A-LEVEL PHILOSOPHY TEACHING AND LEARNING GUIDE: THE METAPHYSICS OF MIND

Amy Kind
Claremont McKenna College

August 2021

Table of Contents

1. Introduction

What is the nature of mind, and how is it related to the brain and to the body? Can we account for the mind within a scientific and naturalistic framework, or does it require a different kind of treatment? How can we know of the existence of minds other than our own? What kinds of entities have minds? For example, would it be possible for machines to have minds? How would we know? All of these questions are at the heart of the philosophical subfield known as Philosophy of Mind.

Though some of these questions raise epistemological worries, in this guide we will focus primarily on those concerning the metaphysics of mind, a set of questions often referred to as the *Mind-Body Problem*. Traditionally, answers to the Mind-Body problem have come in two broad varieties. First there are dualist theories. On this kind of view, the world is thought to consist of two fundamentally different kinds of things, minds and bodies. Second, there are monist theories. On this kind of view, the world is thought to consist of only one fundamental kind of thing. One version of monism, *idealism*, claims that everything that exists is fundamentally mental in nature. Though idealism has had some

proponents throughout the history of philosophy, perhaps most notably the 18th century philosopher George Berkeley, it no longer receives much attention from philosophers of mind and we will not discuss it here. A second version of monism, *materialism*, claims that everything that exists is fundamentally material in nature, i.e., that everything that exists is made of matter. Starting in the 20th century, given that science recognizes all sorts of entities (e.g., forces such as gravity) that are not made of matter, philosophers realized that materialism as traditionally construed was too narrow. In contemporary philosophy, the view has now been reconstrued in terms of the claim that everything that exists is fundamentally physical in nature, and the view is now typically referred to as *physicalism*.

Both dualism and phyiscalism come in several varieties, many of which will be discussed in this guide. We will also discuss a third kind of view, *functionalism*, that became popular in the mid- to late- 20th century. Inspired by developments in computer science and artificial intelligence, functionalism thinks of the relationship between mind and brain similarly to the relationship between software and hardware. On the functionalist view, we can think of the mind as a kind of program being instantiated by the brain. For the functionalist, the mental is best understood not in terms of its physical nature but rather in terms of its functional nature. In considering all three of these views – dualism, physicalism, and functionalism – we will take up both the considerations that have been adduced in support of them and the problems and objections that these views face.

Before we turn to the consideration of these views, however, it will be useful to start with a more general discussion of what is meant by mind.

2. What Do We Mean by Mind?

Consider an Olympic diver who is about to make her final dive in the final round of the women's 3m springboard event. She positions herself on the springboard facing away from the water. Her toes touch the board but her heels do not. Her arms are raised above her head. Her pulse rate is slightly elevated. She is sweating, and she can feel a bead of sweat dripping down her left temple. Her left shoulder aches from an earlier injury. She can hear the buzz of the crowd even though, given how she is facing, she cannot see them. She wants to medal in this event, and she believes that she is in contention. She feels exhilarated, confident, and calm.

These descriptions capture just a handful of the many aspects of the diver's current situation. Some of the descriptions – like those concerning her pulse rate, the position of her arms and toes, and her production of perspiration – pick out her bodily states. But some of the descriptions – like those concerning her sensations, her perceptions, her beliefs, her desires, and her emotions – pick out her mental states.

Let's focus in on some of these mental states. First, consider her desire to medal in this event. Philosophers refer to this state as a *propositional attitude* – the diver has a certain attitude (the "desiring" attitude) toward a certain proposition ("I win a medal in this event.") She might take this same attitude towards other propositions, e.g., "I set a personal best in this event," "I get an endorsement deal with Speedo," "I return to compete at the next Olympic games." And she might also

take different attitudes towards this same proposition – she might believe she will win a medal at this event, hope that she will win a medal at this event, or imagine that she win a medal at this event.

When discussing propositional attitudes, philosopher often classify them based on what's called *direction-of-fit.* The distinction between different directions-of-fit was initially introduced by G.E.M. Anscombe in her influential book *Intention*. Anscombe's example involved two different ways that a list of grocery items might function. A shopper with a grocery list uses the list to determine what items to put in their shopping cart. They want the contents of the cart to track the list. But suppose that a private detective is following the shopper and wants to keep track of what they are buying. Every time the shopper puts something in the cart, the detective adds that item to the list. The detective thus wants the contents of the list to track the contents of the cart. States like belief are like the detective's list – we aim to have our beliefs track (or fit) the world. This is what's called *mind-to-world* direction of fit. In contrast, states like desires are like the shopper's list – we aim to have the world track (or fit) our desires. This is what's called *world-to-mind* direction of fit.

Propositional attitudes are paradigm cases of states with what philosophers call *intentional properties* or *intentionality*. Importantly, the sense of intentionality here invoked has nothing to do with intention and purposefulness. Rather, when a state has intentionality in the sense here in question, it has aboutness, i.e., it is about or directed at something. When the diver is thinking about her family, or Tokyo, or the COVID-19 pandemic, her thoughts are directed at (or about) a particular group of people, a particular place, or a particular global state of affairs.

Of course, not all of our thoughts concern presently existing things. Some of them concern past things (as when the diver might remember her first dive at this Olympics) and some of them concern future things (as when she imagines her first dive at the next Olympics). Sometimes we might also think about things or people or places that don't exist. The diver might just as well be thinking about the lost city of Atlantis as about Tokyo. As Charles Siewert has helpfully put it in his *Stanford Encyclopedia* entry on consciousness and intentionality, "Thoughts, unlike roads, can direct you to a city that is not there."

Do all of our mental states have intentionality? Some philosophers, perhaps most notably the 19th century philosopher Franz Brentano, have claimed that they do. In *Psychology From an Empirical Standpoint*, Brentano claimed that all and only mental states have intentionality, a point that is often put by saying that intentionality is the mark of the mental. But other philosophers disagree. As potential counterexamples, they have often pointed to states like pains and itches. Perhaps, however, these states can be understood to have intentionality insofar as they can be understood as being in some sense about or directed at certain conditions or disturbances of the body. Another potential counterexample stems from the consideration of moods. Though emotions are typically directed at particular states of affairs (the diver is happy about winning a medal or sad about not setting a personal best), moods do not seem to have to be directed at anything at all. Sometimes someone might be a state of free-floating anxiety or pure euphoria that is not really about anything at all.

Even if we don't accept that intentionality is the mark of the mental, however, there's no doubt that intentionality is a central feature of many of our mental states. Another central feature of our mental states concerns what's often referred to as their *what-it's-likeness*. There is something it is like to

experience the sharp pain of stepping on a stray Lego piece, to smell the cookies baking in the oven or to taste a sour pickle, or to see the rich hues of a beautiful sunset. The what-it's-likeness of a mental state concerns its *phenomenal character* or *qualitative character*, and philosophers often refer to the relevant phenomenal properties as *qualia* (singular: *quale*).

Just as we could ask whether all mental states have intentional properties, we might also ask whether all mental states have qualia. Here again we see disagreement among philosophers. Many philosophers argue that though there is nothing that it is like to have a particular belief. What is like to believe that the White House is in Washington, D.C., for example, or that dogs are mammals? The claim that beliefs lack qualia seems especially plausible when we consider that beliefs need not be occurrent before the mind. Right now, before you finished reading this sentence, you presumably have the belief that 2+2=4, but it was not occurrent before your mind. How, then, could there be anything it is like to have that belief? It is hard to see how this non-occurrent belief could have any phenomenal character. More controversial is whether there is phenomenal character associated with occurrent belief. When the diver stands on the springboard and deliberately thinks to herself, "I can do this," is there something that it is like for her to have this belief? The belief may be accompanied by mental imagery, which has phenomenal features, and while she is thinking the thought she may have a phenomenal experience associated with hearing the words run through her mind. But in addressing the question before us, we need to set all this aside and focus on the having the belief itself, or thinking the thought itself, and consider whether there are any proprietary phenomenal properties to be found. Philosophers who do think that there is something it is like to have an occurrent belief or to think a thought refer to the relevant proprietary phenomenal features as *cognitive phenomenology*, and they claim this kind of qualia is another kind of *sui generis* phenomenological feature just like emotional qualia, pain qualia, color qualia, and so on.

Traditionally, qualia have been treated as intrinsic properties of experience and, relatedly, as non-intentional. They have also been thought to be directly or immediately accessible by introspection. It's in large part in virtue of these features that qualia have proved particularly difficult to accommodate within a physicalist or functionalist conception of mind. We will return to issues concerning qualia in considerably more detail when we consider physicalism and functionalism in Sections 4 and 5 below.

3. Dualist Theories

Dualism comes in two main varieties. According to substance dualism, there are fundamentally two different kinds of substances in existence – physical substances like bodies and brains, and nonphysical substances like minds. Put simply, the substance dualist claims that each of us has an immaterial/non-physical mind in addition to, or over and above, our physical brain. Property dualists deny this claim. But even though they accept that the mind can be identified with the brain, they claim that the brain has some mental properties that cannot be identified with or reduced to its physical properties. On their view, though there are not fundamentally two different kinds of substance in existence, there are fundamentally two different kinds of properties in existence. In this section, we will first discuss each of these views in detail. We then turn to issues facing the dualist view. Though some of these issues relate only to a particular version of dualism, some of them apply to all dualist views across the board.

3.1 Substance Dualism

Perhaps the most influential development and defense of substance dualism can be found in the work of 17th century philosopher René Descartes, and most specifically, in his *Meditations on First Philosophy*. Descartes' goal in the *Meditations* is to determine what can be known with certainty, and to achieve this goal, he begins by trying to call all of his prior beliefs into doubt so that he can make a fresh start and build his philosophical theory on a firm foundation. In the first meditation, Descartes supposes that he is the victim of an Evil Demon who has tricked him into believing that the world around him exists. As a result, he comes to doubt almost everything – as he puts it: "I shall think that the sky, the air, the earth, colours, shapes, sounds and all external things are merely the delusions of dreams that he has devised to ensnare my judgement. I shall consider myself as not having hands or eyes, or flesh, or blood or sense, but as falsely believing that I have all these things." But in the second meditation Descartes soon finds a way to climb out of the deep skeptical hole that he'd dug for himself, for he realizes that he can be certain of at least one thing, namely, that he exists. His reasoning for this conclusion is encapsulated in the famous phrase: I think, therefore I am (*Cogito ergo sum*). In the very process of doubting his existence, he is engaging in thought, and in order to be thinking, he must exist. From this kind of reasoning, Descartes is led to his dualist view. Though he can successfully doubt the existence of the body, he cannot successfully doubt the existence of the mind, and so he believes that they must be different things.

By the end of the second meditation, Descartes concludes that he can be certain that the mind exists. Over the course of the remaining meditations he also concludes that he can be certain that God exists and, since God cannot be a deceiver, that material objects exist. He returns to issues concerning the mind and the body in the sixth and final meditation. In this meditation, Descartes not only offers arguments in support of the claim that the mind and body are fundamentally different kinds of substances but he also explains the relationship that he sees holding between them. In fact Descartes sees this relationship between mind and body in terms of casual interaction. The specific version of substance dualism that Descartes offers is thus known as *interactionist dualism*, or *interactionism.* On Descartes' interactionism, there is a two-way causal relationship between the mind and the body. Events in the mind can cause events in the body, as when my desire to get into shape causes me to get up and exercise. Likewise, events in the body can cause events in the mind, as when my elevated pulse rate after only a brief bout of exercise causes me to believe that I am in worse shape than I had thought.

Let's turn to the two arguments for dualism presented in the sixth meditation. At times, Descartes relies on the existence of God to defend his points; after all, he believes himself to have established this point in the previous meditations. Importantly, however, the basic lines of reasoning need not rely on this claim, and insofar as it is good philosophical practice to avoid relying on controversial premises when one need not do so, we will consider secularized versions of the arguments that Descartes gives.

The first of these arguments is what's often referred to as the *argument from indivisibility*. Noting that he cannot discern any parts to the mind, Descartes suggests that the mind is something "quite single and complete." But since it is part of the essential nature of physical things to be divisible, this shows that the mind cannot be a physical thing. We can represent the argument in standard form as follows:

Premise 1. The mind is indivisible.

Premise 2. Everything that is physical is divisible.

Conclusion. Thus, the mind cannot be a physical thing.

Someone might try to object to premise 1 by pointing to various faculties of the mind such as the faculty of the will, the faculty of understanding, and the faculties of sense perception. But though Descartes accepts the existence of these different faculties, he denies that they can be thought of as separate parts of the mind the way we can think of arms and legs as separate parts of the body. Though an arm could continue to exist independently of the body if we were to separate the two, the faculty of the will or the faculty of the understanding can't continue to exist independently of the mind if we were to try to extract them. In fact, talk of extraction in this context doesn't even really make sense. As Descartes notes, it is *the whole mind* that wills, and likewise for understanding and perceptions. Moreover, if we consider any particular idea or thought or perception that we have, we'll see that each of these is really an idea or thought or perception of the whole mind, not of a part of it. Contemporary philosophers often express this basic point by saying that consciousness is *unified*.

That said, even if we grant Descartes these points, it is not clear that this argument is as successful as he might have thought. Each of the premises can be called into question. Against premise 1, we might note that there are other possibilities for how the mind might be divided other than in terms of its faculties. In the tradition of Freud, for example, we might divide the mind into the conscious mind and the unconscious mind. Alternatively, consider a condition like dissociative identity disorder (DID). When someone has DID, they experience discontinuity with respect to their sense of self and agency, and this discontinuity typically manifests itself by way of two or more distinct personality states. These different personality states called *alters* tend to have executive control of the body at different times, and one alter may not have any memories of what's happened when they are not in charge of the body. In cases where the alters have separate streams of consciousness, we might plausibly consider each alter to be a distinct part of the given individual's mind.

Against premise 2, we might note that there are indeed physical things that are incapable of being divided. Objects at the macro level can be divided into smaller and smaller parts, and ultimately into atoms and even sub-atomic particles. But once we get all the way down to particles such as quarks and gluons, we seem to have reached units that cannot be further divided into smaller components.

The second argument that Descartes offers is referred to as *the conceivability argument*. Because Descartes believes that he can conceive of the mind and the body as existing apart from one another, he draws the conclusion that they are actually distinct entities. In standard form, and omitting Descartes' reliance on God in line with the practice we outlined above, we can capture his reasoning as follows:

Premise 1. Whatever is clearly and distinctly conceivable is possible.

Premise 2. I can clearly and distinctly conceive the mind existing without the body.

Sub-conclusion. Thus, it is possible for the mind to exist without the body.

Premise 3. If it is possible for A to exist without B, then A and B are distinct entities.

Conclusion. Thus, the mind and the body are distinct entities, i.e., dualism is true.

The argument proceeds in two distinct stages. In the first stage, Descartes moves from a claim about conceivability to a claim about possibility. In the second stage, Descartes moves from the claim about possibility to a claim about how things are in actuality. As we will see, the kind of reasoning found in each of these stages parallels various kinds of reasoning we do in ordinary life.

Let's first consider stage 1. To understand the reasoning that's going on here, we first need to understand the notion of possibility here in play. Philosophers often distinguish between various different senses of possibility. One distinction often drawn is between *physical possibility* and *metaphysical possibility*. When something is physically possible, it is possible given the laws of physics. For example, it is physically possible for humans to fly when they are wearing jet packs. Moreover, while the jet packs currently in existence max out at about 10 minutes of flight time, longer flight duration is physically possible – the limits come not from our physical laws but from the state of our current technology. Absent jetpacks, however, humans cannot fly. It is not physically possible for humans to fly unaided.

When it comes to metaphysical possibility, we are not restricted by the physical laws. In a universe governed by different physical laws, humans might be able to fly unaided. Unaided flight, though not physically possible, is metaphysically possible. Importantly, not every scenario that we can describe is metaphysically possible. The world might have had different physical laws, and thus might have been different in all sorts of ways from the way that it in fact is. But even in a world that had different physical laws, triangles would still be three-sided. It is not metaphysically possible for a triangle to have four sides. It is not metaphysically possible for an object to be wholly blue and wholly red at one and the same time.

The sense of possibility employed by the argument from conceivability is metaphysical possibility. When Descartes offers the principle stated in premise 1 that whatever can be clearly and distinctly conceived is possible, he means that whatever can be clearly and distinctly conceived is metaphysically possible. It might not be possible given the physical laws of our world, but if we can clearly and distinctly conceive of a given scenario, then that scenario presents a way that our world could have been.

Why should we think this principle is at all plausible? Suppose that a detective is trying to determine whether a given individual could have committed a murder crime under investigation. Given what's known about the distances involved, the detective conceives of a route the suspect might have taken that would have gotten them to the crime scene on time and enabled them to avoid being seen by anyone else. On the basis of this act of conceiving – what we might think of as a *thought experiment* – the detective concludes that it is possible for the suspect to be the murderer. This act of reasoning employed by the detective parallels the kind of reasoning employed in stage 1 of Descartes' argument.

But though the detective's reasoning seems solid up to this point, it does not seem at all reasonable for them to continue on to conclude, solely on this basis, that the suspect actually committed the crime. So it might seem that Descartes' reasoning in stage 2 must be problematic. But there's an important disanalogy between the detective's reasoning structure and Descartes' reasoning structure. While the detective would be moving from a claim of the form *possibly p* to a claim of the form *actually p*, Descartes is moving from a claim of the form *possibly p* to a claim of the form *actually q*. Though Descartes believes that the mind and the body could exist apart from another, he does not conclude that they actually exist apart from one another. He accepts that in actual life they always exist together. He concludes something else, namely, that they are distinct entities. Though it seems implausible to take the fact that two things could exist apart to show that they do so in actuality, it's considerably more plausible to take this fact to show that they are distinct kinds of things; in a sense, what it means to be distinct is to be possibly separable. Consider an analogy. From the fact that it's possible for a glass vase to shatter, we shouldn't conclude that it's actually shattered. But from the fact that it's possible for the vase to shatter, we can conclude that it's fragile.

The discussion thus far helps us to see why Descartes' reasoning has some prima facie plausibility. It should not be simply dismissed out of hand. But should it be accepted? Let's consider three different objections that have often been raised to it.

First, one might deny Descartes' first premise, that is, one might deny that what's conceivable is actually possible. A version of this objection was raised directly to Descartes by Antoine Arnauld. Arnauld pursues the objection by way of a geometric analogy. Someone who is not well acquainted with geometric principles may conceive of a right triangle for which the Pythagorean theorem does not hold. But they would be mistaken to conclude that it is possible for there to be a right triangle for which the Pythagrean theorem does not hold. The Pythagorean theorem holds of all right triangles, that is, in every right triangle, the sum of the squares of the two shorter sides is equal to the square of the hypotenuse.

In response, Descartes notes that premise 1 requires not just that we be able to conceive of something but that we be able to conceive of this *clearly and distinctly*. It's only when we conceive of things clearly and distinctly that we can draw conclusions about possibility. In the triangle case, the conceiving is in some way muddled and confused, and because of this, premise 1 does not apply.

Though this response might provide some protection from Arnauld's worry as we have so far construed it, it is easy to recast the worry in a slightly different way. Doing so gives us the second objection that we'll consider, an objection that targets premise 2. To avoid the counterexamples to premise 1 like the geometric case, we need to put a very strict standard on what counts as conceiving something clearly and distinctly. But once we impose this strict standard, we can no longer have any confidence that it's been met in the mind-body case. Perhaps our thinking in this case is also muddled and confused. For example, one reason that we might be inclined to think that it's conceivable for the mind to exist apart from the body is that we think that humans can exist in a disembodied form. There have been numerous works of fiction that present us with this very scenario, some of which even describe it in

some detail. But now ask yourself this: When you conceive of yourself disembodied, do you conceive of yourself being able to see and hear things around you? How do you do this, if you are completely disembodied? After all, if you had no body then you wouldn't have any eyes or ears. This might seem to suggest that our thinking about disembodied existence is indeed muddled in various ways. According to this objection, then, mind without body is not really conceivable, or at least, not really conceivable in the strict sense needed to support judgments of possibility.

A third objection that has been raised to the conceivability argument concerns the last stage of the argument. One might worry that conclusions about metaphysical possibility don't really tell us anything about the actual world. To develop the worry, one might argue as follows. The fact that there could have been a world in which humans were able to fly unaided doesn't tell us anything about the actual capabilities of actual humans in the actual world. It's still impossible for *us* to fly unaided. Likewise, then, the fact that there could have been a world in which minds were able to exist without bodies doesn't tell us anything about actual minds in the actual world. It's still impossible for *our minds* to exist without a body. Whether this kind of objection can be successful will depend on how exactly we should understand the dualism/physicalism divide, e.g., on whether physicalism must hold in all metaphysically possible worlds in order for it to be true.

3.2 Property Dualism

From the vantage point of the 20th century, many people (including many philosophers) think that substance dualism can be dismissed more or less out of hand. It strikes many as outdated and nonscientific, even supernatural. Accepting the existence of a nonphysical mind seems to commit one to the postulation of "spooky stuff" and tantamount to accepting the existence of ghosts.

In contrast to substance dualism, however, property dualism does not claim that the mind is a separate substance from the body and brain. The property dualist accepts that all substances that exist are physical substances. For this reason, this version of dualism strikes many as considerably more compatible with a naturalistic worldview. In fact, many property dualists explicitly align themselves with naturalism. David Chalmers, for example, refers to the property dualistic view that he offers as *naturalistic dualism.*

On property dualism, the dualism comes in not at the level of substances but at the level of properties. What makes the view dualistic is that some physical substances – in particular, brains – are seen as having mental properties over and above their physical properties. Just as the brain has the property of weighing a certain number of grams, having a certain volume, and being composed of a certain number of neurons, all of which are physical properties, it also has the property of having certain desires, beliefs, and sensations, all of which are non-physical mental properties.

To understand property dualism, it will be helpful to understand two technical notions: *reduction* and *supervenience*. Let's take these in turn. In an effort to achieve greater overall understanding, as well as theoretical coherence and unification, scientific discovery has often led to the reduction of one theory to another by mapping the former's laws and terms onto the latter's laws and terms. For example, when scientists developed statistical mechanics, they were able to map the classical gas laws onto this

new, more explanatory theory. The property *temperature of a gas* was shown to be reducible to – i.e., nothing more than – the property of *mean molecular kinetic energy*. As we will see in Section 4, many physicalists think that mental properties can similarly be reduced to physical properties, namely, brain properties. In saying that mental properties are "over and above" the physical properties of the brain, the property dualist denies this reduction can be achieved, i.e., the property dualist claims that mental properties are *irreducible*.

But some physicalists make a weaker claim than that of reduction. Instead of claiming that mental properties reduce to physical properties, they instead claim that mental properties supervene on physical properties. When a set of properties *S1* supervene on the set of properties *S2*, there can be no change in *S1* without a corresponding change in *S2*. Consider, for example, an aesthetic property like beauty. Suppose there are two paintings that are identical in all their physical properties – they are the same size and shape, they each have exactly the same color and quantity of paint at exactly the same spots, and so on. Given these facts, it seems plausible that these paintings cannot differ in terms of their beauty. If the first painting is beautiful, then so is the second painting, and vice versa. Beauty supervenes on physical composition. But now consider value. Can these two paintings that are physically identical differ in value, even though they can't differ in terms of their beauty? Here the answer seems to be yes. The first might have been painted by Rembrandt, for example, while the second was painted by a mere apprentice who was copying Rembrandt's work, and these facts seem to make the first painting more valuable – even if not more beautiful – than the second. Since there can be a difference in value even without a difference in physical composition, value does not supervene on physical composition. For value, origin also matters. The property dualist takes mental properties to be more like value than like beauty. For the property dualist, the mental does not supervene on the physical.

With this understanding of property dualism before us, in the remainder of this section, we will consider two influential arguments for property dualism that were developed in the late 20th century: *the zombie argument* and *the knowledge argument*.

Importantly, the kind of zombie that features in the zombie argument is a very kind of creature from the zombies of horror films. While Hollywood zombies are dead, philosophical zombies are very much alive. Moreover, they are physically identical to human beings – my zombie twin has exactly the same physical structure that I do, all the way down to the microphysical level, and she correspondingly behaves exactly as I do. She is physically, functionally, and behaviorally identical to me. What distinguishes my zombie twin from me, and, more generally, what distinguishes zombies from ordinary humans, is that a zombie altogether lacks phenomenal consciousness. When I look at a peach and take a bite, and when my zombie twin looks at a peach and takes a bite, we both react the same way – smiling, saying "Wow, this peach is good," and taking another bite. But while I am also having various phenomenal experiences – a certain orangey-yellowy experience when I look at it, a tangy sweet experience when I bite into it, a pang of happiness at how delicious it is – my zombie twin has none. That's not to say that she has no mental life. She does have various psychological states, like beliefs and desires. She will also have analog states to my pains and emotions, states that play the role of my pains and emotions but without

the phenomenal character of my states. She looks exactly like me and acts exactly like me, but for her, all is dark inside.

With this notion of a zombie before us, we can now consider the zombie argument itself. Though the concept of a zombie first appeared in philosophical discussion in the 1970s, the zombie argument rose to philosophical prominence through the work of David Chalmers, who develops it at length in his book *The Conscious Mind*. Like Descartes' conceivability argument, the zombie argument also relies on claims about what is conceivable to motivate its conclusion about the metaphysics of mind.

Premise 1.	I can conceive of zombies, i.e., creatures that are microphysically identical to conscious beings but that lack consciousness entirely.
Premise 2.	If zombies are conceivable then they are metaphysically possible.
Sub-conclusion.	Therefore, zombies are metaphysically possible.
Premise 3.	If zombies are metaphysically possible, then consciousness is non-physical.
Conclusion.	Therefore, consciousness is non-physical.

Sometimes the argument is stated in terms of a zombie world. On this version of the argument, it is claimed that one can imagine an entire zombie world, i.e., a world that is physically identical to ours but in which there is no consciousness.

Given the structural similarity between the zombie argument and the conceivability argument, it's probably unsurprising that similar objections would arise here. When confronted with this argument, many physicalists deny that philosophical zombies are really conceivable. There are various different ways this objection might be pursued. For example, some philosophers argue that someone who thinks they have conceived of their zombie twin is simply fooling themselves. It's likely that what they've actually conceived is something that comes close to being their zombie twin but falls short in some way, perhaps a creature who is very, very physically similar to them but who is not identical to them. Daniel Dennett has forcefully pursued this line of argument:

> Supposing that by an act of stipulative imagination you can remove consciousness while leaving all cognitive systems intact – a quite standard but entirely bogus feat of imagination – is like supposing that by an act of stipulative imagination, you can remove health while leaving all bodily functions and powers intact. If you think you can imagine this, it's only because you are confusedly imagining some health-module that might or might not be present in a body. Health isn't that sort of thing, and neither is consciousness.

Alternatively, other philosophers have argued that conceivings involving assertions about phenomenal consciousness must proceed from an internal perspective. To genuinely conceive of someone who is having a phenomenal experience of happiness, for example, I must imagine the happiness from the inside; that's the only way to really build in the happiness qualia to the conceiving. Likewise, to genuinely conceive of a creature who lacks phenomenal consciousness, I must again imagine this from

the inside. But, it is charged, this is impossible. Given that we have phenomenal consciousness, we cannot imagine phenomenal emptiness from the inside.

A second kind of objection to the zombie argument proceeds by denying that facts about conceivability entail facts about metaphysical possibility. While these discussions can get rather technical, the basic idea arises from the fact that conceivability is primarily an epistemic notion, whereas the relevant kind of possibility in question in the zombie debate is a metaphysical notion. Consider the fact that water is identical to H_2O. As Saul Kripke famously argued in *Naming and Necessity*, claims like this are *a posteriori* necessities. Though this identity claim is necessarily true, it took scientific investigation to determine its truth. It cannot be known *a priori*, independently of experience, but can only be known *a posteriori*. Thus, though we can conceive of water having a different chemical structure, like being composed of XYZ, this conceiving cannot be taken to show that it is metaphysically possible for water to have a different chemical structure.

A third kind of objection to the zombie argument, and one we considered briefly in connection with the conceivability argument, charges that conclusions about metaphysical possibility do not tell us anything about how things are in the actual world. This line of objection is often connected with what's called *the phenomenal concept strategy*. Take a concept like *pain*. This concept is phenomenal; it picks out a certain property by way of its phenomenal features. As such, phenomenal concepts are recognitional. They are what enable us to recognize certain properties. But that does not mean that the properties so recognized are not physical properties.

In the following passage Peter Carruthers and Bénédicte Veillet provide a nice summary of phenomenal concepts and the way that physicalists put them to use in responding to the zombie argument:

> What is said to be distinctive of such concepts is that they are *conceptually isolated* from any other concepts that we possess, lacking any a priori connections with non-phenomenal concepts of any type (and in particular, lacking such connections with any physical, functional, or intentional concepts). Given that phenomenal concepts are isolated, the physicalist argues, then it won't be the least bit surprising that we can conceive of zombies and inverts, or that there should be gaps in explanation. This is because no matter how much information one is given in physical, functional, or intentional terms, it will always be possible for us intelligibly to think, "Still, all that might be true, and still *this* [phenomenal feel] might be absent or different." There is no need, then, to jump to the anti-physicalist conclusion. All of the arguments referred to above are perfectly compatible with physicalist accounts of phenomenal feelings.

The second argument for property dualism that we will consider, the knowledge argument, was developed by Frank Jackson in his 1982 paper, "Epiphenomenal Qualia." At the center of this argument is a thought experiment known as *the Mary case*. Mary is a brilliant color scientist who exists at some time in the future when color science has been completed. Unfortunately, Mary has spent her entire life confined in a black and white room. Everything in the room is in black and white, and Mary has never seen anything in color. (To imagine this, suppose that she wears black and white gloves, that the room contains no mirrored surfaces, etc.) While in her room, she has learned the entirety of color science – including all of the relevant neuroscience – via books and black-and-white videos. She knows

everything there is to know about color and color vision. Now suppose that one day she is released from her room and shown a ripe tomato for the first time. When she experiences the sensation of red for the first time, what happens? According to Jackson, and most people hearing the thought experiment seem to agree, Mary learns something new. She has what we might think of as an "Aha!" moment: "Aha," she might say to herself, "So that's what seeing red is like!"

Jackson then puts this thought experiment to use in showing that physicalism must be false and that we must accept non-physical properties over and above the physical properties. His argument goes as follows:

Premise 1.	While in the room, Mary has acquired all the physical facts there are about color sensations, including the sensation of seeing red.
Premise 2.	When Mary exits the room and sees a ripe red tomato, she learns a new fact about the sensation of seeing red, namely its subjective character.
Sub-conclusion 1.	Therefore, there are non-physical facts about color sensations.
Premise 3.	If there are non-physical facts about color sensations, then color sensations are non-physical events.
Sub-conclusion 2.	Therefore, color sensations are non-physical events.
Premise 4.	If color sensations are non-physical events, then physicalism is false and property dualism is true.
Conclusion.	Therefore, physicalism is false and property dualism is true.

In response to this argument, physicalists have offered various responses. In line with his response to the zombie argument, Daniel Dennett suggests that we should reject premise 2. In his view, we are only inclined to think that Mary has an "Aha!" moment upon seeing the tomato because we have failed to really imagine what we've been asked to imagine. Though we are supposed to imagine that Mary knows *all* the physical facts about color and color vision, we find this so hard to imagine that we instead simply imagine that she knows lots and lots of the facts. According to Dennett, if Mary really knew the entirety of the physical story about color and color vision, then the ending of Jackson's thought experiment would go quite differently. Instead of saying something like "Aha!" when she sees the tomato, Mary would instead nod to herself in satisfaction and indicate that what has happened was exactly as she had predicted. Using her vast knowledge base, she was able to determine exactly what experience she would have when shown a tomato. If you had tried to trick her by showing her a blue tomato instead, she would not have been fooled.

In contrast to Dennett, however, most physicalists share Jackson's intuition that Mary learns something when she leaves her black and white room and has color experiences for the first time. But while they accept that Mary has an "Aha!" moment, they deny that this commits us to the falsity of physicalism. Their responses typically take three forms: *the ability hypothesis*, *the acquaintance hypothesis*, and *the new knowledge/old fact hypothesis*.

The ability hypothesis owes to the work of Laurence Nemirow and David Lewis. This response depends on a distinction commonly drawn between propositional knowledge and ability knowledge. Suppose you want to get certified for scuba diving. The process might start with your taking an e-learning course to get you up to speed on all the basic. At the end of the course, you have learned a lot of facts about scuba diving – you know the names of all the pieces of diving equipment, you know the safety rules, you know the principles of buoyancy and weight managements, and you know that the descent should be controlled so as to equalize the pressure in your ears. But at this point, even though you know an awful lot of facts about scuba-diving, there's a certain kind of knowledge that you lack: you lack know-how. You don't yet have the ability to scuba dive. It's only once you transition from the e-course to the actual, hands-on diving course that you will acquire this ability. In acquiring this ability, however, you don't necessarily learn any new facts. You have the facts; now you need to translate those facts into action. According to the proponents of the ability hypothesis, Mary is like the diver who has only taken the e-learning course. When she finally sees color for the first time, though she does gain knowledge, this knowledge is not propositional in nature. Rather, it's the kind of know-how that the diver gains upon diving for the first time. The diver gains scuba diving abilities. Mary gains abilities to imagine, recognize, and remember color experiences. Knowledge of what an experience is like consists in these abilities. If this hypothesis is correct, then we can accept the basic intuition underlying the Mary case without being forced to reject physicalism and endorse property dualism.

The acquaintance hypothesis takes a similar strategy to the ability hypothesis. On both of these responses to the knowledge argument, it is granted that Mary learns something new when she exits the room, but it is denied that what Mary learns consists of a new *fact*. For proponents of the acquaintance hypothesis, however, the kind of knowledge that Mary gains upon leaving the room is not ability knowledge but rather acquaintance knowledge. It is like the kind of knowledge that one gets when one meets someone new or visits a new city for the first time. You might have read all about Sydney in a guidebook before your trip. You know the layout of the city, how the public transportation system works, what the Opera House looks like, and so on. But there's a sense in which you still don't know Sydney until you actually get off the plane and spend some time in the city. You don't know it until you're acquainted with it. Similarly, there's a sense in which Mary doesn't know the color sensation of red until she's acquainted with it. But just as your coming to gain acquaintance knowledge of Sydney need not consist in the acquisition of any new facts, Mary's coming to gain acquaintance knowledge of the sensation of red need not consist in the acquisition of any new facts. If this hypothesis is correct, then we have another way of blocking the inference from the Mary case to the falsity of physicalism.

The third response to the knowledge argument that we'll here consider, the new knowledge/old fact hypothesis, takes a slightly different approach. Like the ability and acquaintance hypotheses, this hypothesis accepts the "Aha!" intuition. But unlike these other hypotheses, this hypothesis does not deny that knowledge of what an experience is like is propositional knowledge, that it is factual in nature. Rather, what it denies is that the fact Mary learns when she leaves the room is one that is genuinely new to her. What happens is that she comes to apprehend an old fact, one she already knew, in a distinctively new way. Consider the fact that, as depicted in the comic books, Wanda Maximoff is 1.70 meters tall. Given that Wanda Maximoff is one and the same person as the Scarlet Witch, it is also true that the Scarlet Witch is 1.70 meters tall. This is the very same fact, even though it's described

differently. Someone who knows Wanda quite well, perhaps living next door to her, might know her height without knowing that she is the Scarlet Witch. When they encounter the Scarlet Witch in full costume, and they thereby come to know that the Scarlet Witch is 1.70 meters tall, they gain knowledge but they do not learn any new fact. Rather, they simply come to apprehend the old fact, one that they already knew, under a new guise. Proponents of the new knowledge/old fact hypothesis want to say something similar about what's going on with Mary. When she comes to have her first sensation of red, she gains a new way of apprehending an old fact (or facts) that she already knew – perhaps the fact that red has such-and-such wavelength and causes such-and-such neurons to fire when someone is exposed to that wavelength. Note that this hypothesis naturally combines with the phenomenal concept strategy that we encountered earlier. Mary's new way of apprehending the old fact can be understood in terms of her acquisition of phenomenal concepts.

3.3 Problems and Objections

We turn now to general issues that arise when one adopts a dualist view. The first concerns what's often referred to as the *problem of mental causation*. Intuitively, our mental states have causal efficiacy, i.e., they can cause various things to happen. Why did I get up and walk to the kitchen? Because I wanted a drink of water and I believe that the kitchen is the nearest source of water. Why did I scratch my arm? Because I had an itching sensation. Why did I jump in my seat when the villain suddenly appeared on screen? Because I was gripped by fear. In this way, mental states like beliefs, desires, sensations and emotions seem to cause a great variety of bodily states. If mentality is fundamentally nonphysical, it is unclear how any of this is possible.

The basic issue was originally raised in connection with Descartes' interactionism. (Recall that this is a version of substance dualism that claims that there is two-way causal interaction between mind and body.) Following the publication of his Meditations, Descartes carried on a lengthy correspondence with Princess Elisabeth of Bohemia. Over the course of this correspondence, Elisabeth raised a number of insightful criticisms of his view, including a version of the problem we are now discussing. As she put it, given that the human mind "is only a thinking substance, how can it affect the bodily spirits, in order to bring about voluntary actions?" Normally when we think about causation, we think about it in terms of contact: the baseball bat hits the ball, causing it to fly into the air, or the ball then hits a glass window, causing it to shatter. Since it is hard to see how a nonphysical thing could be in contact with a physical thing, it is hard to see how a nonphysical thing could have any causal interactions with a physical thing. As Elisabeth concluded, "I have to say that I would find it easier to concede matter and extension to the soul [the mind] than to concede that an immaterial thing could move and be moved by a body."

As stated by Elisabeth, this problem is a conceptual one. But there is a related problem that arises from empirical considerations. Contemporary science suggests that physics is complete and causally closed, i.e., that all physical events can be given a complete causal explanation in physical terms. When I scratch my arm, for example, the movements can be wholly explained in terms of nerve signals, muscle contractions, and other bodily events. There does not seem to be any room for mental causes. This version of the problem of mental causation is often referred to as the *problem of causal exclusion*. Note that the physicalist has an easy solution to the problem. By identifying the mental causes with the

physical causes, the physicalist can accept the completeness of physics and give a robust picture of mental causation.

How might the dualist respond? One possibility is to treat this as a situation of overdetermination. The movement of my arm has a mental causal explanation in addition to the physical causal explanation, and each of these causal pathways is fully sufficient on its own to bring about the arm movement. But this does not seem particularly satisfactory, as it suggests that my arm would move in exactly the same way even if I did not have the relevant mental states. That does not align very well with how we normally think about the causal power of the mental.

Another option for the dualist is to adopt a view called *epiphenomenalism*. This option is most commonly associated with property dualism. Something that is an epiphenomenon is a mere byproduct of (or accompaniment to) a given process that does not itself have any causal power influence of its own. On an epiphenomenalist view, though the brain gives rise to mental states, the mental states lack causal efficacy altogether. In what is often cited as the first statement of the view, T.H. Huxley referred to consciousness as merely a "collateral product" of the mechanisms of the body. On Huxley's view, just as a steam whistle is produced by a locomotive without having any influence on the workings of the train, so too our mental life is produced by the brain and body without having any influence on their workings.

Though epiphenomenalism avoids the problem of causal closure, it is highly counter-intuitive. The phenomenology of our mental life seems to reveal causal connections to us. In addition to the psycho-physical connections across mental and physical states mentioned earlier, we also see psychological connections between mental states (as when the sensation of itchiness causes annoyance, a desire for the itchiness to stop, and a desire to go to the store to buy anti-itch cream). Consider also my assertion, "I have an itch on my arm." We normally think this is caused by the sensation of itchiness itself, but if epiphenomenalism is true, then that thought would be mistaken. Though we're not quite zombies – we do actually have the phenomenal mental states – we share with zombies the fact that our thoughts and assertions about phenomenal states are not in fact caused by phenomenal states.

Another problem for epiphenomenalism comes from our capacity for introspective self-knowledge. In having mental states like beliefs, desires, emotions, and sensations, we are often able to tell immediately and directly what mental states we are having. Our mental life seems to be responsible for our knowledge of our mental life. This seems ruled out, however, if epiphenomenalism is true.

Finally, epiphenomenalism faces a problem arising from considerations of natural selection/evolution. If mental states play no causal role, then they are not traits that nature could have selected for. This objection, which was originally raised by William James, has been more recently developed by Karl Popper and John Eccles. As they argue, if we deny mental events and experiences any biological function, then we have no way to explain the existence of these events and experiences in Darwinian terms. Moreover, we cannot explain how the existence of these events and experiences played a role in the evolutionary development of the physical world, i.e., we cannot offer any explanation for why creatures with more sophisticated mental lives had an evolutionary advantage.

We turn now to a different issue facing dualism, one that arises from the work of Gilbert Ryle in the first half of the 20th century. In a paper called "Categories," Ryle introduced the notion of a *category mistake*. Category mistakes occur when we try to attribute a given property to a thing to which it is inapplicable. Saying that the attribution is inapplicable is not just to say that it's false. Compare these three claims:

- The 16th President of the United States was born in Kentucky.
- The 46th President of the United States was born in Kentucky.
- The number 46 was born in Kentucky.

The first claim is true. Abraham Lincoln, the 16th President of the United States, was born in a log cabin on his father's farm in Kentucky. The second claim is false. Joseph Biden, the 46th President of the United States, was born in Scranton, Pennsylvania. But what about the third claim? We can dismiss it as false, since the number 46 was not in fact born in Kentucky, but the problem here is not that the number 46 was born somewhere else, as was the case with President Biden. Rather, the problem is that the property of being born doesn't apply to something like a number. In attributing this property to the number 46, we make a category mistake.

On Ryle's view, diagnosing category mistakes can help us to properly delineate our ontological categories. He also thinks that category mistakes often go hand-in-hand with various linguistic confusions that can incline us to think that we are faced with a philosophical problem where in fact none exists. In his famous book, *The Concept of Mind*, Ryle argues that the mind-body problem arises from this sort of linguistic confusion, i.e., that dualism rests on a sort of category mistake. To show this, he first asks us to consider someone who visits Oxford or Cambridge. The visitor tours the campus and is shown the various colleges, libraries, offices, and playing fields. Now suppose that, having seen all of this, the visitor then went on to ask where the University itself was. "I've seen all of these buildings where your staff and students work and reside," the visitor says, "but I have yet to see the University." As Ryle notes, the visitor would be making a category mistake. They think that the University is the same type of thing as a building, that it is just one more entity in the class of things to which the libraries and offices belong. On Ryle's view, the dualist makes a similar kind of mistake. Just as the visitor expects the University to be an extra building, in some ways like the offices and libraries though in other ways different from them, the dualist expects the mind to be an existing thing, in some ways like the body but in other ways different from it.

Consider two sentences, "Joe eats a cookie" and "Joe wants a cookie." Because these two sentences have the same linguistic form, we are inclined to think they must work the same way. This is where Ryle thinks dualism makes its fundamental mistake. The first sentence describes a physical process. Given that the second works the same way as the first but does not seem to describe anything physical, the dualist takes it to describe a nonphysical process and posits a realm of the nonphysical to account for it. But this, says Ryle, is like looking for the University after one has been shown all the other buildings. Just as the word *University* does not function linguistically the same way as the word *building*, the word *wants* does not function linguistically the same way as the word *eats*. It doesn't pick out a nonphysical process but instead points to a behavioral disposition. To say that someone wants a cookie is to say

they are likely to eat a cookie if they're presented with it. Other mental terms refer to other behavioral dispositions. And importantly, behavioral dispositions can be explicated entirely without postulating a non-physical realm, a realm Ryle disparagingly refers to as Descartes' "myth" or as the mistaken doctrine of a "ghost in the machine."

In arguing this way, Ryle contributes to the development of a physicalist theory called *behaviorism*, one that aims to reduce mental states to behavior and behavioral dispositions. We will consider behaviorism in slightly more detail in Section 4. I will postpone assessing its plausibility until then.

In addition to the problem of mental causation and the problem of the category mistake, a third set of issues for dualism arise from what's called the problem of other minds. Generally speaking, we take ourselves not only to know about the existence of our own minds but about the existence of other minds as well. My knowledge of exactly what thoughts and feelings my friends and family are having at any given moment in time is less secure than my knowledge of what thoughts and feelings I myself am having, but that they are thinking and feeling – that they have mental states – seems not to be in any serious doubt. On a physicalist view, we have a relatively straightforward explanation of our knowledge of other minds: we know that other brains exist, so given the mind is just the brain or that mental states are properties of the brain, we know that other people have mental lives. But how can a dualist explain our knowledge of other minds? To develop this worry, we might draw on a point made earlier in our discussion of the problem of mental causation. If your mentality is nonphysical in nature, then regardless of whether it's a separately existing substance or properties of the brain, I cannot have any direct causal interaction with it. So how could I come to have knowledge of it?

In response to the problem of other minds, dualists have at least two possible strategies. First, they might adopt an argument from analogy. When I engage in various behavior, I know that the behavior is explained by various mental states. When I see you engage in similar behavior then, I can conclude that your behavior is explained by your mental states on analogy with my own case. One advantage of this response is that it seems to capture the actual reasoning processes that we use to draw conclusions about what others are thinking and feeling. We see their behavior and ask ourselves, "What would I be thinking and feeling if I were behaving that way?" But critics of this argument from analogy often note that analogies based on a single case, as this one is, are not usually particularly strong ones. Note also that this kind of strategy is not available to the epiphenomenalist, since on their view mental states do not actually play any role in explaining behavior – either one's own behavior or other's behavior.

The second strategy involves an inference to the best explanation. Usually when we are trying to explain a given phenomenon there will be several different hypotheses that we could call upon. To decide among the competing hypotheses, we can look at various factors: Which hypothesis has the greatest explanatory power? Which is the simplest hypothesis? Are any of the hypotheses useful in other contexts as well, so that they have a greater degree of generality and allow us to unify our explanations? Having explored these issues, we can determine which hypothesis is the best overall. Having done so, we then take the fact that it is the best hypothesis overall to give us reason to accept it as true; we infer its truth on the basis of the fact that it provides the best explanation.

This reasoning process, a process called *abductive reasoning*, is common both in the sciences and in everyday life, even if it is not always made explicit. When it comes to the behavior of others, we might think that the best explanation overall is one that posits the existence of mental causes. If this is correct, then the dualist has a solution to the problem of other minds: We know of their existence by way of inference to the best explanation. Adjudicating the success of this response would require a more in-depth look at how the hypothesis involving mentality compares to other possible explanations.

4. Physicalism

We turn now to physicalism, the second major theory of mind that we'll consider in this guide. According to the physicalist, everything that exists is physical. Importantly, however, this claim does not commit them to the claim that the mental does not exist. Rather, they typically accept the claim that mental states exist but go on to deny that these mental states are distinct from physical states. Sometimes this point is put by saying the mental is *nothing over and above* the physical or that the mental *depends* on the physical. Sometimes this point is put by invoking the technical notion of supervenience that we encountered earlier in our discussion of property dualism.

Recall that what it means for one set of properties *S1* to supervene on another set of properties *S2* is for it to be impossible that there is a change in *S1* without a corresponding change in *S2*. To say that the mental supervenes on the physical, then, means that there can be no mental difference without a physical difference. Given two different worlds that are completely identical with respect to all of their physical properties, if physicalism is true then those two worlds must be completely identical with respect to all of their mental properties. In fact, since the physicalist claims that *everything* is physical, not just that the mental is physical, the truth of physicalism entails that those two worlds are completely identical with respect to all of their properties whatsoever. Note that this is something the dualist denies. Recall the philosophical zombies discussed in Section 3. My zombie-twin and I differ with respect to our mental properties. I have phenomenal consciousness, and she does not – even though we do not differ with respect to our physical properties.

An understanding of physicalism in terms of supervenience is often described as *minimal physicalism*, since it is often thought that a commitment to supervenience is required for a view to count as physicalist: Though different versions of physicalism may differ from one another in all sorts of respects, they must share a commitment to the supervenience of the mental on the physical. There is considerably more disagreement about whether supervenience is sufficient for physicalism, but we will not pursue this disagreement here.

In subsection 4.1, we will start by discussing some of the general considerations offered in support of physicalism. The next three subsections focus on a more detailed consideration of three different versions of physicalism: behaviorism, the identity theory, and eliminative materialism. In the course of discussing these theories we will also discuss specific objections that can be levied against them. In the final subsection, we turn to some more the general issues facing physicalist theories.

4.1 General Considerations Supporting Physicalism

Interestingly, the case for physicalism is often made not by offering arguments in support of it but rather by offering arguments to show that dualism is an unacceptable theory. That said, there are two sets of considerations that are generally thought to speak in favor of a physicalist view. The first concerns simplicity and the second concerns the explanatory power of science.

Consider this passage from physicalist J.J.C. Smart in which he defends his own preferred version of physicalism, what he calls *the brain-process theory* but that is generally referred to as *the identity theory*: "If it be agreed that there are no cogent philosophical arguments which force us into accepting dualism, and if the brain process theory and dualism are equally consistent with the facts, then the principles of parsimony and simplicity seem to me to decide overwhelmingly in favor of the brain-process theory." The principles of parsimony and simplicity that Smart here has in mind are often referred to as *Ockham's Razor*. William of Ockham, a philosopher working in the medieval period, suggested that when we are constructing our theories, we should refrain from positing entities beyond what is necessary. Any unnecessary elements in one's theory should be shaved off (hence the razor).

Many people take it to be obvious that Ockham's Razor supports physicalism. Given that dualism posits two kinds of entities or properties where physicalism posits only one, it seems that physicalism is the simpler theory and hence should be preferred. Unfortunately, however, this reasoning proceeds too quickly. Abiding by considerations of simplicity does not mean that one never postulates additional entities. Rather, it means that additional entities should not be postulated unless it is necessary to do so in order to achieve an adequate explanation of the phenomena in question.

In postulating two kinds of entities where the physicalist postulates only one, dualists need not be seen as rejecting Ockham's Razor. On their view, physicalist explanations are inadequate to explain humans and human behavior. Dualists claim that their view has greater explanatory power: It is only by postulating the existence of non-physical mentality that we are able to achieve a fully adequate account of humans and their place in the world. Of course, this is precisely what the physicalists deny. What this suggests, then, is that considerations of simplicity can come into play in the debate between dualism and physicalism only once we have adjudicated the question of the explanatory power of these two theories.

The second set of considerations often put forth in support of physicalism stem from the explanatory power and success of science. We encountered some of these considerations in Section 3 when we discussed the problem of mental causation and the causal closure of the physical. Recall that the causal closure principle states that all physical events can be given a complete causal explanation in terms of physical events. In support of this principle, the physicalist might start by pointing to the laws of conservation of matter and energy. Because these laws are defined in terms of a set of basic, fundamental forces, they do not directly entail causal closure. But if the physicalist can establish that all such forces are physical, then they will have what they need to defend causal closure. In an effort to establish this claim the physicalist might point to the fact that the history of scientific inquiry has never revealed any forces other than physical forces at work. Moreover, in every case that scientists have discovered new forces, these forces have been able to be reduced to the same group of fundamental

forces as previously known forces. For this reason, even though we may seem to come across forces that initially present themselves as being irreducibility mental, the history of science suggests that these forces too will be reducible to the same fundamental forces – forces that are wholly physical – as all other known forces.

More generally, similar reflections about the history of scientific progress count strongly in favor of physicalism. In the past, we often encountered phenomena that initially struck us as inexplicable but were then ultimately explained in scientific terms. While it once seemed utterly mysterious how diseases spread, for example, this mystery has now been largely dispelled. Even though neuroscience is still a comparatively young science, we have already made an incredible amount of progress in advancing our understanding of the workings of the brain. Especially with the advent of neural imagining techniques over the last 50 years, many aspects of human cognition that had previously seemed completely mysterious now seem considerably less so. We are still making further progress in this area day by day and year by year, and all signs point to increased progress on this score. Dualism often seems to gain support from the fact that we cannot now explain various phenomena in scientific terms, and indeed, that we cannot now even conceive of how such an explanation would be possible. But given the record of success that science has compiled, physicalists argue that we should not put too much stock in what we now know or can imagine. As Patricia Churchland has persuasively argued, "the mysteriousness of a problem is not a fact about the problem, it is not a metaphysical feature of the universe – it is an epistemological fact about *us*. It is about where we are in current science." As she urges, we should "learn the science, do the science, and see what happens." Once we do, her prediction is that the mysteries of the brain will be unraveled in the same way that so many other past mysteries have been unraveled. Our past record of scientific success is thus taken to be a powerful testament to the plausibility of physicalism.

4.2 Behaviorism

Though behaviorist theories come in several different varieties, they all in some way attempt to understand the mind in terms of bodily behavior. As we saw in our discussion of the problem of other minds in Section 3.3, we rely on our observations of other people's behavior in order to make judgments about what they are thinking and feeling. As this suggests, there is a tight connection between mental states and bodily behavior. But while we intuitively see this connection as *evidential* in nature, the behaviorist suggests that it should be seen as *constitutive* instead. For the behaviorist, we should not think of bodily behavior as the reflection or manifestation of some inner mental state. Rather, exhibiting such behavior is simply what it is to be in the relevant mental state.

At the start of the 20th century, many psychologists had grown disenchanted with the introspectionist methods that were widely employed throughout their discipline. These psychologists, known as *psychological behaviorists*, aimed to shift psychological study towards more objective methods that would put this study on a par with other sciences. In their view, psychology is best understood not as a science of mind but as a science of behavior.

The psychological behaviorists were not primarily concerned with issues concerning the nature of mind; rather, their primary focus concerned scientific methodology. But the advent of behaviorism in psychology gave rise to a similar movement in philosophy, one that was concerned with the nature of

mind. Philosophical behaviorists approach this issue by thinking about the meaning of statements involving mental expressions. Consider a claim like "Nayeli believes that it will rain" or "Nayeli has an itch on her right arm." Typically, these claims are taken to refer to inner states, in the first case to a belief and in the second case to an itch. The logical behaviorists offer an alternative understanding of these claims. On their view, the meaning of these claims does not consist in facts about inner states, whether mental or physical, but rather in facts about behavior – both the actual behavior that one has manifested or about the one's behavioral dispositions, i.e., the behavior that one is disposed to manifest. As put by Gilbert Ryle, "To find that most people have minds ... is simply to find that they are able and prone to do certain sorts of things."

Some philosophical behaviorists thought that the relevant behavioral facts can all be expressed in terms of the language of physics. This form of behaviorism, associated with philosophers such as Carl Hempel, is often called *logical behaviorism* (or *hard behaviorism*). But other philosophical behaviorists like Ryle thought that the facts about behavior can be expressed in ordinary language terms. This form of behaviorism, associated with Ludwig Wittgenstein as well as with Ryle, is often called *ordinary language behaviorism* (or *soft behaviorism*). On this version of the view, when we say that Nayeli has an itch on her right arm, what we mean is something like: Nayeli rubs and scratches at her arm and when asked "what's wrong?" she utters the words, "I have an itch on my arm."

One advantage of the behaviorist view is the solution that it provides to the problem of other minds. Behaviorists like Wittgenstein worried that, if mental states like pains and itches are inner states that we each only know on the basis of our own experience, we can never have any knowledge of other minds. For example, since my word "pain" refers to my private experience and your word "pain" refers to your private experience, how do I know that our states are of the same type? More generally, I can never know that you are in the same kind of state that I am in; indeed, I can never really know anything about your mental states at all. These problems dissipate if we embrace behaviorism. Since all that it is to be in a given mental state is to manifest certain behavior, we can know everything there is to know about mental states simply by observing that behavior.

That said, though behaviorism nicely handles the problem of other minds, in doing so it opens itself up to a different kind of epistemological worry. It is generally agreed that there is an asymmetry between the way that we know about our own mental states and the way we know about the mental states of others. Though I can only know about your mental states by observation of your behavior, I can know about my own mental states by way of introspection. (Perhaps there are particular cases in which I lack this kind of immediate self-knowledge, as when I am self-deceived, but in the typical case my mental states are introspectively accessible to me.) If the behaviorist were right, however, and mental states were just behavioral dispositions rather than inner states of any sort, there is nothing for me to access by way of introspection. I can't know my own mental states immediately and directly but could only achieve this knowledge by way of inference from observations of my own behavior. Thus, the only way I could know about my own states would be the very same way that I know about yours. This treatment of our self-knowledge has struck many philosophers as deeply counter-intuitive.

In addition to this problem about self-knowledge, philosophers have raised several other problems for behaviorism. One problem stems from the behaviorist's treatment of claims such as "I am in pain" or "I want an ice cream." Normally we think of such claims as providing information about our mental life, i.e., as reports of inner states. For the behaviorist, however, such claims are not reporting on anything;

in fact, there is nothing to report on. They are just one instance of the general class of pain-behavior, no different from winces and grimaces. The inability to treat such claims as mental state reports has struck many philosophers as problematic.

A further criticism owes to the work of Hilary Putnam. In an influential paper called "Brains and Behavior," Putnam asked readers to consider a thought experiment involving a community of stoic individuals called "Super-Spartans." All of the adult members of the community have trained themselves to entirely suppress all of their pain behavior. When they have pain, they refrain from wincing or grimacing or rubbing at the affected bodily part. Yet they still feel pain just as we do, and will admit to being in pain if asked. Taking this one step further, Putnam asks us to imagine a community of Super-Super-Spartans who don't make any verbal reports of their pain and won't even admit to being in pain if asked. Insofar as this case is conceivable, and most philosophers seem to agree that it is, then we have a case of pain without any pain-behavior or even any disposition towards pain behavior, a situation that should be impossible if behaviorism were true.

Notice also that we can have cases of the reverse sort, i.e., cases where we have pain-behavior even without any pain. Think of an actor on stage playing a part of someone who has just been shot. Though not in pain, the actor drops to their knees, clutches the wound, grimaces and cries out. This behavior may well be indistinguishable for the behavior exhibited by an actual gun-shot victim, a victim who is in pain. Given that we can have cases of pain without pain-behavior and pain-behavior without pain, philosophical behaviorism must be rejected.

One last objection concerns a worrisome circularity inherent in the behaviorist specifications of mental states. In articulating the behavior and behavioral dispositions that someone has when they are in pain, it looks like we will have to make reference to other mental states. For example, though someone might typically cry out when they are in pain, they will not do so when they are in an environment where they need to stay quiet, like a library or a classroom. Likewise, though someone might typically scratch their arm when they have an itch, they will refrain from doing so if they believe that their scratching will irritate the skin. To specify the behavior and behavioral dispositions, then, we have to do so more carefully. Rather than saying that Nayeli will scratch her arm, we need to say something like: she will scratch her arm as long as she believes that doing so will not irritate her skin. But now our definition of itch makes reference to the state of belief. More generally, it seems likely that any adequate specification of the behavior and behavioral dispositions for a given mental state will have to make reference to others, and we are soon led to overlapping specifications and, in some cases, to circular specifications. If this is right, then the behaviorist project will not be successful in eradicating reference to irreducible mentality.

Taken together, these criticisms are generally viewed as so devastating that behaviorism has long since fallen out of favor. Though at this point there are a few philosophers who have behaviorist inclinations, it would be fair to say that the view is no longer regarded as viable by the vast majority of philosophers.

4.3 Identity Theory

The identity theory entered philosophical discussion in the middle of the 20th century through the work of U.T. Place, Herbert Feigl, and J.J.C. Smart. Like the behaviorist theory just considered, the identity theory is a reductive theory. But while the behaviorists attempt to reduce mental states to behavior, the identity theorists attempt to reduce mental states to brain states.

As we saw when we first introduced the notion of reduction in Section 3, when scientists reduce one theory to another, the entities postulated by the former theory can often be reduced to, or identified with, the entities postulated by the latter theory. For example, when scientists reduced the classical gas theory to statistical mechanics, they were able to reduce the property *temperature of a gas* to the property of *mean molecular kinetic energy*. Other developments in science led us to identify water with H_2O and lightning with electrical discharge. Drawing an analogy with these scientific identities, the identity theorists argue that mental states can be identified with particular brain states. As one example of a psychophysical identity, we might identify the state of pain with some particular brain state – call it *C fiber firing*.

(Though this example is widely used in the literature, and though we will continue to use it here, it's worth noting explicitly that this is really meant to be just a place-holder. While the brain does contain nerve fibers called C fibers, and while they are identified as nociceptors and thought to be involved when we experience pain, it is unlikely that it will turn that pain reduces smoothly and directly to the firing of C fibers – even if the identity theory is true. Only once neuroscience has advanced to a sufficient point would identity theorists be able to refine this identity claim to one that is more accurate.)

Further clarification of the identity theory requires us to take note of the distinction between tokens and types. Consider Spelling Bee, the popular (and addictive) word puzzle offered daily by the *New York Times*. In Spelling Bee you are given seven letters. They are arranged in a hexagonal shape, with one letter in the middle, and you have to make as many words as possible using those letters. According to the instructions: "Words must contain at least 4 letters. Words must include the center letter. … Letters can be used more than once." To understand these instructions and succeed at the game, you have to understand the distinction between letter types and letter tokens. Though words must contain at least four letter tokens, they need not contain at least four letter types. So, for example, in yesterday's puzzle the letters were **O D L T A I N**, where **O** was the center letter. One of the acceptable words was LOOT. This word is four letter tokens long, though it only uses three letter types. Another acceptable word was ADDITIONAL. This is a ten-letter word, i.e., it uses ten letter tokens. But it contains only seven letter types, as the **A**, **D**, and **I** are each used twice.

Just as we can talk of both letter state types and letter state tokens, we can also talk about mental states types and mental state tokens. In identifying pain with C fiber firing, or more generally, in identifying mental states with brain states, the identity theory is best seen as a theory about mental state types. For this reason, the identity theory is often referred to as *type physicalism*. Identity theorists are not just claiming that this particular pain state token happens to be identical to this particular token state of C fiber firing. Rather they are claiming that all instances of the mental state type pain will be an instance of the brain state type C fiber firing; the mental state type pain can be identified with the brain state type C fiber firing.

The identity theorists were motivated in part by a desire to correct a problem that arose for the behaviorists, namely, that reports about mental states seem to be genuine reports. Note that dualism was not beset by this problem. On the dualist view, a statement like "I am in pain" is a genuine report. But the identity theorists were troubled by the idea that such claims report on something as nonscientific as an irreducible psychical thing. On their view, such reports avoid this trouble. When I report that I am in pain, or that I have some other mental state, I am making a report about a given

brain state. In this way, mental state reports can be seen as genuine, and what they're reporting on is scientifically respectable.

But this point may seem to give rise to a troubling objection. People can make reports about their pains and other mental states even if they don't have a sophisticated understanding of the brain; in fact, we can talk about our pains without any knowledge of brains at all. Lots of people who know that they are in pain have never even heard of C fibers! But the identity theorist has an easy answer to this objection. In claiming that pain is identical to C fiber firing, the identity theorists are not claiming that the words "pain" and "C fiber firing" are synonymous. Consider a claim like: a square is an equilateral rectangle. This claim is true by definition. It is an analytic truth. But the psychophysical identities posited by the identity theory are not meant to be true by definition. They are not analytic truths, and the identity theorist is not making an analytic reduction. Rather, they are making an ontological reduction, i.e., a reduction of one ontological category to another. This connects to a point mentioned earlier: The identity theorists take psychophysical identities to be scientific discoveries on the model of other scientific discoveries such the identification of lightning with electrical discharge. These scientific claims are not knowable *a priori*, independent of experience, and they are not analytic. Thus, just as one can use the word "lightning" without any knowledge of electricity, one can use the word "pain" without any knowledge of the brain.

There are other criticisms of the identity-theory that cannot be handled quite so easily. One important problem arises from what we might call *the multiple realizability objection.* In short, mental states appear to be realizable in multiple different physical structures. Consider the example of Martian pain offered by David Lewis. Suppose a (friendly) Martian were to land on earth. When we do a CT scan and other tests on the Martian, we discover that it doesn't have a brain in any way like ours. It doesn't have any neurons but instead has some kind of heat exchanger inside its head. Its system seems to operate entirely hydraulically: "there are varying amounts of fluid in many inflatable cavities, and the inflation of any one of these cavities opens some valves and closes others." But the Martian nonetheless exhibits behavior that looks very much like pain. When exposed to painful stimuli, it grimaces, cries, takes evasive action, and (as we can tell once we get our universal translator working) reports that it is in pain. We might even suppose that when exposed to painful stimuli it produces states that are behaviorally indistinguishable from our own. It seems extremely plausible that the Martian is in pain, even though the Martian is not in a state of C fiber firing. Though our pain and mental states are realized in our neural structures, its pain and mental states are realized in its hydraulic structures. But if this right, if mental states like pain can be multiply realized, then they cannot be identical to brain states.

Sometimes people dismiss the Martian example because they worry that we cannot really be sure that the Martin is in pain. But we can also use a first-personal example to demonstrate multiple realizability. Just as scientists have developed artificial organs and other kinds of artificial implants, it seems plausible that they might one day soon develop inorganic artificial neurons. Laboratory tests show that the artificial neurons integrate perfectly into the overall neural framework and function exactly the way that organic neurons do. Now consider the following thought experiment:

> *While driving home one rainy night, you're in a car accident and suffer neural damage that affects your C fibers. In a complicated brain surgery, the surgeons replace your damaged C fibers with the newly developed artificial neurons. When you wake up in the hospital, you discover that*

> *you have a headache. The headache feels just like headaches that you've had on other occasions in the past.*

Assuming this thought experiment is a coherent one, and many people hearing it agree that it is, then we have another instance of the multiple realizability objection to the identity theory. You are in pain even though you are not in a state of C fiber firing.

But perhaps this case too seems problematic, too much an invention of science fiction. We might then consider one last case. Consider an animal that has a very different neural structures from our own, like a shark. Given the way sharks respond to painful stimuli, it seems plausible that they experience pain even though they don't have any C fibers. If this is right, then the mental state type pain cannot be identified with the mental state type C fiber firing.

Confronted with these cases, one might worry whether the identity theorist would do better to retreat to a weaker position. Even if it is not true that mental state types can be identified with physical state types, perhaps every token mental state token can still be identified with some token physical state. Our token pains are identical to C fiber firings, the Martians token pains might be identical with hydraulic state H1, the accident victim's token pains might be identical with the firing of the artificial neuron, and the shark's token pains might be identical with some state of its brain. In all of these cases the relevant states are physical in nature. They would all be tokens of some physical state or other even if not of the same physical type. Though this would is incompatible with type physicalism, it is fully consistent with a view we might call *token physicalism*.

Unfortunately, however, there are reasons to think that the retreat to token physicalism is unsatisfactory. If the identity theorist has to construe their view in terms of token identity statements rather than in terms of type identity statements, it looks like they will be unable to satisfy their explanatory goals. In evaluating theories of mind, we expect to be provided with an explanation of mentality, an explanation of what makes a mental state mental. A type physicalist theory gives us an answer to this question: a mental state is mental in virtue of being a brain state. This answer might not be entirely satisfying, as there are various brain states that we don't consider to be mental states – the unconscious processes involved in the control and regulation of breathing for example. But the explanation offered is nonetheless a unifying one, and there is potential down the road for a more complete and satisfying answer as type physicalism is further refined. If the identity theorist adopts token physicalism, however, then they really can't provide any answer to this question. All that can be said about what makes a mental state mental is that it is a physical state. But of course there are many physical states that are not mental states – the physical states of the books and pens and pencils on my desk, for example. Token physicalism has not provided us with any explanation of mentality and it does not seem well positioned to do so.

4.4 Eliminative Materialism

Like the identity theory, eliminative materialism takes its inspiration from the history of science and the development of scientific understanding. But while the identity theorist points to cases in which our increased scientific understanding of a given phenomenon allows us to reduce one entity to another, the eliminative materialist points to cases in which our increased understanding allows us to eliminate an entity altogether.

Consider the theory called *vitalism*. At one time, before it was well understood what accounted for life and how to explain the difference between living things and non-livings, vitalists postulated that life could be explained in terms of a vital force, an élan vital, that flowed through living things. It was only with the development of chemistry in the 18th and 19th centuries that scientists were able to understand why vitalism should be rejected. In particular, when experiments by Friedrich Wohler in the early 19th century showed that organic substances could be synthesized entirely from inorganic substances, the foundations of vitalism were overturned. Over the course of the 19th century and the early 20th century, the properties of life that had previously seemed explicable unless we postulated a vital force came to be explained by biochemistry, genetics, and natural selection. Importantly, however, these new explanations were not achieved by coming to understand the real nature of the vital force. Though the development of statistical mechanics could be said to have revealed the true nature of temperature such that we came have a better understanding of what it is, the development of biochemistry led to the rejection of the very notion of a vital force. The notion was eliminated from our scientific discourse.

For the eliminative materialist, once we have achieved an adequate understanding of the working of the brain and neural system, we will see that talk of mental states is more like talk of a vital force than talk of temperature. In the true scientific understanding of cognition, mental states are not likely to be reducible to physical states. Rather, what's much more likely is that they will be altogether eliminated.

In arguing for their theory, eliminative materialists often note that the reason that we believe in mental states is that they are embedded in our intuitive understanding of ourselves – what they refer to as *folk psychology*. When we attempt to explain behavior in terms of beliefs, desires, and other mental states, we are using the theory of folk psychology to do so. But as the history of science has shown, folk theories are often false. To give just one example, consider folk physics. Aristotle famously claimed that heavier objects fall faster than lighter objects. Though this was disproven in the 16th century by Galileo, who is said to have tested the Aristotelian claim by dropping two spheres of different masses from the Leaning Tower of Pisa, many people continue to find it intuitive. It remains part of their folk understanding of the world, along with numerous other false claims as well. Why should we think that folk psychology fares any better than folk physics?

Perhaps the best-known proponents of eliminative materialism are Paul and Patricia Churchland. In developing the line of criticism that we have just been discussing, the Churchlands offer an argument for eliminative materialism that we might call *the argument from the falsity of folk psychology*. We can put it in standard form as follows:

Premise 1. Mental states like beliefs and desires are theoretical posits of folk psychology.

Premise 2. Folk psychology is false.

Premise 3. The theoretical posits of false theories do not exist.

Conclusion. Therefore, beliefs and desires do not exist.

The discussion thus far has suggested some general reasons in support of premise 2, namely, that folk theories have a bad track record. The Churchlands also offer some reasons to be concerned about folk psychology in particular. First of all, it is stagnant. Productive theories are developed, revised, and extended over time. When it comes to folk psychology, however, we are still working with the same

basic folk psychological notions and principles that have been used for millennia. Second, folk psychology does not have the kind of explanatory power that we want from our theory of the mind. There are numerous questions about mental phenomena that folk psychology does not seem very well positioned to answer. Paul Churchland offers an impressive list: What is mental illness? What is the psychological function of sleep? How do we construct a three-dimensional visual image from the two-dimensional stimulations in our retina? What accounts for our ability to retrieve information from memory at such a blazingly fast speed?

Despite these considerations in its favor, eliminative materialism is not a very widely-held view, and it faces several persistent objections. The first problem that is typically raised for eliminative materialism is simple: the theory simply seems wildly counterintuitive. Many people have a reaction to it that might be seen as analogous to a common reaction to philosophical skepticism. How do I know I have hands? Look here, one might say, waving one's hands in the air. How do I know the physical world exists? Look here, one might say, kicking a rock. When it comes to mental states, we can't wave them or kick them, but we have a great deal of confidence in their existence nonetheless. Just as we feel more secure in our judgment that we have hands than in any of the hypotheses proposed by the skeptic, we feel more secure in our judgment that we have mental states than in any of the arguments put forth by the eliminative materialist. Our certainty about the existence of our own mental states takes priority over any of the theoretical concerns that might be raised.

A second objection concerns the fact that eliminative materialism seems to be a self-refuting theory. As a general matter, we take assertions to be statements of belief, so in asserting their theory, the eliminative materialists seem to be expressing a belief. In effect, an eliminative materialist can be seen to be saying something akin to: "I believe that there are no beliefs." This claim is an incoherent one.

The third objection that we'll consider attacks the argument from the falsity of folk psychology and, in particular, premise 2 ("Folk psychology is false"). To defend the truth of folk psychology, one might note that it has considerably more predictive and explanatory power than the Churchlands allow. As we navigate our daily interactions with other people, we are constantly calling upon folk psychology to predict what they're going to do or to explain what they've done, and the theory thus proves useful to us. Though our predictions and explanations are not always perfect, oftentimes the problem lies not with folk psychology itself but with us. Just as some of our scientific mistakes come from the misapplication of a perfectly good scientific theory, some of interpersonal mistakes come from the misapplication of folk psychology. Insofar as these mistakes do seem to be the result of the theory itself, one need not conclude that it must be altogether discarded. While folk psychology may well be in need of refinement in various ways, it serves us pretty well in ordinary life and we thus have no reason to reject it as false.

Relatedly, one might worry that acceptance of a view like eliminative materialism seems premature in light of the current state of neuroscience. Perhaps once we have a better understanding of the brain, we will discover that folk psychology is problematic and that some of its posits need to be eliminated, but perhaps not. Moreover, even if folk psychology does prove problematic, that does not show that we will need to throw out all of our mental state concepts. It might be that some of them can be retained as is and that others can withstand the necessary refinements. We have no reason to think that the problems with folk psychology will require its wholescale rejection. Thus, to suggest that we should

already be abandoning mental state talk – as the Churchlands themselves recommend, and as they try to do in their own personal lives – seems at this stage to be both rash and unnecessary.

4.5 General Problems and Objections

In our discussion thus far, we have already noted various problems and objections that are specific to particular versions of physicalism on offer. Before we close our discussion of physicalism, it will be worth noting some general problems and objections that all versions of physicalism seem to face.

Many of these worries concern qualia. In short, physicalism seems unable to account for the phenomenal character of our mental states. Sometimes these worries are developed by way of a famous example given by Thomas Nagel in a paper from 1974. As Nagel notes, though it seems fairly certain that bats, being mammals, have phenomenal experience, it seems impossible for us to know what it's like to be a bat. Bats navigate the world by way of echolocation, something that Nagel alleges we can neither experience nor imagine. Though facts about bat experience concern subjective properties, science can only capture objective properties. Facts about experience thus seem to lie outside the domain of science.

As this brief discussion of Nagel suggests, the bat example has much in common with Jackson's knowledge argument. Since we have already discussed that argument in detail in the context of property dualism, we will not rehearse the considerations further here. But it will be worth noting one other way that considerations about phenomenal experience are often thought to count against physicalism. As famously argued by David Chalmers, though there are many problems relating to the mind that will not be easy to solve, phenomenal consciousness seems to present us with a particularly difficult problem, so much so, in fact, that he has termed it the *Hard Problem*.

Consider first various psychological processes such as information processing and behavioral control. Though we have not yet achieved a full understanding of how these processes work, Chalmers notes that there is good reason to be optimistic that continuing scientific investigation into the mechanisms of our cognitive system will provide answers. In contrast, when it comes to questions about how the brain gives rise to qualitative experience – why a physical process should be accompanied by one particular qualitative feel rather than another, or indeed, why it should be accompanied by any qualitative feel at all – we are faced a with a mystery that seems virtually unsolvable.

Take an emotion like jealousy, for example. Some aspects of explaining jealousy seem to be well within the reach of science. For example, science will undoubtedly one day be able to explain which brain processes are associated with jealousy and what role it plays in the human system. But the problem of accounting for the fact that these brain processes come with that particular phenomenal feeling of jealousy, or accounting for the fact that they come with any feeling at all, is one that science does not seem well positioned to answer. The way that jealousy feels seems to lie outside the objective realm in which science operates. As such, it does not look like physicalism in any of it forms will be able to provide an adequate account of the phenomenal aspects of jealousy, or of phenomenal consciousness more generally.

A second set of general worries concerning physicalism arise from what's often referred to as *Hempel's Dilemma*, named after the 20th century philosopher Carl Hempel who first proposed it. (We encountered Hempel earlier in our discussion of behaviorism.) The specification of physicalism relies crucially on the notion of *physical*. To explicate this notion, physicalists tie it to the science of physics:

What it is for an entity to be a physical entity is for it to fall within the domain of physics or to be explained by those that fall within the domain of physics. An important question then arises: What is meant by the domain of physics? Does the physicalist want to rely on physics in its current form or does the physicalist want to rely on an idealized future physics? According to Hempel's Dilemma, neither of these options will be satisfactory, and so the physicalist cannot provide an adequate specification of their own theory.

To see why neither option is satisfactory, let's start with the specification in terms of current physics. Consider quarks. Quarks were first proposed in 1964. Several years later experiments conducted at Stanford provided some evidence for their existence. Prior to the 1960s, then, physics did not include quarks within its domain. But surely quarks are the kinds of entities that are meant to be captured by it, and so when they were discovered physics expanded to accommodate them. Though our physics has advanced since the 1960s, there is no good reason to think that at this point in time it is entirely complete. New particles may well be discovered. After all, in the decades following the discovery of quarks, scientists have also discovered the existence of gluons and bosons. It thus seems that what the physicalist means by "physics" cannot be current physics. If it were, then physicalism would surely be false.

This suggests that the physicalist should define their theory in terms of an idealized future physics. But this also seems problematic. Here the problem is not one of falsity but of triviality. At this time, it's hard to see what the domain of future physics will involve. For all we know, it might be expanded in such a way that it would include properties that we currently think of as mental. Suppose, for example, that future physicists were to find strong evidence for the existence of immaterial ghosts, evidence that can't be explained in terms of properties that are already accepted by physics – properties like having mass, having charge, etc. As a result, the future physicists add the property "being a ghost" to the domain of physics. Though intuitively the existence of ghosts should be incompatible with physicalism, if we define physicalism in terms of a future physics, we don't get this result. Physicalism thus becomes trivial. In short, if anything that can't be explained in terms of existing physical theory simply gets added as a new primitive element of that physical theory, then the notion of *physical* on which the physicalist needs to rely becomes an empty one.

Granted, we might think that the problem raised by Hempel's dilemma arises for the dualist as well. After all, just as the physicalist needs a way to understand the notion of the *physical*, the dualist needs a way to understand the notion of the *nonphysical*. But given that the dualist may be able to fall back on other ways of defining their theory, perhaps in terms of claims about fundamental mentality, the problem seems more pressing for the physicalist than for the dualist.

5. Functionalism

We come now to the final theory of mind that we will consider: functionalism. Functionalism, which emerged onto the philosophical scene in the second half of the 20th century, draws on some important insights gleaned from both behaviorism and the identity theory. From behaviorism, we learn that there is a tight connection between mental states and behavior. In thinking about where the theory went wrong, we see first of all that the connection is not a constitutive one, and we also see that mental states appear to be interdefinable. From the identity theory, we learn that sentences involving mental

states seem to be genuine reports. In thinking about where this theory went wrong, we see that mental states appear to be multiply realizable.

Functionalism also takes inspiration from developments in computer science and artificial intelligence. Just as a computer program can be run on many different types of machines, a mental program can be run on many different types of physical constitutions. Mental states should not be defined in terms of their physical structure but rather in terms of their functional structure, in terms of the functional role that they play. For the functionalist, the mind can be seen as the software of the brain.

The functional role of mental states can be specified in terms of three different components: (1) the inputs to the system (e.g., what's seen, heard, and so on); (2) the behavioral outputs of the system; (3) the relations to other mental states. Specifying these functional roles will thus give us a large, interconnected network of states. Just as a computer program will yield different outputs depending on what machine state it is in upon receiving a given input, a human will produce different behavior depending on what mental state it is in upon receiving certain perceptual information. If I am hungry and thirsty, I will behave differently when I pass by a Pret than I will if I have just had lunch; if I am angry with a colleague I will behave differently when they ask me for a favor than I will if I am feeling grateful for something they've just done.

Strictly speaking, functionalism is compatible with both dualism and physicalism. In defining a functional state, we make no reference to the make-up of the system in which it is instantiated. An artificial heart made of plastic, metal, or ceramic can be in the same functional state as a heart made of organic material. Theoretically, then, a system made of neurons, a system made of silicon, plastic and copper, and a system made of ectoplasm could all be in the same functional state as well. That said, most functionalists are sympathetic to physicalism, and functionalism is often treated as compatible with token physicalism.

Functionalism has been thought to work very well for beliefs and other propositional attitudes. It works less well, however, when it comes to phenomenal states. Just as qualia-based objections have proved a significant threat to physicalist theories, qualia-based objections prove a significant threat to functionalist theories. Notice that some of the arguments that we've already seen are just as problematic for the functionalist as for the physicalist. Recall the knowledge argument. From her perspective inside the room, Mary knows all the functional facts about color and color vision as well as all of the physical facts. So, if successful, the knowledge argument counts against functionalism as well as physicalism. In what follows we will consider two additional arguments relying on considerations about qualia that are raised specifically against functionalism: the inverted qualia argument and the absent qualia argument.

Consider two students, Grace and Nirel, both of whom are working in the campus dining hall. They have been asked to sort the ripe tomatoes from the unripe tomatoes. They each judge the green tomatoes to be the unripe ones and put them in a pile on the left, and they judge the red tomatoes to be the ripe ones and put them in a pile on the right. Moreover, each of them judge that unripe tomatoes share the same color as limes, grass, and emeralds, whereas ripe tomatoes share the same color as strawberries, London Double-decker buses, and rubies. When driving, they each stop when the stoplight turns red and start moving when the stoplight turns green. They each associate red with Valentine's Day and green with St. Patrick's Day. In short, they are functionally identical with respect to the states "having a red experience" and "having a green experience." It seems possible, however, that whenever Grace is

having a red experience, she has the same qualia as the qualia that Nirel has when having a green experience, and vice versa. Their qualia could be completely inverted to one another. Nothing in the functional specification of the states rules this the possibility of this kind of inverted spectrum. Given that having green qualia is essential to the state "having a green experience," and having red qualia is essential to the state "having a red experience," functionalism thus seems unable to account adequately for the nature of our mental states. If two systems can be in identical states while being in qualitatively different states, then functionalism cannot be an adequate theory of mind.

In standard form, the argument can be captured as follows:

Premise 1.	The inverted spectrum scenario is conceivable.
Premise 2.	If the inverted spectrum scenario is conceivable, then it is possible.
Sub-conclusion 1.	Therefore, the inverted spectrum scenario is possible.
Premise 4.	If the inverted spectrum scenario is possible, then functionalism does not adequately account for qualitative states.
Conclusion.	Therefore, functionalism does not adequately account for qualitative states.

While the inverted qualia argument suggests that functionalism allows for cases where two individuals' mental states are qualitatively inverted with respect to one another, the absent qualia argument suggests that functionalism allows for case where an individual's mental states lack qualia altogether. In a sense, a case of absent qualia is akin to the kind of zombiehood that we encountered when discussing property dualism. But since the absent qualia argument has been presented differently in the context of functionalism from the way the zombie argument is presented in the context of property dualism, it will be worth considering it here. More specifically, we will consider the argument presented by Ned Block in his 1978 paper "Troubles with Functionalism."

Block's argument begins by describing what he calls a *homunculi-headed robot*. The robot has a body that from the outside looks very much like a human body but whose internal structure is very different. Inside a hollow cavity in the robot's head where a brain would normally be found, the robot has a very large number of little individuals (the homunculi). Each of these homunculi has an extremely simple task. For example, homunculus 7694 has to produce push a button marked O_{19} when the light marked I_{17} comes on, while homunculus 8329 has to push a button marked O_{783} when the light marked I_{92} comes on. These tasks each correspond to what a single neuron in the brain would do. Working together, the homuculi produce a system that is functionally identical to the human brain. According to Block, however, it's obvious that the homunculi-headed robot is not having states with qualia. Even if it behaves just like a human, it is a system that is entirely absent of qualia.

Anticipating worries that such homunculi are impossible, Block next offers us a different version of the same rough scenario, one that simulates the brain on a much larger scale. We recruit a very large number of volunteers, as many as we would need to assign each volunteer the role of a single neuron in the human brain. At the time that he was writing, Block thought that about one billion volunteers would be enough. Given that it's now thought that the human brain has approximately 86 billion neurons, we would need a lot more volunteers than Block thought (and in fact, much more than the

current population of the earth), but that doesn't make a difference to the structure of the argument. Once we have all the volunteers, we assign each individual the kinds of tasks that were previously assigned to the homunculi. These 86 billion volunteers take up a lot of space, and so won't fit inside the robot's hollow cavity. But they could be connected up to the robot by a system of two-way radios. With everything up and running, we again have a system that is functionally identical to a human being. Perhaps we couldn't keep the system running for very long – the volunteers get tired and bored – but, says Block, we could probably have the system in place for at least an hour. Just as with the previous case, the system consisting of the robot plus the volunteer network lacks qualia and thus presents us with a counterexample to functionalism. If there can be two functionally equivalent systems, one that has qualitative states and one that does not, then functionalism cannot be an adequate theory of mind.

In standard form, the absent qualia can be captured as follows:

Premise 1.	We can conceive of an absent qualia scenario, i.e., a scenario where a system lacking qualitative states is functionally equivalent to a system that has qualitative states.
Premise 2.	If the absent qualia scenario is conceivable then it is possible.
Sub-conclusion.	Therefore, the absent qualia scenario is possible.
Premise 4.	If the absent qualia scenario is possible, then functionalism does not adequately account for qualitative states.
Conclusion.	Therefore, functionalism does not adequately account for qualitative states.

Like the conceivability argument and the zombie argument that we considered in Section 3, these two qualia-based arguments against functionalism move from claims about conceivability to claims about possibility. Thus, the kinds of worries about the legitimacy of this inference that we saw earlier will apply here as well. But there are other strategies that the functionalist might use in responding to these arguments in an attempt to save their theory. Typically, these strategies involve attacking the first premise of each argument, that is, attacking the conceivability of the scenario being proposed.

In response to the inverted qualia argument, the functionalist might insist that any such spectrum inversion would have to manifest in behavior. Even if Nirel and Grace use color words the same way, for example, there may be more subtle differences in their behavior that would show that they are in functionally different states. For example, perhaps one of them categorizes red as a warm shade and green as a cool shade, while one of them might categorize green as a warm shade and red as a cool shade.

In response to the absent qualia argument, the functionalist might point out that the system being operated by the network of human volunteers will be operating on a much slower timescale than the human brain. Even when executing a single simple task, a human cannot respond as quickly as a neuron can. According to the functionalist, the difference in response times shows that the robot is not really functionally identical to you. Block himself dismisses this kind of objection. On his view, the time scale of the system shouldn't matter. If we encountered an alien species who operated on a dramatically

different time scale from us, this would not prevent us from comparing their states to ours and attributing states like belief and pain to them.

6. Suggested Further Reading

On what we mean by mind

For a more detailed discussion of these general issues:

- Kind, Amy (2020). *Philosophy of Mind: The Basics*. Routledge Press. (See chapter 1.)

For an influential development of worries about the notion of qualia:

- Dennett, Daniel C. (1988). Quining qualia. In Anthony J. Marcel & E. Bisiach (eds.), *Consciousness and Contemporary Science*. Oxford University Press.

On dualism

For the classic presentation of substance dualism:

- Descartes, René (1641/1986). *Meditations on First Philosophy: With Selections From the Objections and Replies*. Translated by John Cottingham. Cambridge University Press.

For Princess Elisabeth's objection to Descartes' interactionism:

- Shapiro, Lisa (ed.) (2007). *The Correspondence Between Princess Elisabeth of Bohemia and René Descartes*. University of Chicago Press.

For discussion of the problem of mental causation:

- Yoo, Julie. (2018) "The Mental Causation Debates in the 20th Century." In Amy Kind, ed., *Philosophy of Mind in the Twentieth and Twenty-First Centuries*. Routledge Press.

For a recent development of the conceivability argument:

- Gertler, Brie (2008). "In Defense of Mind-Body Dualism." In Joel Feinberg and Russ Shafer-Landau, eds., *Reason and Responsibility: Readings in Some Basic Problems of Philosophy*, 285-297. Thomson Wadsworth, 2008.

For discussion of the divisibility argument:

- Brook, Andrew & Stainton, Robert J. (2000). *Knowledge and Mind: A Philosophical Introduction*. Bradford. (See chapter 5.)

For a development and defense of property dualism:

- Chalmers, David J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

For the original presentation of the knowledge argument and influential criticisms of it:

- Conee, Earl (1985). Physicalism and phenomenal properties. *Philosophical Quarterly* 35 (July):296-302.
- Jackson, Frank. (1982). Epiphenomenal qualia. *Philosophical Quarterly* 32: 127–136.
- Lewis, David K. (1988). What experience teaches. *Proceedings of the Russellian Society* 13: 29–57.

- Nemirow, Laurence (1990). Physicalism and the cognitive role of acquaintance. In William G. Lycan (ed.), *Mind and Cognition*. Blackwell.

For the zombie argument and criticisms of it:

- Chalmers, David J. (1995). The puzzle of conscious experience. *Scientific American* 273 (6):80-86.
- Dennett, Daniel C. (1991). *Consciousness Explained*, Boston: Little, Brown, & Co.
- Dennett, Daniel C. (1995). The unimagined preposterousness of zombies. *Journal of Consciousness Studies* 2 (4):322-26.

On physicalism

For a classic treatment of psychological behaviorism:

- Skinner, B.F. (1974). *About Behaviorism*. New York: Alfred A. Knopf, Inc.

For classic treatments of philosophical behaviorism:

- Hempel, Carl G. (1980). "The Logical Analysis of Psychology." In Ned Block, ed., *Readings in the Philosophy of Psychology*, Volume 1. Cambridge, Mass.: Harvard University Press, 1-14.
- Ryle, Gilbert (1949/2000). *The Concept of Mind*. Penguin Classics.

For a defense of the identity theory:

- Feigl, Herbert (1958). "The 'mental' and the 'physical'." *Minnesota Studies in the Philosophy of Science* 2:370-497.
- Smart, J.J.C. (1959). "Sensations and brain processes." *Philosophical Review* 68 (April):141-56.
- Place, Ullin T. (1956). "Is consciousness a brain process." *British Journal of Psychology* 47 (1):44-50.

For discussion of causal closure:

- Papineau, David. 2002. *Thinking About Consciousness*. Oxford: Oxford University Press.

For discussion and development of eliminative materialism:

- Churchland, Patricia. (1986). *Neurophilosophy: Toward a Unified Science of the Mind/Brain*. Cambridge, MA: MIT Press.
- Churchland, Paul (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy* 78: 67–90.
- Churchland, Paul (1988). *Matter and Consciousness*, revised edition. Cambridge, MA: MIT Press.

For qualia-based worries about physicalism:

- Nagel, Thomas (1974). What is it like to be a bat? *Philosophical Review* 83: 435-50.
- Chalmers, David J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.
- Chalmers, David J. (1995). The puzzle of conscious experience. *Scientific American* 273 (6):80-86.

For a discussion about how to define physicalism and the mind/body debate (relevant to the discussion of Hempel's Dilemma):

- Montero, Barbara (2001). Post-physicalism. *Journal of Consciousness Studies* 8 (2):61-80.

On functionalism

For development of the theory:

- Fodor, Jerry A. (1981). The mind-body problem. *Scientific American* 244:114-25.
- Putnam, Hilary (1960). "Minds and machines." In Sidney Hook (ed.), *Dimensions of Minds*. New York University Press,. pp. 138–164.

For criticisms of the theory:

- Block, Ned (1978). Troubles with functionalism. *Minnesota Studies in the Philosophy of Science* 9:261-325.

***

Author's Note: In writing this guide, I have occasionally drawn lightly on content from my *Philosophy of Mind: The Basics* (Routledge 2020).