. Fourth, worries about "reducibility" have led some to suppose that supervenience is not generally a conceptual relation (Hellman and Thompson 1975); but it is unclear that there are any good arguments against reducibility that are also good arguments against logical supervenience. Fifth, the very phenomenon of conscious experience is sometimes invoked to demonstrate that supervenience relations cannot be logical in general (Seager 1988); but we have seen that conscious experience is almost unique in its failure to supervene logically. Finally, the claim that supervenience relations are not generally logical is often stated without argument, presumably as something that any reasonable person must believe (Bacon 1986; Heil 1992). [50]

It is plausible that every supervenience relation of a high-level property upon the physical is ultimately either (1) a logical supervenience relation of either the primary or secondary variety, or (2) a contingent natural supervenience relation. If neither of these holds for some apparent supervenience relation, then we have good reason to believe that there are no objective high-level facts of the kind in question (as, perhaps, for moral facts). I will argue further in Chapter 4 that there is no deep variety of supervenience intermediate between the logical and the natural.

This provides a unified explanatory picture, in principle. Almost every phenomenon is reductively explainable, in the weak sense outlined earlier, except for conscious experience and perhaps indexicality, along with the rock-bottom microphysical facts and laws, which have to be taken as fundamental.

It is worth taking a moment to answer a query posed by Blackburn (1985) and Horgan (1993): How do we explain the supervenience relations themselves? For a logical supervenience relation based on the primary intension of a concept, this is a simple matter of giving an appropriate analysis of the concept, perhaps in functional or structural terms, and noting that its reference is invariant across physically identical worlds. Here, the supervenience conditional is itself an *a priori* conceptual truth. For a logical supervenience relation based on a secondary intension, the supervenience can be explained by noting that the primary intension of the concept picks out some actual-world referent that is projected (by rigidification) invariantly across physically identical worlds. All we need here for an explanation is an *a priori* conceptual analysis combined with contingent facts about the actual world. [51] On the other hand, a mere natural supervenience relation will itself be a contingent law. At best it will be explainable in terms of more fundamental laws; at worst, the supervenience law will itself be fundamental. In either case, one explains certain regularities in the world by invoking fundamental laws, just as one does in physics, and as always, fundamental laws are where explanation must stop. Mere natural supervenience is ontologically expensive, as we have seen, so it is fortunate that logical supervenience is the rule and natural supervenience the exception.

PART II—
THE IRREDUCIBILITY OF CONSCIOUSNESS

3—
Can Consciousness Be Reductively Explained?

1—
Is Consciousness Logically Supervenient on the Physical?

Almost everything in the world can be explained in physical terms; it is natural to hope that consciousness might be explained this way, too. In this chapter, however, I will argue that consciousness escapes the net of reductive explanation. No explanation given wholly in physical terms can ever account for the emergence of conscious experience. This may seem to be a negative conclusion, but it leads to some strong positive consequences that I will bring out in later chapters.

To make the case against reductive explanation, we need to show that consciousness is not logically supervenient on the physical. In principle, we need to show that it does not supervene *globally*—that is, that all the microphysical facts in the world do not entail the facts about consciousness. In practice, it is easier to run the argument *locally*, arguing that in an individual, microphysical facts do not entail the facts about consciousness. When it comes to consciousness, local and global supervenience plausibly stand and fall together, so it does not matter much which way we run the argument: if consciousness supervenes at all, it almost certainly supervenes locally. If this is disputed, however, all the arguments can be run at the global level with straightforward alterations.

How can we argue that consciousness is not logically supervenient on the physical? There are various ways. We can think about what is conceivable, in order to argue directly for the logical possibility of a situation in which the physical facts are the same but the facts about experience are different. We can appeal to epistemology, arguing that the right sort of link between knowledge of physical facts and knowledge of consciousness is absent. And we can appeal directly to the concept of consciousness, arguing that there is no analysis of the concept that could ground an entailment from the physical to the phenomenal. In what follows I will give arguments using all three of these strategies. The first two are essentially arguments from conceivability, the second two are arguments from epistemology, and the fifth is an argument from analysis. There is some element of redundancy among the five arguments, but together they make a strong case.

One can also do things more directly, making the case against reductive explanation without explicitly appealing to logical supervenience. I have taken that route elsewhere, but here I will give the more detailed analysis to allow a fuller case. All the same, the case against reductive explanation and the critique of existing reductive accounts (in section 2 onward) should make sense even without this analysis. Some readers might like to proceed there directly, at least on a first reading.

(A technical note: The burden of this chapter is to argue, in effect, that there is no *a priori* entailment from physical facts to phenomenal facts. The sort of necessity that defines the relevant supervenience relation is the *a priori* version of logical necessity, where primary intensions are central. As we saw in Chapter 2, this is the relation that is relevant to issues about explanation; matters of *a posteriori* necessity can be set to one side. In the next chapter, issues of ontology rather than explanation are central, and I argue separately that there is no *a posteriori* necessary connection between physical facts and phenomenal facts.)

Argument 1:
The Logical Possibility of Zombies

The most obvious way (although not the only way) to investigate the logical supervenience of consciousness is to consider the logical possibility of a *zombie:* someone or something physically identical to me (or to any other conscious being), but lacking conscious experiences altogether. [1] At the global level, we can consider the logical possibility of a *zombie world:* a world physically identical to ours, but in which there are no conscious experiences at all. In such a world, everybody is a zombie.

So let us consider my zombie twin. This creature is molecule for molecule identical to me, and identical in all the low-level properties postulated by a completed physics, but he lacks conscious experience entirely. (Some might prefer to call a zombie "it," but I use the personal pronoun; I have grown quite fond of my zombie twin.) To fix ideas, we can imagine that right now I am gazing out the window, experiencing some nice green sensations from seeing the trees outside, having pleasant taste experiences through munching on a chocolate bar, and feeling a dull aching sensation in my right shoulder.

Image not available.

Figure 3.1.
Calvin and Hobbes on zombies.
(Calvin and Hobbes © Watterson.
Distributed by Universal Press Syndicate.
Reprinted with permission. All rights reserved)

What is going on in my zombie twin? He is physically identical to me, and we may as well suppose that he is embedded in an identical environment. He will certainly be identical to me *functionally:* he will be processing the same sort of information, reacting in a similar way to inputs, with his internal configurations being modified appropriately and with indistinguishable behavior resulting. He will be *psychologically* identical to me, in the sense developed in Chapter 1. He will be perceiving the trees outside, in the functional sense, and tasting the chocolate, in the psychological sense. All of this follows logically from the fact that he is physically identical to me, by virtue of the functional analyses of psychological notions. He will even be "conscious" in the functional senses described earlier—he will be awake, able to report the contents of his internal states, able to focus attention in various places, and so on. It is just that none of this functioning will be accompanied by any real conscious experience. There will be no phenomenal feel. There is nothing it is like to be a zombie.

This sort of zombie is quite unlike the zombies found in Hollywood movies, which tend to have significant functional impairments (Figure 3.1). The sort of consciousness that Hollywood zombies most obviously lack is a psychological version: typically, they have little capacity for introspection and lack a refined ability to voluntarily control behavior. They may or may not lack phenomenal consciousness; as Block (1995) points out, it is reasonable to suppose that there is something it tastes like when they eat their victims. We can call these *psychological zombies;* I am concerned with *phenomenal zombies,* which are physically and functionally identical, but which lack experience. (Perhaps it is not surprising that phenomenal zombies have not been popular in Hollywood, as there would be obvious problems with their depiction.)

Page 96

The idea of zombies as I have described them is a strange one. For a start, it is unlikely that zombies are naturally possible. In the real world, it is likely that any replica of me would be conscious. For this reason, it is most natural to imagine unconscious creatures as physically different from conscious ones—exhibiting impaired behavior, for example. But the question is not whether it is plausible that zombies could exist in our world, or even whether the idea of a zombie replica is a natural one; the question is whether the notion of a zombie is conceptually coherent. The mere intelligibility of the notion is enough to establish the conclusion.

Arguing for a logical possibility is not entirely straightforward. How, for example, would one argue that a mile-high unicycle is logically possible? It just seems obvious. Although no such thing exists in the real world, the description certainly appears to be coherent. If someone objects that it is not logically possible—it merely seems that way—there is little we can say, except to repeat the description and assert its obvious coherence. It seems quite clear that there is no hidden contradiction lurking in the description.

I confess that the logical possibility of zombies seems equally obvious to me. A zombie is just something physically identical to me, but which has no conscious experience—all is dark inside. While this is probably empirically impossible, it certainly seems that a coherent situation is described; I can discern no contradiction in the description. In some ways an assertion of this logical possibility comes down to a brute intuition, but no more so than with the unicycle. Almost everybody, it seems to me, is capable of conceiving of this possibility. Some may be led to deny the possibility in order to make some theory come out right, but the justification of such theories should ride on the question of possibility, rather than the other way around.

In general, a certain burden of proof lies on those who claim that a given description is logically *impossible.* If someone truly believes that a mile-high unicycle is logically impossible, she must give us some idea of where a contradiction lies, whether explicit or implicit. If she cannot point out something about the intensions of the concepts "mile-high" and "unicycle" that might lead to a contradiction, then her case will not be convincing. On the other hand, it is no more convincing to give an obviously false analysis of the notions in question—to assert, for example, that for something to qualify as a unicycle it must be shorter than the Statue of Liberty. If no reasonable analysis of the terms in question points toward a contradiction, or even makes the existence of a contradiction plausible, then there is a natural assumption in favor of logical possibility.

That being said, there are some positive things that proponents of logical possibility can do to bolster their case. They can exhibit various indirect arguments, appealing to what we know about the phenomena in question and the way we think about hypothetical cases involving these phenomena,

in order to establish that the obvious logical possibility really is a logical possibility, and really is obvious. One might spin a fantasy about an ordinary person riding a unicycle, when suddenly the whole system expands a thousandfold. Or one might describe a series of unicycles, each bigger than the last. In a sense, these are all appeals to intuition, and an opponent who wishes to deny the possibility can in each case assert that our intuitions have misled us, but the very obviousness of what we are describing works in our favor, and helps shift the burden of proof further onto the other side.

For example, we can indirectly support the claim that zombies are logically possible by considering ***nonstandard realizations*** of my functional organization. [2] My functional organization—that is, the pattern of causal organization embodied in the mechanisms responsible for the production of my behavior—can in principle be realized in all sorts of strange ways. To use a common example (Block 1978), the people of a large nation such as China might organize themselves so that they realize a causal organization isomorphic to that of my brain, with every person simulating the behavior of a single neuron, and with radio links corresponding to synapses. The population might control an empty shell of a robot body, equipped with sensory transducers and motor effectors.

Many people find it implausible that a set-up like this would give rise to conscious experience—that somehow a "group mind" would emerge from the overall system. I am not concerned here with whether or not conscious experience would *in fact* arise; I suspect that in fact it would, as I argue in Chapter 7. All that matters here is that the idea that such a system lacks conscious experience is *coherent*. A meaningful possibility is being expressed, and it is an open question whether consciousness arises or not. We can make a similar point by considering my silicon isomorph, who is organized like me but who has silicon chips where I have neurons. Whether such an isomorph would *in fact* be conscious is controversial, but it seems to most people that those who deny this are expressing a coherent possibility. From these cases it follows that the existence of my conscious experience is not logically entailed by the facts about my functional organization.

But given that it is conceptually coherent that the group-mind set-up or my silicon isomorph could lack conscious experience, it follows that my zombie twin is an equally coherent possibility. For it is clear that there is no more of a *conceptual* entailment from biochemistry to consciousness than there is from silicon or from a group of homunculi. If the silicon isomorph without conscious experience is conceivable, we need only substitute neurons for silicon in the conception while leaving functional organization constant, and we have my zombie twin. Nothing in this substitution could force experience into the conception; these implementational differences are simply not the sort of thing that could be conceptually relevant to experience. So consciousness fails to logically supervene on the physical.

The argument for zombies can be made without an appeal to these non-standard realizations, but these have a heuristic value in eliminating a source of conceptual confusion. To some people, intuitions about the logical possibility of an unconscious physical replica seem less than clear at first, perhaps because the familiar co-occurrence of biochemistry and consciousness can lead one to suppose a conceptual connection. Considerations of the less familiar cases remove these empirical correlations from the picture, and therefore make judgments of logical possibility more straightforward. [3] But once it is accepted that these nonconscious functional replicas are logically possible, the corresponding conclusion concerning a physical replica cannot be avoided.

Some may think that conceivability arguments are unreliable. For example, sometimes it is objected that we cannot really imagine in detail the many billions of neurons in the human brain. Of course this is true; but we do not need to imagine each of the neurons to make the case. Mere complexity among neurons could not conceptually entail consciousness; if all that neural structure is to be relevant to consciousness, it must be relevant *in virtue* of some higher-level properties that it enables. So it is enough to imagine the system at a coarse level, and to make sure that we conceive it with appropriately sophisticated mechanisms of perception, categorization, high-band-width access to information contents, reportability, and the like. No matter how sophisticated we imagine these mechanisms to be, the zombie scenario remains as coherent as ever. Perhaps an opponent might claim that all the unimagined neural detail is conceptually relevant in some way independent of its contribution to sophisticated functioning; but then she owes us an account of what that way might be, and none is available. Those implementational details simply lie at the wrong level to be conceptually relevant to consciousness.

It is also sometimes said that conceivability is an imperfect guide to possibility. The main way that conceivability and possibility can come apart is tied to the phenomenon of *a posteriori* necessity: for example, the hypothesis that water is not H_2O seems conceptually coherent, but water is arguably H_2O in all possible worlds. But *a posteriori* necessity is irrelevant to the concerns of this chapter. As we saw in the last chapter, explanatory connections are grounded in *a priori* entailments from physical facts to high-level facts. The relevant kind of possibility is to be evaluated using the primary intensions of the terms involved, instead of the secondary intensions that are relevant to *a posteriori* necessity. So even if a zombie world is conceivable only in the sense in which it is conceivable that water is not H_2O, that is enough to establish that consciousness cannot be reductively explained.

Those considerations aside, the main way in which conceivability arguments can go wrong is by subtle conceptual confusion: if we are insufficiently

reflective we can overlook an incoherence in a purported possibility, by taking a conceived-of situation and *misdescribing* it. For example, one might think that one can conceive of a situation in which Fermat's last theorem is false, by imagining a situation in which leading mathematicians declare that they have found a counterexample. But given that the theorem is actually true, this situation is being misdescribed: it is really a scenario in which Fermat's last theorem is true, and in which some mathematicians make a mistake. Importantly, though, this kind of mistake always lies in the *a priori* domain, as it arises from the incorrect application of the primary intensions of our concepts to a conceived situation. Sufficient reflection will reveal that the concepts are being incorrectly applied, and that the claim of logical possibility is not justified.

So the only route available to an opponent here is to claim that in describing the zombie world as a zombie world, we are misapplying the concepts, and that in fact there is a conceptual contradiction lurking in the description. Perhaps if we thought about it clearly enough we would realize that by imagining a physically identical world we are thereby *automatically* imagining a world in which there is conscious experience. But then the burden is on the opponent to give us some idea of where the contradiction might lie in the apparently quite coherent description. If no internal incoherence can be revealed, then there is a very strong case that the zombie world is logically possible.

As before, I can detect no internal incoherence; I have a clear picture of what I am conceiving when I conceive of a zombie. Still, some people find conceivability arguments difficult to adjudicate, particularly where strange ideas such as this one are concerned. It is therefore fortunate that every point made using zombies can also be made in other ways, for example by considering epistemology and analysis. To many, arguments of the latter sort (such as arguments 3–5 below) are more straightforward and therefore make a stronger foundation in the argument against logical supervenience. But zombies at least provide a vivid illustration of important issues in the vicinity.

Argument 2: The Inverted Spectrum

Even in making a conceivability argument against logical supervenience, it is not strictly necessary to establish the logical possibility of zombies or a zombie world. It suffices to establish the logical possibility of a world physically identical to ours in which the facts about conscious experience are merely *different* from the facts in our world, without conscious experience being absent entirely. As long as some positive fact about experience in our world does not hold in a physically identical world, then consciousness does not logically supervene.

It is therefore enough to note that one can coherently imagine a physically identical world in which conscious experiences are *inverted,* or (at the local level) imagine a being physically identical to me but with inverted conscious experiences. One might imagine, for example, that where I have a red experience, my inverted twin has a blue experience, and vice versa. Of course he will call his blue experiences "red," but that is irrelevant. What matters is that the experience he has of the things we both call "red"—blood, fire engines, and so on—is of the same kind as the experience I have of the things we both call "blue," such as the sea and the sky.

The rest of his color experiences are systematically inverted with respect to mine, in order that they cohere with the red-blue inversion. Perhaps the best way to imagine this happening with human color experiences is to imagine that two of the axes of our three-dimensional color space are switched—the red-green axis is mapped onto the yellow-blue axis, and vice versa. [4] To achieve such an inversion in the actual world, presumably we would need to rewire neural processes in an appropriate way, but as a *logical* possibility, it seems entirely coherent that experiences could be inverted while physical structure is duplicated exactly. Nothing in the neurophysiology dictates that one sort of processing should be accompanied by red experiences rather than by yellow experiences.

It is sometimes objected (Harrison 1973; Hardin 1987) that human color space is asymmetrical in a way that disallows such an inversion. For instance, certain colors have a warmth or coolness associated with them, and warmth and coolness appear to be directly associated with different functional roles (e.g., warmth is perceived as "positive," whereas coolness is perceived as "negative"). If a warm color and a cool color were switched, then the "warm" phenomenal feel would become dissociated from the "warm" functional role—a "cool" green experience would be reported as positive rather than negative, and so on. In a similar way, there seem to be more discriminable shades of red than of yellow, so swapping red experiences with yellow experiences directly might lead to the odd situation in which a subject could functionally discriminate more shades of yellow than are distinguishable phenomenologically. Perhaps there are enough asymmetries in color space that any such inversion would lead to a strange dissociation of phenomenal feel from the "appropriate" functional role.

There are three things we can say in response to this. First, there does not seem to be anything *incoherent* about the notion of such a dissociation (e.g., cool phenomenology with warm reactions), although it is admittedly an odd idea.[5] Second, instead of mapping red precisely onto blue and vice versa, one can imagine that these are mapped onto slightly different colors. For example, red might be mapped onto a "warm" version of blue (as Levine [1991] suggests), or even onto color not in our color space at all. In the

red–yellow case, we might imagine that red is mapped onto an extended range of yellow experiences, in which more discrimination is available. There is no reason why spectrum inversion scenarios *must* involve colors drawn from the usual color space. Third, perhaps the most compelling response is to argue (with Shoemaker [1982]) that even if our own color space is asymmetrical, there certainly *could* be creatures whose color space is symmetrical. For example, there is probably a naturally possible creature who sees (and experiences) precisely two colors, *A* and *B*, which correspond to distinct, well-separated ranges of light wavelengths, and for which the distinction between the two exhausts the structure of the color space. It seems entirely coherent to imagine two such creatures that are physically identical, but whose experiences of *A* and *B* are inverted. That is enough to make the point.

Even many reductive materialists (e.g., Shoemaker [1982]) have conceded that it is coherent that one's color experiences might be inverted while one's functional organization stays constant. It is allowed that a system with different underlying neurophysiological properties, or with something like silicon in place of neurobiology, might have different color experiences. But once this is granted, it follows automatically that inversion of experiences in a physical replica is at least conceptually coherent. The extra neurophysiological properties that are constrained in such a case are again not the kind of thing that could logically determine the nature of the experience. Even if there is some sort of *a posteriori* identification between certain neurophysiological structures and certain experiences (as Shoemaker believes), we must still allow that a different pattern of associations is conceivable, in the sense of conceivability that is relevant to reductive explanation.

While the possibility of inverted spectra and the possibility of zombies both establish that consciousness fails to supervene logically, the first establishes a conclusion strictly weaker than the second. Somebody might conceivably hold that inverted spectra but not zombies are logically possible. If this were the case, then the *existence* of consciousness could be reductively explained, but the specific *character* of particular conscious experiences could not be.

Argument 3:
From Epistemic Asymmetry

As we saw earlier, consciousness is a surprising feature of the universe. Our grounds for belief in consciousness derive solely from our own experience of it. Even if we knew every last detail about the physics of the universe —the configuration, causation, and evolution among all the fields and particles in the spatiotemporal manifold —*that* information would not lead us to postulate the existence of conscious experience. My knowledge of consciousness,

in the first instance, comes from my own case, not from any external observation. It is my first-person experience of consciousness that forces the problem on me.

From all the low-level facts about physical configurations and causation, we can in principle derive all sorts of high-level facts about macroscopic systems, their organization, and the causation among them. One could determine all the facts about biological function, and about human behavior and the brain mechanisms by which it is caused. But nothing in this vast causal story would lead one who had not experienced it directly to believe that there should be any *consciousness.* The very idea would be unreasonable; almost mystical, perhaps.

It is true that the physical facts about the world might provide some indirect evidence for the existence of consciousness. For example, from these facts one could ascertain that there were a lot of organisms that *claimed* to be conscious, and said they had mysterious subjective experiences. Still, this evidence would be quite inconclusive, and it might be most natural to draw an eliminativist conclusion—that there was in fact no *experience* present in these creatures, just a lot of talk.

Eliminativism about conscious experience is an unreasonable position *only* because of our own acquaintance with it. If it were not for this direct knowledge, consciousness could go the way of the vital spirit. To put it another way, there is an *epistemic asymmetry* in our knowledge of consciousness that is not present in our knowledge of other phenomena. [6] Our knowledge that conscious experience exists derives primarily from our own case, with external evidence playing at best a secondary role.

The point can also be made by pointing to the existence of a problem of other minds. Even when we know everything physical about other creatures, we do not *know* for certain that they are conscious, or what their experiences are (although we may have good reason to believe that they are). It is striking that there is no problem of "other lives," or of "other economies," or of "other heights." There is no epistemic asymmetry in those cases, precisely because those phenomena are logically supervenient on the physical.

The epistemic asymmetry in knowledge of consciousness makes it clear that consciousness cannot logically supervene. If it were logically supervenient, there would be no such epistemic asymmetry; a logically supervenient property can be detected straightforwardly on the basis of external evidence, and there is no special role for the first-person case. To be sure, there are some supervenient properties—memory, perhaps—that are more easily detected in the first-person case. But this is just a matter of how hard one has to work. The presence of memory is just as accessible from the third person, in principle, as from the first person. The epistemic asymmetry associated with consciousness is much more fundamental, and it tells us that no

collection of facts about complex causation in physical systems adds up to a fact about consciousness.

Argument 4:
The Knowledge Argument

The most vivid argument against the logical supervenience of consciousness is suggested by Jackson (1982), following related arguments by Nagel (1974) and others. Imagine that we are living in an age of a completed neuroscience, where we know everything there is to know about the physical processes within our brain responsible for the generation of our behavior. Mary has been brought up in a black-and-white room and has never seen any colors except for black, white, and shades of gray. [7] She is nevertheless one of the world's leading neuroscientists, specializing in the neurophysiology of color vision. She knows everything there is to know about the neural processes involved in visual information processing, about the physics of optical processes, and about the physical makeup of objects in the environment. But she does not know what it is like to see red. No amount of reasoning from the physical facts alone will give her this knowledge.

It follows that the facts about the subjective experience of color vision are not entailed by the physical facts. If they were, Mary could in principle come to know what it is like to see red on the basis of her knowledge of the physical facts. But she cannot. Perhaps Mary could come to know what it is like to see red by some indirect method, such as by manipulating her brain in the appropriate way. The point, however, is that the knowledge does not follow from the physical knowledge alone. Knowledge of all the physical facts will in principle allow Mary to derive all the facts about a system's reactions, abilities, and cognitive capacities; but she will still be entirely in the dark about its experience of red.

A related way to make this point is to consider systems quite different from ourselves, perhaps much simpler—such as bats or mice—and note that the physical facts about these systems do not tell us what their conscious experiences are like, if they have any at all (Nagel focuses on this sort of issue). Once all the physical facts about a mouse are in, the nature of its conscious experience remains an *open question:* it is consistent with the physical facts about a mouse that it has conscious experience, and it is consistent with the physical facts that it does not. From the physical facts about a bat, we can ascertain *all* the facts about a bat, except the facts about its conscious experiences. Knowing all the physical facts, we still do not know what it is like to be a bat.

Along similar lines we can consider a computer, designed as a simple cognitive agent (perhaps it has the intelligence of a dog), but similar to us in certain respects, such as its capacity for perceptual discrimination. In

particular it categorizes color stimuli in a manner quite similar to ours, grouping things that we would call "red" under one category and things we would call "green" under another. Even if we know every detail about the computer's circuits, questions remain: (1) Is the computer experiencing anything at all when it looks at roses?; (2) If it is, is it experiencing the same sensory color quality that we have when we look at a rose, or some quite different quality? These are entirely meaningful questions, and knowing all the physical facts does not force one answer rather than another onto us. The physical facts therefore do not logically entail the facts about conscious experience.

Jackson put his argument forward as an argument against materialism rather than against reductive explanation. There have been many replies to the argument; I will discuss them in the next chapter, where materialism rather than reductive explanation will be at issue. But for now it is interesting to note that most of the objections to the argument against materialism have *conceded* the point that is relevant to the argument against reductive explanation: that knowledge of what red is like is factual knowledge that is not entailed *a priori* by knowledge of the physical facts. The only way that the conclusion can be evaded is to deny that knowing what red experience is like gives knowledge of a *fact* at all. This is the strategy taken by Lewis (1990) and Nemirow (1990), who argue that all Mary is lacking is an *ability,* such as the ability to recognize red things. I discuss this suggestion in the next chapter; here, I simply note that insofar as it seems clear that when she sees red for the first time, Mary is *discovering* something about the way the world is, it seems clear that the knowledge she is gaining is knowledge of a fact.

Argument 5:
From the Absence of Analysis

If proponents of reductive explanation are to have any hope of defeating the arguments above, they will have to give us some idea of how the existence of consciousness *might* be entailed by physical facts. While it is not fair to expect all the details, one at least needs an account of how such an entailment might *possibly* go. But any attempt to demonstrate such an entailment is doomed to failure. For consciousness to be entailed by a set of physical facts, one would need some kind of analysis of the notion of consciousness—the kind of analysis whose satisfaction physical facts could imply—and there is no such analysis to be had.

The only analysis of consciousness that seems even remotely tenable for these purposes is a functional analysis. Upon such an analysis, it would be seen that all there is to the notion of something's being conscious is that it should play a certain functional role. For example, one might say that all there is to a state's being conscious is that it be verbally reportable, or that it be the result of certain kinds of perceptual discrimination, or that it make information available to later processes in a certain way, or whatever. But on the face of it, these fail miserably as analyses. They simply miss what it means to be a conscious experience. Although conscious states may play various causal roles, they are not *defined* by their causal roles. Rather, what makes them conscious is that they have a certain phenomenal feel, and this feel is not something that can be functionally defined away.

To see how unsatisfactory these analyses are, note how they trivialize the problem of explaining consciousness. Suddenly, all we have to do to explain consciousness is explain our ability to make certain verbal reports, or to perform certain sorts of discrimination, or to manifest some other capacity. But on the face of it, it is entirely conceivable that one could explain all these things without explaining a thing about consciousness itself; that is, without explaining the *experience* that accompanies the report or the discrimination. To analyze consciousness in terms of some functional notion is either to change the subject or to define away the problem. One might as well define "world peace" as "a ham sandwich." Achieving world peace becomes much easier, but it is a hollow achievement.

Functional analyses of consciousness can also be argued against on more specific grounds. For example, any functionally analyzed concept will have a degree of semantic indeterminancy. Does a mouse have beliefs? Do bacteria learn? Is a computer virus alive? The best answer to these questions is usually in a sense yes, in a sense no. It all depends on how we draw the boundaries in the concepts, and in any high-level functional concepts the boundaries will be vague. But compare: Does a mouse have conscious experience? Does a virus? These are not matters for stipulation. Either there is something that it is like to be a mouse or there is not, and it is not up to us to define the mouse's experience into or out of existence. To be sure, there is probably a continuum of conscious experience from the very faint to the very rich; but if something has conscious experience, however faint, we cannot stipulate it away. This determinacy could not be derived from any functional analysis of the concepts in the vicinity of consciousness, as the functional concepts in the vicinity are all somewhat vague. If so, it follows that the notion of consciousness cannot be functionally analyzed.

Another objection is that the functionalist analysis collapses the important distinction, outlined in Chapter 1, between the notions of awareness and consciousness. Presumably if consciousness is to be functionally analyzed, it will be analyzed roughly as we analyzed awareness then: in terms of a certain accessibility of information in later processing and in the control of behavior. Awareness is a perfectly good concept, but it is quite distinct from the concept of conscious experience. The functionalist treatment collapses the two notions of consciousness and awareness into one, and therefore does not do justice to our conceptual system.

The alternatives to functional analysis look even worse. It is most unclear that there could be any other kind of analysis appropriate for reductive explanation. The only alternative might be a structural analysis—perhaps consciousness could be analyzed as some sort of biochemical structure—but that analysis would be even more clearly inadequate. Whether or not consciousness *is* a biochemical structure, that is not what "consciousness" *means*. To analyze consciousness that way again trivializes the explanatory problem by changing the subject. It seems that the concept of consciousness is irreducible, being characterizable only in terms of concepts that themselves involve consciousness.

Note that this is quite unlike the sort of irreducibility that is sometimes supposed to hold for high-level concepts in general. We have seen that many high-level notions have no crisp definitions, and no manageable analyses in terms of necessary and sufficient conditions. Nevertheless, as we saw in the last chapter, these concepts at least have rough-and-ready analyses that get us into the ballpark, although they will inevitably fail to do justice to the details. Most importantly, it is easy to see that properties such as life, learning, and so on can be analyzed as functional properties, even if spelling out the details of just *which* functional property is a difficult matter. Even though these properties lack crisp functional definitions, they are nevertheless quite compatible with entailment by the physical facts.

The problems with consciousness are in a different league. Here, the purported analyses do not even get into the ballpark. In a much starker way, they completely fail to characterize what needs to be explained. There is no temptation to even *try* to add epicycles to a purported functional analysis of consciousness in order to make it satisfactory, as there is with similar analyses of life and of learning. Consciousness is simply not to be characterized as a functional property in the first place. The same goes for analyses of consciousness as a structural property, or in other reductive terms. There is therefore no way for an entailment from physical facts to consciousness to get off the ground.

2—
The Failure of Reductive Explanation

The failure of consciousness to logically supervene on the physical tells us that no reductive explanation of consciousness can succeed. Given any account of the physical processes purported to underlie consciousness, there will always be a further question: Why are these processes accompanied by conscious experience? For most other phenomena, such a question is easily answered: the physical facts about those processes *entail* the existence of the phenomena. For a phenomenon such as life, for example, the physical facts imply that certain functions will be performed, and the performance of those functions is

all we need to explain in order to explain life. But no such answer will suffice for consciousness.

Physical explanation is well suited to the explanation of *structure* and of *function*. Structural properties and functional properties can be straightforwardly entailed by a low-level physical story, and so are clearly apt for reductive explanation. And almost all the high-level phenomena that we need to explain ultimately come down to structure or function: think of the explanation of waterfalls, planets, digestion, reproduction, language. But the explanation of consciousness is not just a matter of explaining structure and function. Once we have explained all the physical structure in the vicinity of the brain, and we have explained how all the various brain functions are performed, there is a further sort of explanandum: consciousness itself. Why should all this structure and function give rise to experience? The story about the physical processes does not say.

We can put this in terms of the thought experiments given earlier. Any story about physical processes applies equally to me and to my zombie twin. It follows that nothing in that story says why, in my case, consciousness arises. Similarly, any story about physical processes applies equally to my inverted twin, who sees blue where I see red: it follows that nothing in that story says why my experience is of one variety rather than another. The very fact that it is logically possible that the physical facts could be the same while the facts about consciousness are different shows us that as Levine (1983) has put it, there is an *explanatory gap* between the physical level and conscious experience.

If this is right, the fact that consciousness accompanies a given physical process is a *further fact*, not explainable simply by telling the story about the physical facts. In a sense, the accompaniment must be taken as brute. We might try to systematize and explain these brute facts in terms of some simple underlying pattern, but there will always remain an element here that is logically independent of the physical story. Perhaps we might get some kind of explanation by combining the underlying physical facts with certain further *bridging* principles that link the physical facts with consciousness, but this explanation will not be a reductive one. The very need for explicit bridging principles shows us that consciousness is not being explained reductively, but is being explained on its own terms.

Of course nothing I have said implies that physical facts are *irrelevant* to the explanation of consciousness. We can still expect physical accounts to play a significant role in a theory of consciousness, giving information about the physical *basis* of consciousness, for example, and perhaps yielding a detailed correspondence between various aspects of physical processing and aspects of conscious experience. Such accounts may be especially useful in helping to understand the *structure* of consciousness: the patterns of similarity and difference between experiences, the geometric structure of phenome-

nal fields, and so on. I say much more about these and other things that physical explanation can tell us about experience in a nonreductive framework in Chapter 6. But a physical account, alone, is not *enough.*

At this point, a number of objections naturally arise.

Objection 1:
Are We Setting the Standards Too High?

Some might argue that explanation of *any* high-level phenomena will postulate " bridge laws" in addition to a low-level account, and that it is only with the aid of these bridge laws that the details of the high-level phenomena are derived. However, as the discussion in the last chapter suggests (and as is carefully argued by Horgan [1978]), in such cases the bridge laws are not further facts about the world. Rather, the connecting principles themselves are logically supervenient on the low-level facts. The extreme case of such a bridging principle is a supervenience conditional, which we have seen is usually a conceptual truth. Other more "localized" bridging principles, such as the link between molecular motion and heat, can at least be derived from the physical facts. For consciousness, by contrast, such bridging principles must be taken as primitive.

It is interesting to see how a typical high-level property—such as life, say—evades the arguments put forward in the case of consciousness. First, it is straightforwardly inconceivable that there could be a physical replica of a living creature that was not itself alive. Perhaps a problem might arise due to context-dependent properties (would a replica that forms randomly in a swamp be alive, or be human?), but fixing environmental facts eliminates even that possibility. Second, there is no "inverted life" possibility analogous to the inverted spectrum. Third, when one knows all the physical facts about an organism (and possibly about its environment), one has enough material to know all the biological facts. Fourth, there is no epistemic asymmetry with life; facts about life in others are as accessible, in principle, as facts about life in ourselves. Fifth, the concept of life is plausibly analyzable in functional terms: to be alive is roughly to possess certain capacities to adapt, reproduce, and