2
The Argument against Physicalism

2.1 Introduction

Physicalism says that the fundamental physical facts are the only fundamental facts. All other facts, whether about rocks, tables, morals, or minds, are derivative on these physical facts. In this chapter, I argue that physicalism is false by arguing that a purely physical world could not contain facts of experience. Others have given arguments of this kind, but I hope to look at this kind of argument in a fresh way. In chapter 3 I defend the argument against objections.

My argument is not a form of conceivability argument or knowledge argument. It is a direct argument that the phenomenal facts are of a type that cannot be entailed, either a priori or a posteriori,[i] by the physical facts. To diagnose precisely why entailment fails, I produce a working analysis of physical facts as a type. This working analysis is central to this chapter, and it recurs in part II. Because the specific lessons of this chapter's argument hold recurring importance, I ask even readers who are familiar (or impatient) with the debate over physicalism to pay some attention to this chapter.

2.2 The Dialectic

Recent antiphysicalist arguments rely on thought experiments that claim to show limits on the physicalist program for explanation and, by implication, the metaphysical status of physicalism. In his seminal paper, "What Is It Like to Be a Bat?" (1974), Thomas Nagel argues that any physicalist account of the universe, by being inherently objective, will leave out the subjectivity of points of view. Nagel argues that this omission is reflected in the fact that even when we know all about the physiology of creatures that are very different from us, we do not know what it is like to be them.

Among others, Frank Jackson (1982) and David Chalmers (1996) have refined

Nagel's guiding intuitions. In Jackson's well-known Knowledge Argument, he asks that we consider a superneuroscientist named Mary. From within a black-and-white room, through books and observation of a black-and-white TV, Mary learns everything there is to know about the functioning of the visual system. Jackson maintains that, nevertheless, Mary learns something the first time she is exposed to color. She learns what the experience of blue is like, for instance. Jackson claims that it follows that physicalism must be false because we can know all the physical facts without being able to know, even in principle, *all* the facts.

Chalmers's Conceivability Argument asks us to conceive of a universe physically identical to ours from Big Bang to Big Crunch, but with the twist that our counterparts have no conscious mental life. They are subjective zombies. Chalmers argues that such a universe is conceivable and, furthermore, metaphysically possible. He argues that this shows the falsity of physicalism by showing that the facts about qualitative consciousness are further facts, not determined in the appropriate way by the physical facts.

By using thought experiments, the antiphysicalists aim to show that there is no *entailment* from physical facts to facts about experience, where an entailment is understood as an a priori implication (*A* a priori entails *B* if one can rule out a priori that *A* is true and *B* is false). That is, they aim to show that facts about experience cannot in principle be deduced from physical facts by a priori reasoning. From there, the antiphysicalists argue that physicalism is false. Later I argue against entailment in a different and more general way, using an analogy to an artificial world with a toylike physics. This analogy allows us to diagnose exactly why no kind of entailment can hold in the real world. So, despite being introduced with an analogy, the result is a direct argument against entailment that does not rely on a conceivability claim or the knowledge argument.

2.3 The Game of Life

The rest of this chapter develops a detailed analogy between the physical facts and the facts about *cellular automata*. *Cellular automata* names a certain class of artificial, digital worlds. A cellular automaton consists of points, or "cells," located in an abstract space, all of which can have kinds of "causal" properties. Computer modelers define various physics for these worlds and study the behaviors they exhibit. To start an automaton, one assigns an initial distribution of causal properties to the cells, perhaps at random. The automaton then evolves, changing states according to rules that apply pointwise to the space. These rules are its physics. Typically, the rules that determine which properties a cell will have at a given time are a function on the properties of neighboring cells at an immediately preceding time. One then studies what kinds of entities can evolve and what sorts of properties these entities can have, given the physics that the modeler has created.

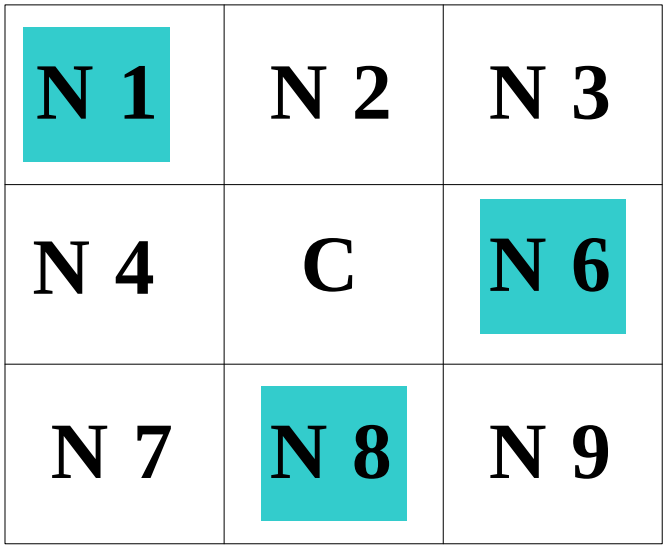
Life is the name of a kind of cellular automaton that evolves on a two-dimensional grid. The *Life* world has been used in discussion of the mind-body problem before, most notably in Dennett (1991a), and its physics is extremely simple

and easy to understand. For these reasons, I am also going to use the *Life* world as my example cellular automaton. I define a *pure Life world* as follows:

Definition: A world is a *pure Life world* if, and only if, it is a *Life* world of which no fundamental facts are true except those stipulated in its physics.

In *Life*, we are supposed to think of each cell on the grid as a square and as having eight neighbors: a neighbor touching it on each side and a neighbor touching it on each corner. The location of a cell never changes. Additionally, a cell can host exactly one of two mutable causal properties, being *on* or being *off*, at any given time step. To illustrate the basic scheme, figure 2.1 depicts a cell and its neighbors. Three simple rules govern the evolution of a *Life* automaton:

1. If a cell has exactly two *on* neighbors, it maintains its property, *on* or *off*, in the next time step.
2. If a cell has exactly three *on* neighbors, it will be *on* in the next time step.
3. Otherwise, the cell will be *off* in the next time step.



<FN> Figure 2.1 </FN> <FT> A depiction of a center cell, C, and its "neighborhood" on a Life world grid. Neighbor's N1, N6, and N8 are depicted as "on". </FT>

Imagine a *Life* universe consisting of an infinite grid. The two properties possessed by grid cells, *on* and *off*, are the basic physical properties in the *Life* universe. The rules governing the grid's evolution are that universe's laws of physics. When thought about in this way, *Life* becomes a good modeling ground for understanding how physical facts can entail other kinds of facts.

Despite its simple physics, the *Life* automaton can evidence a tremendous variety of patterns. For instance, John Conway, the mathematician who invented it, proved that a *Life* grid can be a universal Turing machine.[ii] More remarkably, he has proven that the grid can support extremely complex patterns that are self-replicating in von Neumann's sense of nontrivial self-replication (Poundstone 1985). These patterns have functional properties similar to DNA and provide the motivation for the name *Life*. In general, it is the interesting patterns like these in *Life* that create entailments from its basic physical facts to facts of other kinds.

Entities called gliders serve as a simple example of how entailment works in the *Life* universe. A glider consists of a sequence of patterns, each containing exactly five contiguous cells, which move across the grid in a characteristic fashion (see figure 2.2). Gliders make for a useful example because other cellular automata can also produce gliders. This means that *Life* can present *sufficient* conditions for the existence of gliders but cannot present *necessary* conditions, so we cannot *define* the property of being a glider in terms of *Life* physics. To be a glider just means to have a certain structure and to evolve in a certain way, regardless of the underlying physics. The glider structure produces a predictable range of successive states that, lacking interference, move across the grid.

Even before seeing what a stage in the glider pattern actually does when we instantiate it in a *Life* world, we know that *Life* will allow for structure to arise and for the evolution of those structures. Seeing this, it then becomes obvious that *Life* worlds (epistemically) *might* support conditions that entail the existence of gliders. To rule out the (epistemic) possibility that gliders could exist in a *Life* universe, we would need a specific proof that the physics could not produce them.

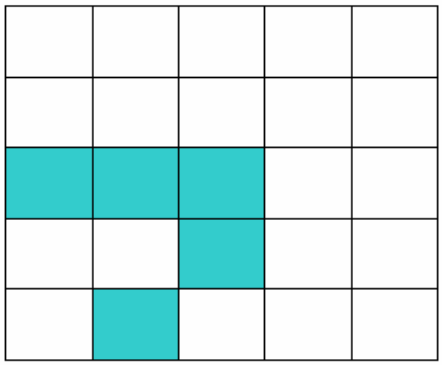
As it turns out, the *Life* physics can produce gliders. One can prove this by taking a pure *Life* world, producing one of the configurations in the life cycle of a glider, and checking that it evolves correctly over time. It does, so we see that *Life* worlds can entail the existence of gliders.[iii]

In this example, entailment acts as a determination relation: The basic facts in *Life* are the facts about the distribution of the "on" and "off" properties and how they redistribute over time. Also, the basic *Life* facts necessitate the facts about gliders without our having to introduce any new fundamental ontology. Instead, the necessity is grounded in conceptual truths about what it means to be a glider combined with the empirical truths about the configurations of the basic properties in the *Life* world and the evolution of those configurations. Given a situation in the *Life* world, these interpretive truths are enough to determine the truth of facts about gliders.[iv]

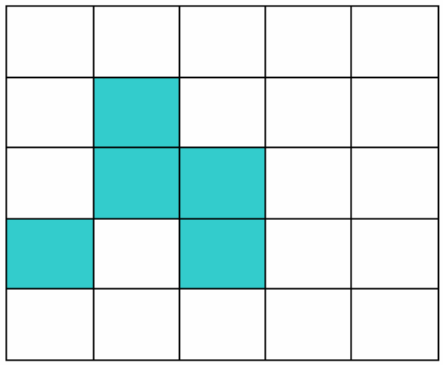
2.4 The Form of the Argument against Physicalism

With this understanding of the *Life* world in mind, the argument against physicalism that I defend is:

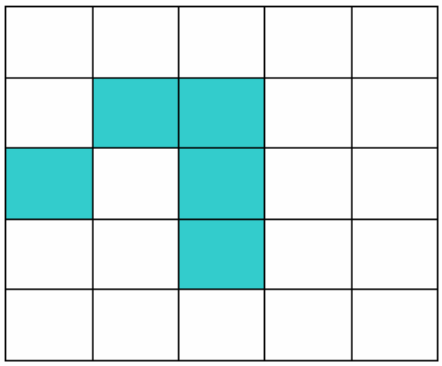
Figure 2. 2



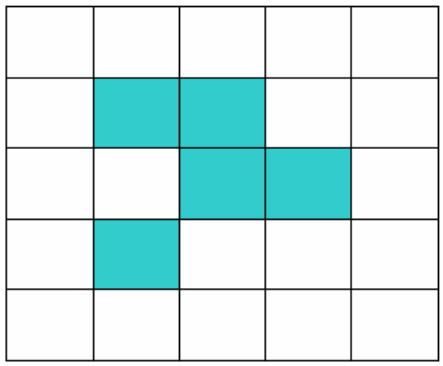
(a) One state of a glider



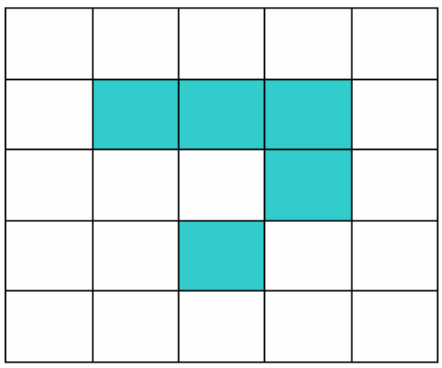
(b) The next state of the glider



(c) The next state of the glider



(d) The next state of the glider



(e) The next state of the glider

A full cycle of the repeating states in the life of a glider. Notice that the glider in figure (e) is a mirror of the glider in figure (a), only moved up the grid by one cell.

<FN> Figure 2.2 </FN> <FT> One full cycle of states in the existence of a glider. </FT>

1. Facts about a pure *Life* world do not entail facts about phenomenal consciousness (either a priori or a posteriori).
2. If facts about a pure *Life* world do not entail facts about phenomenal consciousness, then facts about a pure physical world do not entail facts about phenomenal consciousness.
3. Therefore, facts about a pure physical world do not entail facts about phenomenal consciousness.

This is my overall argument. By presenting it, I will lock onto a theoretical conception of what it means to be physical and to be entailed by the physical facts. My strategy is to use the physics of the *Life* world to draw out the categorical structure of physical theories in general, identifying the kinds of information physical theories convey and exposing the kinds of conditions that make physical properties the kinds of properties they are.

2.5 The Argument against Life Entailing Consciousness

Facts about a pure Life world do not entail facts about phenomenal consciousness. I defend the first premise of my argument against physicalism by defending something that I call *the Skeptic's Claim*. The Skeptic's Claim is that the facts about a pure *Life* universe cannot entail facts about consciousness. The skeptic's use of "entail" includes both a priori and a posteriori entailment. Thus we may consistently acknowledge any kind of structure and functionality for *Life* objects and still deny the presence of consciousness in a *Life* universe. The argument I defend for the skeptic is:

1. The fundamental properties of a pure *Life* world consist of bare differences.
2. Facts about phenomenal consciousness include facts about qualitative content.
3. Facts about bare difference cannot entail facts about qualitative content.
4. Therefore, some facts about phenomenal consciousness are not entailed by pure *Life* facts.

Premise 1: Pure Life worlds consist of bare differences What is a *bare difference*? I mean the phrase *bare difference* to express an intuitive idea that can be loosely explained by saying that *Life*'s physics leaves us in the dark about what the "on" and "off" properties are themselves. It just tells us that they are different and enter into certain dynamic relations.

What is an "on" property? It is not the "off" property. What is the "off" property? It is not the "on" property. That, plus the rules of evolution, is all *Life's* physics specifies about the "on" and "off" properties. In this way, bare differences are defined circularly in terms of their difference from each other. Moreover, if the *Life* world is pure, we know that there are just no other facts about those properties because we know that the physics tells us everything there is to know. I say the difference is *bare* because it does not rest on any further categorical facts about the properties (if the world is pure). It is a difference that is

ungrounded by any further facts about internal structural differences between those entities or internal relations of difference or contrast between unspecified structureless intrinsic contents.

Postulating facts about intrinsic natures in the *Life* world would violate the purity condition we are working under, because no facts about intrinsic natures are specified by its physics. Thus a *Life* world with any basis other than bare difference would be an impure *Life* world. For now, I think the best way to conceive of a bare difference between two properties, *x* and *y,* is to think of the relation as primary rather than implied by other facts, with the existence of the relata, such as they are, derivative on their participation in an ungrounded relation of difference.

Bare differences are difficult to conceive of. Some readers may reject the idea altogether, insisting that a *Life* world must have some kind of intrinsic basis. With an intrinsic basis, there would be a contentful difference where the existence of the difference would be derivative on further facts of intrinsic difference between some unspecified natures of the relata. It is clear, however, that such facts about an intrinsic basis would go beyond what is specified by the bare laws of the *Life* world. I argue later that such an intrinsic basis is crucial to the production of consciousness, but to presuppose it now would beg the question about whether pure physics can specify an adequate basis for the world. So I stay with the "bare" understanding of *Life* for now and examine it more critically later in the book.

My defense of the Skeptic's Claim begins with a closer look at the materials available in the *Life* universe. To reiterate: What does it mean to be an *on* or *off* property? The only two requirements are that (1) they should be distinct and (i2) they should be instantiated in patterns conforming to the rules set down by the three dynamical laws. In short, the distinction between being *on* or being *off* is a merely formal one. *On* and *off* specify bare, content-free difference.

Because it specifies only bare difference, the *Life* specification is, at heart, a structural schema for a universe. It specifies certain patterns of contrast between kinds of being, patterns that must hold for a universe to count as a *Life* universe. As we ascend to higher levels of organization in the pure *Life* world, we do not escape from the circle of bare difference. In pure *Life* we have a world potentially consisting of a huge number of simple, bare differences lying side by side, with reliable, regular transitions between them. A *Life* structure is a pattern of bare difference, mere contrast.

Premise 2: Consciousness contains qualitative content The skeptic claims that we have observational knowledge that consciousness contains qualitative content, usually called *qualia*. The claim that knowledge of qualitative content is observable is critical to the force of the skeptic's arguments. Without it, there is no strong reason to resist performing a *modus tollens* on the conclusion, simply eliminating phenomenal consciousness and its troublesome qualities from our list of explanatory targets. In the following I support the Skeptic's Claim by provid

ing a direct argument that qualia are indeed observables. By calling qualia *observables,* I mean that they meet four conditions:

1. They belong to a type whose members are potential objects of awareness.
2. We can become aware of them without the aid of special instruments.
3. The dubitability of our belief in facts of the relevant type is almost zero.
4. Our awareness of instances of the type is reliable.

Some people do raise objections to the claim that qualia are observables (e.g., Wilkes 1988; Dennett 1991; Akins 1993). The most common worry is that modeling our knowledge of qualia on perception is misleading, so people are unsure how we can be observing them. Minimally, opponents sympathetic to these eliminativist worries hold that the knowledge grounding the skeptic's conclusions is highly refined, theoretical, and corrigible.

To these worries, the skeptic replies that the objector seems to have an unreasonably narrow concept of observation. By insisting that something can achieve the status of an observable only if we obtain the information about it through ordinary perception, the objector is making too strong a claim. The objector rules out of court a huge amount of information about consciousness that we have access to and that a theory of consciousness should have to explain. I defend the following argument that qualia are observables:

1. Some thoughts and memories are observables.
2. If thoughts and memories are observable, then the evidence for them is observable.
3. Phenomenal contents (i.e., qualia) provide evidence for observable kinds of thoughts and memories.
4. Therefore, qualia are observable.

As examples of observable thoughts and memories, here are two that statements most would agree express observable facts:

(A) Last night I thought about my childhood.
(B) Sometimes I think about my childhood when no one else is around.

The previously defined characteristics of observables allow facts such as (A) and (B) to attain the status of useful falsifiers for scientific and other theories. For example, a theory of mind fails to account for some of our evidence about ourselves if it fails to account for how we can sometimes think about our childhoods when no one is around.

Facts such as (A) and (B) are no *more* problematic than many other facts we count as observable. Also, they are *introspectively* observable, and the fact that perception does not mediate our awareness of them seems like a red herring. Basically, if *anything* counts as observable, then (A) and (B) must count as observable, too. Our skeptic firmly insists that a science of mind must recognize observables such as these if it wants to be treated as legitimate. Because facts such as (A) and (B) turn out to be no more problematic as observables than are per

ceptually mediated facts, a straightforward argument delivers the phenomenal qualities as observable also.

Last night, I lacked behavioral evidence that I was thinking about my childhood. I was not writing about it, nor talking about it, nor acting on it. I was, in fact, scouring my bathtub. How do I know what I was thinking about? What was the evidence of my thoughts? I introspectively observed my thoughts, and my evidence was the presence of certain kinds of conscious phenomenal imagery, verbal, imagistic, and kinematic: phenomenal images of childhood scenes, spoken and heard sounds, and remembered emotions. That imagery may have been identical to the thoughts or it may just be a concomitant of thinking that gives evidence for thoughts the way that snow on the ground gives evidence for cold weather.

In either case, my awareness of the conscious phenomenal imagery cannot be considered more doubtful than my awareness of my thoughts. Because the phenomenal imagery is the evidence for such thoughts, it is easy to argue that[v]: the sentence (A) has the status of an observation claim only if the phenomenal imagery that is my evidence for it has the status of being observable. Similarly, I obtain my knowledge of *types* of thoughts such as those referred to in sentence (B) from observables only if I also obtain my knowledge of types of phenomenal qualities from observables.

> Arguments such as this, the skeptic maintains, establish that we obtain knowledge of what the phenomenal qualities (colors, feelings, sounds, imagery, other sensations) are like through observation. For example, I obtain my knowledge of what the shades of blue look like to me by consciously experiencing them. Consequently, phenomenal qualities are observables (which is not to say observation of them is always either easy or incorrigible). As scientists, we must hold explanations and theories accountable for phenomenal information obtained through observation.

This conclusion does not cross David Lewis's (1995) recommendation that physicalists must deny that we have special, unmediated access to the true nature of qualia. To possess phenomenal information, our skeptic does not need to have a more direct access to qualia than to any other kind of observable. The skeptic is chiefly concerned with the character of the connection between phenomenal qualities, as disclosed through the phenomenal information we *do* have available, and their hidden natures, if they have hidden natures.

Premise 3: Bare difference does not entail qualitative content Could conscious experience with its qualitative content arise from bare difference? Bare differences within cellular automata are a surprisingly fruitful ground for the emergence of an incredibly large number of kinds of things. *Life* itself can exhibit phenomena of indefinite complexity. For instance, because we already know that *Life* may contain self-replicating phenomena, we cannot rule out that it could exhibit some kind of genuine life. Because *Life* supports the existence of objects that dynamically evolve, it is at least an epistemic possibility that these entities might eventually lead to the existence of animate objects. We also have to hold it

as epistemically possible that these objects might metabolize elements of their environment, act in a goal-directed manner, adapt to be increasingly complex, and generally possess a suite of functional properties sufficient for regarding them as alive.

Given that life might exist, ecologies might exist. Given that ecologies might exist, even economies might arise in a *Life* universe. We can analyze economies into kinds of functional relations between objects within an ecological system, and functional relations are a combination of evolutionary and interactive properties. So, overall, requiring entailments from lower levels to higher levels in a *Life* world does not give us grounds to rule out many kinds of phenomena in it. Nevertheless, the skeptic holds that no pure *Life* world can entail the existence of consciousness or the specific character of its qualities.

The skeptic maintains that facts about bare difference are always consistent with the absence of experience, because qualitative contents are not merely structures of bare difference. If we consider that our taste space, for instance, contains different tastes and that our color space contains different colors, the relevant premise is that these tastes and colors are contents instantiating a structure of difference relations, not structures instantiated merely by difference relations.

Of course the skeptic knows that we can catalogue the differences between different colors and different tastes along relevant dimensions. If we do this, we can surely abstract out a content-free difference structure. The skeptic's objection is to the further move of analyzing conscious qualities into these abstract patterns of difference between them. Rather, our acquaintance with the phenomenal qualities yields information about them as contents occupying slots within these difference structures. Reification of the difference structure as basic ignores the grounding of those differences in each specific case and so ignores the content instantiating those structures. Given this observation that phenomenal qualities are not themselves bare, our question then reduces to whether or not individual qualitative contents such as the shades of green might be constituted by patterns of bare difference.

We can observe that a pattern of differences between colors can produce another color. For instance, a field of tightly packed yellow and red dots may yield an experience of phenomenal orange under the right viewing conditions. However, we can also observe that the shade of orange that results is not produced by the mere pattern of difference. It has to be a pattern of difference between the appropriate colors, thus providing no explanation of color in terms of *mere* patterns of difference.

If we try to abstract the patterns of difference from their contentful bases, viewing colors as mere difference structures, we see that the result is multiply realizable and that some of the realizations do not yield orange. For example, one can instantiate the same structure of differences between two other colors whose hues lie at the same distance from each other as red and yellow (e.g., yellow and green). A pattern of dots of these colors will yield a different color from orange. Therefore, we can observe an identical structure of formal difference but differ

ent colors. The example shows that, even allowing that we start with colors, one cannot reduce some colors to the mere difference structure among other colors.

The preceding observation is suggestive. After all, the skeptic is maintaining a much weaker position. The position the skeptic is defending is that patterns of *bare* differences do not entail the facts about the phenomenal qualities. Patterns of bare differences are difference structures whose identity obtains because of a mere formal difference, ungrounded by content at all. The skeptic notes that orange cannot even be reduced to the structure of difference between red and yellow once we allow substitutions for the phenomenal content of red and yellow. We can observe more straightforwardly that red and yellow are not constituted by patterns of mere difference, without any content at all.

The skeptic can even recruit Frank Jackson's argument about Mary, the superneuroscientist who spends most of her life trapped in a black-and-white room, to bolster this point. Most find it hard to deny that Mary learns something factual the first time she sees red (even if it is just a fact involving a new mode of presentation for an already known fact). By knowing all the physical facts, Mary certainly had all the information about the patterns of contrast and difference that are relevant to conscious sight. Yet these facts are not enough to yield, even in principle, whatever it is she learns on first seeing red. Whatever one thinks this implies about *physicalism*, it certainly implies something about *phenomenal redness*. It follows inevitably that whatever she learns about the experiencing of red is not just a fact about bare difference or patterns of bare difference. Because those are the only kinds of facts a pure *Life* world could entail, it follows that such a world could not entail the facts about conscious experience.

As an analysis of phenomenal content, the idea that something like a shade of red is a pattern of bare, merely formal differences conflicts with empirical observation. To make it work, something must be added. The only other tool *Life* presents for constructing phenomenal content out of patterns of bare difference is its counterfactual content. Unfortunately, known logics of counterfactuals add only such things as consistency constraints or metrics over similar possible worlds to our tool kit. These are not even the right kinds of things to add to a collection of formally distinct properties to make them add up to properties that are not merely formally distinct. A pattern of bare differences does not become a phenomenal content because another possible world contains a similar pattern or because it is consistent with patterns that occur elsewhere in that same world. Yet that is all we have here. If one tells a skeptic that a pattern of bare differences transitions to another pattern of bare differences, the skeptic can consistently deny that either pattern has to support experiencing. Nothing in the logic of counterfactuals requires that the *transition* should feel like something, either. The *Life* schema thus seems to underdetermine the story about qualitative content. We seem to have good reason for believing that the Skeptic's Claim is true.

Indexicality Indexical facts are facts specifying an honored place in space and time that counts as the center of a world or an honored object that counts as the

reference of terms such as "I" in that world. By being the center of the world, an indexed point or region of space and time provides a point of view from which we can understand the other facts in the world. For example, indexing a point or region as a *Life* world's center would provide the necessary point of view from which we could partition a world's history into past, future, and present; it allows us to partition its spatial coordinates into the place which is *here* and other places that are identified by their distances and directions from *here*; and it allows us to partition the world into physical information that is available at that place (because signals from other places may have traveled to it) and information that is not available (because signals have been lost or have not had time to reach it).

My argument in defense of the Skeptic's Claim has been run without appealing to indexical facts about potential *Life* worlds. Some people believe that facts about consciousness are essentially indexical, and so it would not be possible to derive facts about consciousness from any nonindexical base of facts. This might be true, although I think that is not clear, but it cannot be the fundamental problem with our analysis of the *Life* world. The kinds of information added by including indexical facts are either honorary (this is the *center* of the world), or are relational facts that follow straightforwardly from discursive knowledge of the honorary fact (*this* is true relative to the center), or are elusive in a way that seems quite different from how phenomenal qualities elude entailment by facts of bare difference (*this* moment is *now).* If others wish to insist that adding indexical facts to a pure *Life* world would turn a world of bare difference into a world able to entail facts about phenomenal qualities, then there is some gap between our understandings of indexicality and phenomenal quality that I do not know how to address. Perhaps they have some very substantial theory of indexicality that I cannot imagine and cannot answer, or perhaps their understanding of their phenomenal information is far less substantial than my understanding of mine.

Warning: We cannot transfer information between worlds When considering the skeptic's argument, we must resist beliefs about our world that might tempt us to smuggle phenomenal facts into the *Life* universe. For example, qualia in our world may perform some functions or correspond to some specific internal structures or processes. I want to emphasize that we must remain constantly aware that the *Life* universe is not our universe. We are to imagine an alien dimension, a dimension fully described by *Life's* physics. No one can decide the question of whether any conscious feeling at all can exist in a pure *Life* universe by an appeal to first-person evidence, analogy, or verbal reports. This takes out of play certain ground-level intuitions that affect the discussion about consciousness in our universe.

For instance, we cannot claim that we will empirically discover that different kinds of descriptions coordinate (Flanagan 1992) in a way that allows us to attribute an identity or determination relation between conscious feeling and the functionality of *Life* objects. Establishing such coordination would require us to access facts of both kinds, and the problem is precisely to access the phenomenal

facts in a pure *Life* world, if any. Conversely, if we had access to the phenomenal facts, if any, we would obviously not need any process of "coordination" between them and other kinds of facts. Those other kinds of facts would have been the entailment base from which we obtained our phenomenal information. So the pure *Life* universe is alien to us, and only entailment could bring consciousness into existence within it.

What phenomenology, if any, would cognitively functioning objects in the *Life* world require us to attribute to them? We do not have first-person knowledge that even one conscious state exists in such a world. Without some supporting story about how the facts in this alien world can be sufficient to support facts about consciousness, we cannot assume the existence of consciousness. And the supporting story must go beyond a coincidence of facts in our world if we want to generalize from our world to a hypothetical *Life* universe. Our alienation from the *Life* world blocks us from transferring the information so naively.

At this point, the existence of an explanatory gap in our world, admitted even by many physicalists, is evidence that mere coextensiveness, or "coordination," is all we really have. If this is so, then the functional information in the *Life* world by itself cannot be the whole story that we would need to attribute consciousness to *Life* objects. It follows that a skeptic is consistent if he admits to any kind of functioning at all in the *Life* world and denies that the activity supports consciousness.

2.6 From Life's Physics to Earth's Physics

If a pure Life world cannot entail facts about phenomenal consciousness, then a pure physical world cannot entail facts about phenomenal consciousness. Cellular automata such as *Life* very closely capture the essential character of our scientific concepts of the physical world and physical properties. In fact, it is not too difficult to imagine that our world might be a giant cellular automaton, albeit perhaps one with complicated stochastic causal-role properties. By using genetic algorithms to discover evolution rules, researchers at the Santa Fe institute have even discovered automata that produce particle-like elements capable of moving from cell to cell and interacting. An automaton can use these particles as information-bearing elements useful in solving problems encoded in its initial state (Das, Mitchell, and Crutchfield 1994). More recently, Wolfram (2002) has reported results of his twenty-year study of cellular automata, arguing from a tremendous amount of data that understanding our world in terms of cellular automata throws light on fundamental and unsolved problems in almost every branch of science, including fundamental physics.

Even if the concept of a cellular automaton does not perfectly capture our notion of the physical world, our concept of the physical is sufficiently close to that of cellular automata that it seems as if the same restrictions apply. They seem to be the same in the relevant respects. In particular, they share a common commitment to bare differences in their fundamental postulates. In *Life* we have *off* and *on* properties. In physics we have *spin*, *color*, *flavor*, *charge*, and *mass*.

The theoretical character of the basic properties is just the same in both cases: One stipulates at first that they are distinct and fleshes out their natures by designating laws that describe how they behave. The only real differences between *Life* and physics lie in such attributes as the complexity of the laws, the number and kind of dimensions the cells exist in, and perhaps nonlocal causation. In our world, the structure of the basic entities is more exotic. Instead of squares, we have particle waves and fields, maybe ten-dimensional strings with six of their dimensions rolled into an inscrutable knot, or other such exotica. Perhaps causation in the physical world requires infinite calculation, and so a Turing machine cannot simulate it.

At best, these differences add degrees of vagueness or complexity to the notions of structure, interactive property, and so forth that already are present in cellular automata but that do not seem to make a fundamental saving difference. The failure in the *Life* universe does not seem to arise from the facts that the basic objects were squares rather than strings or that the causal role properties were related simply and locally rather than complexly and nonlocally. Rather, the failure was rooted categorically in the stark geometric and bare counterfactual nature of the properties and of the world they made. Our experienced world is a world of felt tone: warm and warring, whirling and worrying, color and cadence. The pure *Life* world is a ghostly crystal home to phantasms.

So it seems that *any* phenomena that the physical facts entail must be analyzable into one of the basic classes of properties or some combination of them. Combinations of these classes support properties such as location, causal role properties, historical properties, structural properties, evolutionary properties, and properties of interaction. Such properties ultimately need nothing more than patterns of bare difference to exist. Again, the complexity of our universe introduces some vagueness into these general concepts, a vagueness that our specific concepts mirror; but vagueness hardly seems like the kind of thing that will allow us to escape the trap.

Some (e.g., Churchland and Churchland 1990) argue that antiphysicalist arguments using thought experiments are arguments from failure of imagination:

> The negative arguments here all exploit the very same theme, viz. our inability to imagine how any possible story about the objective nuts and bolts of neurons could ever explain the inarticulable subjective phenomena at issue.

It should be noted that the argument I have given does not have the form, "I cannot imagine how such-and-such could possibly explain consciousness, so such-and-such cannot explain consciousness." Its form is, "We have reasons for thinking that the physics of *Life* can only entail facts about bare difference and patterns of bare difference. We have observational evidence that the facts of consciousness are not facts of this type. Therefore, we have good reason to believe that a pure *Life* world cannot entail the existence of consciousness." Its essential form is an argument from insight. It argues from facts about a theory to that theory's failure of prediction, making a direct argument from what Chalmers

calls the "absence of analysis" and not an indirect argument from conceivability or new knowledge. To paraphrase Dennett, a perception of failure is not the same thing as a failure of perception. The successful perception, in this case, is of a failure of prediction. All that is then required to make the move against physicalism is pointing out that a pure physical world shares exactly the same damning characteristic that causes failure in the pure *Life* world.

<2>*A possible disanalogy: The intrinsic properties of the physical*</2> One disanalogy suggested by some people is that physical things have an intrinsic nature, whereas the entities and properties in the *Life* world are defined in a way that disregards such intrinsic facts. These facts about the intrinsic nature of the physical, some feel, may be responsible for the existence of phenomenal consciousness. If so, proponents of this objection argue, consciousness would be physical after all. Stoljar (2001) gives an interesting defense of this position.

I am sympathetic to this kind of view and defend something like it in part II. But I believe it fails to salvage physicalism, instead yielding a dual-aspect theory in which nature has both extrinsic physical and intrinsic phenomenal aspects. In short, a logical gap exists between (1) the observation that the proposed intrinsic properties are properties *of the physical* and (2) the conclusion that *they are physical properties*. After all, not all properties of biological things are biological properties, nor are all properties of economic things economic properties. I argue that physics really does specify only the relational network. Because physical theories are committed only to the existence of the facts they specify, the proposal that physical things have an intrinsic character implies that the commitments of our scientific physical ontology incompletely catalogue the world's properties. At best, physical theories may get at the intrinsic base only indirectly by needing it as part of a metaphysical, or at least extraphysical, framework.

One can take the view either that physical properties such as mass and spin are actually relational properties (e.g., if one believes no reference frames are privileged) or that they are intrinsic properties that physics specifies relationally (e.g., if one believes that the rest frame delivers the intrinsic value of mass). In either case, it is difficult to reconcile the view that physics describes intrinsic natures with the hard fact that general relativity is a fundamental physical theory.

Physics describes the outcomes of potential measurements. If measurements of mass were measurements of an intrinsic property, one would expect to find that an instance of it has the same value in all frames of reference. After all, intrinsic properties are the paradigm case of context-invariant properties, and the measurement of something intrinsic will not vary according to the frame of reference from which it is measured. But this is not what physical theory tells us the physical properties are like. According to general relativity, the very same instance of mass (for instance) has different values from different frames of reference. The same is true for some other physical attributes, such as velocity and shape.

If measurements of mass were instead measurements of the potential effective *differences* mass makes to processes in the frame of reference, finding out that

mass measures differently in different frames of reference poses no intellectual puzzles. Nothing is paradoxical about the idea that a dispositional property, mass, may have a different effective impact in different contexts. Even if one believes that "rest mass" provides a preferred intrinsic value for mass, it is most accurate to think of the measurements of mass in physics as the measurement of pure masslike differences that exist relative to frames of reference. This second, relativistic understanding of mass is the bare difference conception: Mass is first something distinct from the other named properties, and it is further differentiated as the property capable of having a certain kind of dynamical impact relative to other properties. It is these things and not any "intrinsic character" to which physics is committed. The methods of physics suggest that the measurement of all physical properties is essentially similar in intent and outcome, even if as a matter of fact some (like spin) happen to be invariant.

This feature of our physical concepts is very deep and fundamental, and it survives revolutions in theory. For example, in his wonderful book *Three Roads to Quantum Gravity* (2001), Lee Smolin gives exciting details about the progress being made on the successor theory to quantum mechanics and relativity. Here is how Smolin describes what our new version of the fundamental character of the physical world will probably be like, according to the latest breakthroughs:

> In the earlier chapters I argued that our world cannot be understood as a collection of independent entities living in a fixed, static background of space and time. Instead, it is a network of relationships the properties of every part of which are determined by its relationship to the other parts. In this chapter we have learned that the relations that make up the world are causal relations. This means that the world is not made of stuff, but of processes by which things happen….processes carrying little bits of information between events at which they interact, giving rise to new processes. They are much more like the elementary operations in a computer than the traditional picture of an eternal atom.

This new physics is directly a picture of bare difference (just as the current picture is). This difficulty with the relativity of physical predicates leads directly to another problem with the suggestion that the physical facts are facts about intrinsic properties. It is extraordinarily difficult to show that physical things *must* have intrinsic properties. All that our best physical theories describe is a network of effective dispositions, with each element typed according to its place in a network of relations to other such dispositions. This kind of purely relational world intuitively feels absurd to some, but arguments showing it to be incoherent do not seem at hand. In part II I produce plausibility arguments against the purely relational view. I believe the arguments there give strong reasons for preferring an alternative view, but they fall short of ruling out the relational view altogether. The claim that intrinsic properties must carry the dispositions described by physics ultimately must be added to theory as an intuitively justified metaphysical axiom. As such, it stands out as a primitive further fact relative to our scientific knowledge of the physical.

Finally, Stoljar (2001) distinguishes between the physicalism I have described here—which he calls *t-physicalism* (theory-based physicalism)—and an alternative that he calls *o-physicalism* (object-based physicalism). O-physicalism holds that a property is physical if, and only if, it is a "property required by a complete account of the intrinsic nature of paradigmatic physical objects and their constituents or else is a property which metaphysically (or logically) supervenes on the sort of property required by a complete account of the intrinsic nature of paradigmatic physical objects and their constituents."

Notice that even if one accepts the existence of some intrinsic character possessed by the basic physical entities, and even if one extends one's notion of physical fact to cover facts about intrinsic character, science is still left with a bootstrapping problem. The bootstrapping problem concerns how to get from (1) the claim that electrons, photons, and quantum chromodynamic quarks have an intrinsic character to (2) the conclusion that there could exist a human consciousness with further intrinsic character all its own to (3) an explanation of why this intrinsic character associated with a middle-level object should be experiential in nature to (4) an explanation of why this experiential context has various features that it seems to have, such as its peculiar unity and coherence with information processing characteristics of the brain. Why shouldn't intrinsic character be delimited right at the boundaries of the microphysical, with the rest of nature, including us, being mere abstraction off the patterns of their interaction? And why shouldn't intrinsic character be *merely* intrinsic, a kind of categorical nature not experienced at all? Even if one agrees that nature needs an intrinsic basis, it looks as if nature ordered an ordinary Volkswagen and God delivered a top-of-the-line Mercedes, which is very odd.

The proposal that o-physicalism really is a kind of physicalism can sound deceptively reasonable at first glance, but the devil is in the details. An o-physicalist could view this book as an attempt to present the details needed to make an o-physicalist view really work. For now, I merely want to point out that we have no reason to believe that we can solve the challenge of placing consciousness without appealing to some kind of new facts about the world, over and above those that science recognizes as the physical facts. Thus I use *physicalism* to mean t-physicalism.

To an extent, whether o-physicalism also deserves the name is a disagreement about labeling. Knowing the end of the story, I believe the o-physicalist path takes us so far beyond what physicalists have traditionally seemed to mean by *physicalism* that it is unreasonable to think that the view that results is physicalist. For readers who consider themselves o-physicalists, I recommend absorbing the full story and then deciding. Even if the specific development of Liberal Naturalism in this book is not ultimately accepted, it is in the same family of theories as those an o-physicalist will have to develop and eventually endorse. Therefore, it helps to make clear the magnitude of the departure from a physics-based physicalism required by such a view.

2.7 Summary

The failure of the *Life* world to entail facts about consciousness seems to be a result of the fact that the phenomenal aspect of experience is not the sort of thing that is entailed by the existence of patterns of bare difference. In particular, it possesses a qualitative content whose existence is not entailed by facts of a functional, structural, or evolutionary sort. Given that, the failure should hold in our world for basically the same reasons it holds in *Life*. Indeed, in many ways the *Life* example gives us a better proving ground for making the determination, because, unlike in our world, we do not start with the knowledge that

consciousness exists in that universe. Therefore, the temptation to see entailments where they do not exist is greatly lessened.

<FN>[i] "A priori" is a philosophical term for a conclusion that can be justified independently of an appeal to historical or scientific facts. "A posteriori" is the complementary term for conclusions that can be justified only by appealing to such facts.

[ii] A universal Turing machine is a kind of computer that can simulate any other computer.

[iii] Note that this entailment obviously holds even though gliders have not been defined in terms of *Life's* physics.

[iv] More exact analyses of "conceptual," "empirical," and "interpretive" are given in sections 2.7 and 3.2.

[v]The relevant premise for the argument is (P): If **x** has the status of being an observable, then the evidence for **x** must also have the status of being observable. For example, if I can observe that it is cold outside based on the evidence that there is snow on the ground, the snow on the ground must also be something that I can observe. Similarly, if I observe a meson in the cloud chamber based on the evidence that a cloud track has appeared, then the cloud track must be an observable also. The premise (P) gains its plausibility from the principle that the epistemic status of evidence cannot be less secure than the status of that which it is evidence for.</FN>