What is the mind and who has one?

Gregory Johnson

1. Introduction

You have a mind. This you know. But let's consider two other cases.

> Jeff is walking home through the woods near his family's farm in northern Mississippi. Off in the distance he notices a weird glow. As he approaches it, he sees a large object in the middle of a clearing. Suddenly a hatch on the object swings open. An instant later a strange looking creature jumps out.
>
> 'Whoa,' Jeff thinks, 'that's an alien.'
>
> The creature walks over to Jeff and through a series of gestures indicates that he needs directions to some other planet. Jeff tries to explain that he doesn't know the planet, and even if he did, he wouldn't know how to get there. The alien nods and returns to his craft.

Does this alien have a mind? Can it think and understand?

> Jeff continues on his way home. A few minutes later he hears crashing in the woods behind him. Someone, or something, is coming toward him. Suddenly, he sees a robot sprinting in his direction. He's scared, but there is no time to react. The robot stops as he gets to Jeff and holds out his hand. Jeff looks. He sees that the robot is holding a wallet. Jeff stares at the wallet, but doesn't reach for it. "Oh wait," the robot says, "wrong wallet." He pulls out a different wallet, which Jeff recognizes as his. He must have left it somewhere earlier. Jeff takes the wallet and thanks the robot. The robot nods and asks how to get to Highway 145. Jeff points in the direction of the road, and the robot begins walking in that direction.

Does this robot have a mind? Can it think and understand?

How should we decide if these two creatures, the alien and the robot, have minds? We might notice, first, that the robot is not a living creature. But what does it mean to be alive? The robot burns energy, can move around, and can react to its environment. It doesn't, however, need or consume nutrients. It also doesn't grow, and it can't reproduce. The alien, on the other hand, we're likely to assume, does all of those things. But needing nutrients, growing, and reproducing don't seem especially relevant to having a mind—after all, plants, which don't have minds, need nutrients, grow, and reproduce. If anything, moving around and reacting appropriately to the environment, which the robot can do, seems more indicative of having a mind.

A second issue that might concern us is what the alien and the robot are made of. The "brain" of the robot is a computer, which means it's made of various metals, perhaps some plastic, and, importantly, silicon chips. In contrast, your brain is composed of mostly hydrogen, oxygen, and carbon. Sodium, potassium, calcium, and chloride, although present in small amounts, have an especially important role transmitting electrical signals throughout your brain and the rest of your nervous system. As for the alien, we don't know. Its brain could be silicon-based like the robot's or it could be carbon-based like ours. Or it could be something else entirely—although there are only so many elements in the universe, and, as far as we know, only some of them can be used in a system that functions as a brain. But that being said, it might not matter what material the creature's brain is made of when deciding if it—or he or she—has a mind. This is a point to which we will return in section 4.

A different way to proceed is to think about the qualities that minds have. A mind has to be able to process information so that the creature can absorb stimuli from its environment and react and behave appropriately. If a creature can do that, should we say that it has a mind?

In 1950, the mathematician Alan Turing proposed the following test, now known as the *Turing test*. The test involves three participants: a judge, another person, and a computer. The judge puts questions to the person and to the computer. (So as not to reveal which is the person and which is the computer, the judge cannot see the person or the computer, and he or she types the questions and receives the answers on a screen.) After a period of questioning, the judge has to decide which is the person and which is the computer. If the computer can successfully fool the judge into believing that it is the person, then, according to the Turing test, the computer can think.

Turing predicted that by 2000, computers would be able to pass this test 30 percent of the time, but as early as 1966 a computer running a relatively simple program passed an informal Turing test and many more have since. So, if our robot and alien generally respond appropriately—in other words, respond as a person would respond—does that mean that they have minds? Many, although not all, philosophers, psychologists, and cognitive scientists think that it does.

A related issue that we might want to consider—particularly with respect to the robot—is how independent its thought is from that of its creator. The calculator on my phone, for instance, is only going to produce the results that it has been programed to generate. It can't contemplate an unsolved math problem and then produce a proof for it. Similarly, many other programs can only produce the relatively limited set of outputs that are determined by their programing. But the day when that was all that computer programs could do has passed. There are now—as we all know—artificial intelligence programs that can not only generate new information and hold sophisticated conversations but can perform tasks—such as translating documents and completing proofs of previously unsolved math problems—that their creators cannot.

A second important feature of minds is consciousness. *Consciousness* can mean different things. Sometimes it refers to "being awake." Sometimes it means "being aware or focused." The meaning we're after, though, is the experience that occurs in one's mind. When I bite into a lemon, hear Chopin's *Funeral March*, or smell coffee, what I taste, hear, and smell, is accompanied by a particular experience. That experience is what we mean by *consciousness* or *conscious experience*. Meanwhile, for the robot, even if it responds appropriately when biting into a lemon, hearing the *Funeral March*, or smelling coffee, it doesn't have the accompanying experience. It's awake and aware, but it lacks consciousness, in this sense of consciousness. Or, at least, I'm guessing that it does. I don't know for sure.

Consciousness might seem like a good way of deciding who has a mind and who doesn't, but with it comes what philosophers call *the problem of other minds*. When thinking about which creatures are conscious, I start with myself. I know that I am a conscious creature because I can, as it were, look inward and note that I have conscious experiences. But after that, I hit a wall. I can't look inside anyone else's mind and check whether or not they have similar conscious experiences—or any conscious experiences at all. All that I can do is observe other people's behavior. In philosophical parlance, beings who look and act just like you and me, but lack consciousness, are called *zombies* (or *philosophical zombies* to differentiate them from the zombies on tv and in movies). The human beings sitting in front of me in a classroom seem similar enough to me, and so I assume that they are not zombies. But I can't check that my students are conscious beings the same way that I check a pulse or someone's height. All that I can do is assume that they are. That's the problem of other minds.

Notice that the Turing test can be used to determine if a computer can think, but it doesn't tell us anything about consciousness. As of yet, we don't have a test for consciousness, and it's not clear how we would devise one. A sufficiently intelligent creature that lacked consciousness, would, or

at least could, respond in every situation just like a creature with consciousness. If you ask a zombie whether being burned hurts, she'll say yes, and she'll pull her hand away from a flame. If you ask her if the lemon is bitter, she'll say yes and grimace when she bites into one. Never having been a conscious creature, she won't even know that she lacks consciousness.

A moment ago, I assumed that the robot lacked consciousness. That was based on the thought that my phone, my computer, the calculator in my desk, and other similar devices aren't (as far as I know) conscious. The robot's brain is made out of the same sorts of materials as the computer on my desk. It's just running a much more sophisticated program on more powerful hardware. But maybe, as the software and hardware got more complex, consciousness was introduced at some point. That can't be ruled out, but at the same time, most people's intuition is that the robot, however complex and intelligent it might be, isn't consciousness.

What about the alien? Our intuitions about whether the alien is conscious can, it seems, go either way. Even though the alien looks remarkably different than a human being, we might assume that, since it is an intelligent, living creature, it is conscious. On the other hand, it evolved in an environment unknown to us, and there is no known law of evolution mandating that cognitive abilities have to be accompanied by consciousness. So, it too could be a philosophical zombie.

2. Dualism

Investigations of the different theories about the mind typically begin with the 17th century philosopher René Descartes. In this passage from his *Discourse on the Method*, published in 1637, Descartes explains the process he used for determining the nature of the mind.

> Next I examined attentively what I was. I saw that while I could pretend that I had no body and that there was no world and no place for me to be in, I could not for all that pretend that I did not exist. I saw on the contrary that from the mere fact that I thought of doubting the truth of other things, it followed quite evidently and certainly that I existed; whereas if I had merely ceased thinking, even if everything else that I had ever imagined had been true, I should have had no reason to believe that I existed. From this I knew I was a substance whose whole essence or nature is simply to think, and which does not require any place, or depend on any material thing, in order to exist. Accordingly this 'I'—that is, the mind by which I am what I am—is entirely distinct from the body, and indeed is easier to know than the body, and would not fail to be whatever it is, even if the body did not exist.[1]

This passage encapsulates many of the central ideas in Descartes's theory of the mind. The foremost being that he—what he really is—is a mind and that the mind is "entirely distinct from the body" and "does not require any place, or depend on any material thing, in order to exist." This idea that the mind and the body are separate—two different, what he calls, *substances*—is what gives this theory its name, *dualism*. Or, to distinguish it from more recent versions of dualism, it is sometimes called *substance dualism* or *Cartesian dualism*.

Cartesian dualism is, interestingly, both easy and difficult to grasp, depending on how we look at it. Many movies have been made about two characters who swap bodies—*The Change Up* (2011), *Freaky Friday* (2003), *Vice Versa* (1988), *18 Again!* (1988), *Like Father, Like Son* (1987), and others. In these movies, the characters' brains aren't switched from one body to another. Rather, a wish is made at an inadvertent moment, and each

[1] Descartes, R. (1637). *Discourse on the Method*, part 4, pp. 32 – 33.

person—his or her mind, in other words—ends up with the other person's body. This could only happen if minds are separate and independent from our bodies. Of course, movie audiences don't usually probe the details about how the switch could happen, but the basic idea is one that we can grasp. Our minds typically inhabit our own bodies, but if somehow an exchange was made, we can conceive of a mind inhabiting another body, even as the brain stays behind.

More seriously, almost all religions have, as a core principle, the belief that when our bodies die we will continue to exist, either in an afterlife or reincarnated with a different body. Religions might call the part of us that survives death the soul, but if the soul contains our personalities, memories, habits of thought, and so forth, then it is the mind. (If it doesn't contain your memories and other cognitive qualities, then it's not a mind. But it also won't have your identity, so it won't be you who survives death.) The idea that we—that is, our minds—will exist after our bodies die is grounded in Cartesian dualism.

So, most of us are familiar with the theory that our minds might be separate from our bodies—in particular, separate from our brains. Something also seems right about Descartes's contention that "I could pretend that I had no body." It seems, at least at first glance, that my body, while important to me, is not essential for either my identity or my existence. Thinking, and a working mind, on the other hand, do seem both necessary and sufficient for my existence.

All this being said, when we probe the idea that our minds are separate from our bodies, things get murky. According to Descartes, the mind has no location and does not take up space—in other words, it's an *immaterial substance*. In contrast, our bodies, and all other objects in the world, are material objects. They exist in particular locations and, among other qualities, they have width, length, depth, and shape—what Descartes and

his contemporaries called "extension"—which means that they take up space.

Having no location is, almost by definition, impossible to imagine. After all, if something has no location, then we normally take that to mean that it doesn't exist. Similarly, not having width, length, depth, mass, or energy also suggests, in some sense, that the mind isn't really there—or, at least, it's impossible to picture or conceptualize in any of the ways that we do for everything else in the world. Furthermore, if our minds are immaterial, then our thoughts are also immaterial. There should be a difference between having one thought, or five thoughts, or 1,000 thoughts. But counting anything requires that there be objects or events that exist in some location and which can then be counted. Plus, more of something should take up a greater amount of space than fewer of the same kind of thing. But that can't apply to immaterial thoughts, which makes them quite mysterious.

Moreover, although it might seem that our minds are separate from our bodies, it's equally obvious that our brain and mind are intimately connected. Phineas Gage is one of the most well-known cases of damage to the brain affecting the mind. Gage was a railroad foreman, and in 1848, while working on a construction project, he had a serious accident with a tamping iron—a 43-inch-long iron rod that was pointed at one end and used for packing gunpowder into holes drilled into rock. As Gage was tamping down some gunpowder, he was distracted and dropped the rod, which created a spark when it hit the side of the rock. The spark ignited the gunpowder, and the pointed end of the rod was sent through his left cheek, behind his left eye, through the frontal lobe of his brain, and out the top of his skull. Remarkably he survived, and after a period of convalescence, he seemed to have recovered. But, as his physician, Dr. Harlow, recounted,

> His contractors, who regarded him as the most efficient and capable foreman in their employ previous to his injury,

> considered the change in his mind so marked that they could not give him his place again. The equilibrium or balance, so to speak, between his intellectual faculties and animal propensities, seems to have been destroyed. He is fitful, irreverent, indulging at times in the grossest profanity (which was not previously his custom), manifesting but little deference for his fellows, impatient of restraint or advice when it conflicts with his desires . . . Previous to his injury, though untrained in the schools, he possessed a well-balanced mind, and was looked upon by those who knew him as a shrewd, smart business man, very energetic and persistent in executing all his plans of operation. In this regard his mind was radically changed, so decidedly that his friends and acquaintances said he was "no longer Gage."[2]

There are many other examples of damage to the brain affecting a person's mind and cognitive abilities. But even those of us who haven't had any part of our brain removed still have first-hand experiences that suggest that the mind is, in one way or another, located in the brain. The most straightforward evidence for this comes from receiving a blow to the head. If the mind really was separate from the brain, then being knocked "unconscious" wouldn't have any effect on the mind. The mind might be somewhat restrained by a limp body and a bruised brain, but it would be as clear and functional as always. Similarly, if the mind were really an immaterial substance, then alcohol and recreational drugs would be unable to have any effect on our thinking and judgment. But clearly they do. Also, brain scans reveal the activity in our brains when we are engaged in cognitive tasks. If those activities were happening in an immaterial mind instead of in the brain, then they wouldn't be captured by positron

[2] Harlow, J. (1868). Recovery from the passage of an iron bar through the head. *Publications of the Massachusetts Medical Society*, 2, pp. 339 – 340.

emission tomography (PET), functional magnetic resonance imagining (fMRI), or any other kind of neuroimaging.

3. Problems for Cartesian dualism

Problems with Descartes' theory were apparent to his contemporaries, and by the twentieth century, dualism—at least Descartes' version of it—came to be viewed by most philosophers and scientists as an untenable theory.

The main problem concerns this idea that minds take up no space and have no location. If that is so, it's a remarkable fact that my mind only ever causes my body to react and behave. If my mind is not located near my body (because it has no location), then my mind could, it seems, just as well cause another person's body to go to the kitchen and get a beer, text my friends, or turn off the alarm and continue sleeping. But, of course, outside of movies, that never happens. Without being able to refer to the mind's location, there doesn't seem to be any way to explain the pairing of my mind and my body.

Two other ways of explaining the problem with an immaterial mind focus on the interactions between the mind and the brain. If I have a thought about reaching for a book, that thought, perhaps along with some other mental states, will cause my arm to reach toward the book. Somehow the thought has to set in motion a causal chain of events that starts in my mind and reaches, eventual, the muscles in my hand. But how can an immaterial mind interact with a physical body? The closest that Descartes came to answering this question was to suggest where it might happen: in the pineal gland near the base of the brain. (He chose this structure because there is only one pineal gland on the center line of the brain. Most other brain structures—e.g., the hippocampus, amygdala, and temporal lobe—occur in pairs, with one in each hemisphere of the brain.) But stating where the interaction between the mind and the brain might happen not only contradicts the thesis that mind has no location, it doesn't address the

fundamental difficulty that this theory faces, which is *how* the interaction happens.

One way of getting more precise about this problem is to invoke a fundamental principle of physics: the law of conservation of energy. According to this law, energy cannot be created or destroyed and the amount of energy in a closed system remains constant. Since the universe is a closed system, this law tells us that energy cannot be introduced into the universe or removed from it. According to Descartes's theory, however, when my immaterial mind (which doesn't contain any energy) causes activity in my brain, new energy is introduced into some part of my brain—which thus violates the law of the conservation of energy.

A second, perhaps simpler, way of explaining the problem is just to think about how an immaterial mind could trigger activity in the brain. According to Descartes, minds have no mass or energy or any other physical qualities. If that is so, then there is no way for a mind to "get a grip" on anything physical. It has no qualities that will allow it to push or pull or otherwise set in motion activity in the brain. And even if it did, since an immaterial mind contains no energy, it has no energy to transfer to the brain to trigger activity there. This is not a new criticism. It was pointed out to Descartes by, among others, Princess Elisabeth of Bohemia in 1643. In a letter to Descartes, she wrote,

> So I ask you please to tell me how the soul of a human being (it being only a thinking substance) can determine the bodily spirits, in order to bring about voluntary actions.[3] For it seems that all determination of movement happens through [*a*] the impulsion of the thing moved, [*b*] by the manner in which it is pushed by that which moves it, or else [*c*] by the particular qualities and shape of

[3] 'Bodily spirits' refers to something like human physiology. Princess Elizabeth is not using 'spirit' in the immaterial sense.

> the surface of the latter. Physical contact is required for the first two conditions, extension for the third. You entirely exclude the one [extension] from the notion you have of the soul, and the other [physical contact] appears to me incompatible with an immaterial thing.[4]

In a sense, this isn't a very deep problem. It presents itself as soon as we begin thinking about immaterial minds. For all of the intuitive appeal of Descartes' theory of the mind, it conflicts with some of the basic things that we know about ourselves and the world.

In response to those conflicts, beginning in the 19th century, dualism was largely replaced by *monism*. Whereas dualism, with respect to the mind, is the view that there are two kinds of substances, mental substance and physical substance. Monism holds that there is only one kind of substance. Minds, bodies, and everything else in the universe are all made of matter.

4. Functionalism and cognitive psychology

Monism—the idea that the universe is composed of only one substance, matter—is really a category of theories. Perhaps surprisingly, especially given the path taken by biology and its evident successes during the 19th and 20th centuries, explaining the mind as activity in the brain has been a peripheral view until relatively recently. Instead, one of the earliest, prominent versions of monism, *behaviorism* (which will be familiar to anyone who remembers his or her Introduction to Psychology course) explains behavior in terms of the agent's environment, history, and learning. This solves the problem of explaining the mind by replacing mental states with tendencies or dispositions to behave in certain ways given the circumstances. The mind, as a thing inside the head, doesn't exist

[4] Princess Elisabeth to Descartes, May 6, 1643.

in this theory. But at the same time, behaviorism has trouble explaining anything but the simplest behavior. We will, therefore, turn to the version of monism that dominated the second half of the twentieth century both in philosophy and psychology.

Before we launch ourselves into this new theory, let me emphasize that it—especially the philosophical version—tries to be consistent with how we normally think about the mind. It is built around beliefs, desires, intentions, thoughts, ideas, memories, emotions, and our other mental states, and it doesn't wish to or try to say anything about the brain.

Philosophy's contribution to this theory began with the philosopher Hilary Putnam's observation that a mental state such as pain can be experienced by very different kinds of creatures. His examples were mammals, reptiles, octopuses (which are a type of mollusk), and aliens. The first three have, here on earth, taken different evolutionary paths, and so their brains are not that similar. (Of course, the brain of a cat and the brain of a primate are not that similar either, but since they are both mammals and share a relatively recent evolutionary history—having diverged less than 100 million years ago—their brains are more similar to each other than either is to a reptile or a mollusk.) Still, mammals, reptiles, and octopuses can all experience pain. And an alien will have yet another type of brain but can still, presumably, experience pain.

Putnam's response to this observation was to suggest that pain and all other mental states should not—and, in fact, could not—be defined as cellular or molecular or chemical states of the brain. Rather they should be defined in terms of how they *function*. Pain is not "c-fibers firing" (to use a popular example in the philosophical literature). Rather, it is the mental state that causes me to say "ouch" and to pull back from the stimulus causing the pain.

Meanwhile, around the same time, psychologists in the emerging field of cognitive psychology (and in conjunction with researchers working in

linguistics and AI) began modeling the mind as a system that processes information. Atkinson and Shiffrin's memory model is one prominent example (figure 1). In this model, information from the environment passes through a multi-components process. These interactions of stimuli and previously stored information, then, generate behavior. Notice that Atkison and Shiffrin's model explains this part of the mind in terms of processing, storing, and manipulating information, and it doesn't refer to the brain at all.

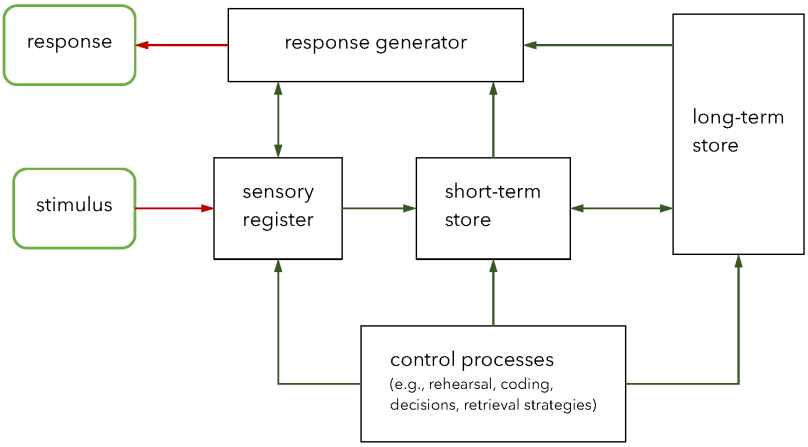

figure 1. Atkinson and Shiffrin's memory model (1968). Each component in this model is defined in terms of its role or function in this system.

Before going any further, let's think a little more about the difference between *function* and *structure*. Sitting on my desk is a pen that is mostly made of plastic. The cylinder is clear, the cap is blue, inside the cylinder is a thin tube filled with blue ink, and, at the tip of the pen, is a small ball made of tungsten carbide. Those features of the pen—a plastic cylinder a little over 5 inches long and a quarter of an inch in diameter, a small ball at one end, and so forth—are structural features. Without them, the pen wouldn't exist. But what makes a pen a pen is the *function* (or the *task* or the *job*) that

it performs, not its specific structural features. It has only one function: *facilitating the manual application of ink to a surface*, but various structures can perform this function. Instead of plastic, a pen can be made of metal, reed, or a large feather. Instead of a quarter inch in diameter, it can be wider or narrower. Instead of a ball on the end it, it can have a nib (as fountain pens do), a felt tip, or the sharpened end of a feather. But whatever its structure, as long as it performs the correct function, it's a pen. This insight, that certain things are defined in terms of their function, is the core idea for this theory of the mind, which is called, appropriately enough, *functionalism*.

mental states

By *mental states*, we mean, for instance, beliefs, desires (i.e., wants), thoughts, ideas (although thoughts and ideas may be the same as beliefs), intentions, sensations, and emotions. Most mental states, although not all, have content. For instance, my *belief that today is Thursday* is a belief that has the content *today is Thursday*. Similarly, my *desire that it snow this weekend* is a desire with the content *it snow this weekend*.

Functionalism holds that mental states are functional states. It is still a version of monism, though, and so it agrees that—just as a pen has to be instantiated in some physical object—mental states are instantiated in the brain. But, according to functionalism, we don't have to—in fact, it would be wrong to—define mental states as particular states of the brain. That gives functionalism a certain appeal. First, because it is a version of monism, it doesn't have any of the problems that dualism encountered. Second, it allows us to characterize the mind in a way that is very familiar to us. I feel (a mental state!) as though I have beliefs, desires, hopes, fears, and so forth. Furthermore, those mental states, just as Descartes said, seem to define who I am. I might be disappointed if the best theory of the mind told me that the mind is really just a series of neurons firing in the brain. Some people would be more than just disappointed. The philosopher Jerry

Fodor, who along with Putnam was instrumental in developing functionalism, says at one point,

> if it isn't literally true that my wanting is causally responsible for my reaching, and my itching is causally responsible for my scratching, and my believing is causally responsible for my saying . . . if none of that is literally true, then practically everything I believe about anything is false and it's the end of the world.[5]

But luckily for Fodor, functionalism tells us that beliefs, wants and desires, sensations (for instance, itching), emotions, and so forth are real and have scientific credibility.

Let's look more closely at how functionalism describes a simple part of one person's mind—let's say that it's my mind.

> I look at the clock and see that it is 6:00 pm. Seeing that it is 6:00 pm causes **the belief that it is 6:00 pm** which causes the **thought that it's time to stop working** and the **desire for a beer**. I already have the **belief that there is a beer in the refrigerator**, and so **the belief that there is a beer in the refrigerator** plus my **desire for a beer** cause me to get up and walk toward the kitchen.

These are the mental states in this small story about my mind:

> *the belief that it is 6:00 pm*
> *the thought that it's time to stop working*
> *the desire for a beer*
> *the belief that there is a beer in the refrigerator*

According to functionalism, this kind of description of how these mental states are causally related is how we define them. So, what is a desire according to functionalism? The *desire for a beer*, at least for me,

---

[5] Fodor, J. (1989). Making mind matter more. *Philosophical Topics*, *17*, p. 77.

is the mental state that is caused by *the belief that it is 6:00 pm* and causes this action: *walking into the kitchen*. That's a functional characterization of my desire for a beer. (Similarly, if we turn back to Atkinson and Shiffrin's model, we find that, for instance, the short-term store is the component that (*a*) takes information from the sensory register and the long-term store, (*b*) manipulates it, and (*c*) delivers it to the long-term store or response generator. That is what it does, and, as far as this model is concerned, what it does establishes what it is.)

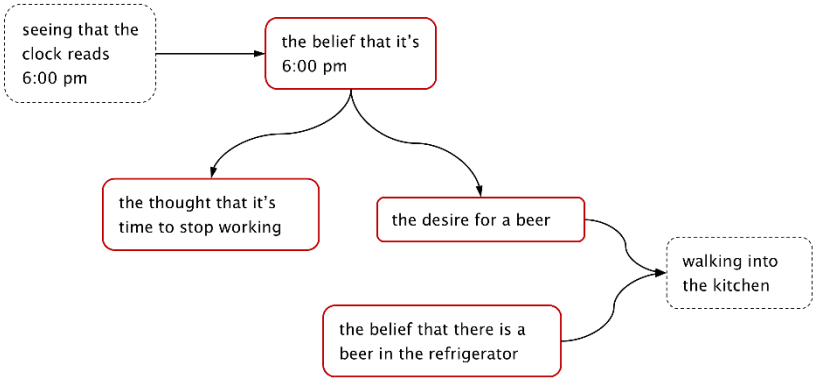

figure 2. Since what really matter according to functionalism are the causal interactions between stimuli in the environment, the mental states, and behaviors, we can create a diagram much like Atkinson and Shiffrin's.

Of course, to be made complete, this partial description of the mind would have to include all of the mental states that can cause the *desire for a beer* and all of the other mental states and actions that this desire causes. The full story for this desire and for all my other mental states is going to get quite complicated, but if we wanted to do the work, functionalism provides the framework for explaining the entire mind. Since, each mental state is defined by what causes it and what it causes, all that is needed for a complete description of the mind is a

description of every mental states' causal interactions with inputs from the environment, other mental states, and our reactions and behaviors.

5. Functionalism: Consequences and a problem

It's not a coincidence that functionalism and cognitive psychology developed and gained moment at the same time that electronic computers were becoming widely used. Functionalism is often explained by analogy with computer programs, which are also functionally described processes for generating outputs in response to inputs. Given this theory of the mind, then, we have a straightforward answer to the question Can a computer or a robot have a mind? The answer is yes. If the mind is just a series of functionally defined internal states, then not only can a computer have a mind, but our minds are essentially just programs.

This seems to successfully explain many aspects of the mind, but a significant problem remains. This problem turns on our primary reason for being skeptical that the robot has a mind, namely, our intuition that it does not have conscious experiences. It can, perhaps, get angry for the right reasons and display angry behavior, but most of us would still believe that it can't *feel* angry.

Before diving into conscious experiences, it must be emphasized that, according to functionalism (and cognitive psychology) a robot can have many legitimate mental states. To see why this is, recall that earlier I said that the mental state that is caused by the belief that it is 6:00 pm and causes walking to the kitchen is *the desire for a beer*. We can diagram these interactions, with arrows indicating 'causes', this way:

the belief that it is 6:00 pm → ***desire for a beer*** → *walking into the kitchen*

Similarly, the mental state that is caused by seeing that it is 6:00 pm and causes the desire for a beer is *the belief that it is 6:00 pm*. We can diagram those interactions this way:

seeing that it is 6:00 pm → ***the belief that it is 6:00 pm*** → *desire for a beer*.

The mental state in the middle would still be the same mental state if I had called it anything else or simply labeled it *x*. For instance, in this process:

seeing that it is 6:00 pm → *x* → *desire of a beer*

x is still caused by the same perception, and it still causes the same desire. According to functionalism, that's all that there is to *the belief that it is 6:00 pm*, or whatever we want to call it.

Functionalism embraces the implication that a robot can have all of the parts of this process: *seeing that it is 6:00 pm* (which is just a perception), *the belief that it is 6:00 pm* (which is just the mental state that is caused by that perception), *the desire for a beer* (which is a mental state caused by that belief), and *walking into the kitchen* (which is an action caused by that desire). It may seem a little odd to say that a robot can have *the belief that it is 6:00 pm* or *the desire for a beer*, but functionalism may be right that those mental states are nothing more than how they function in this process. If that's so, then a robot can have these mental states in its robot mind.

Now, consider the following. Let's say that as I'm entering the kitchen to get my beer, I hit my elbow on the door frame. This causes pain, which causes me to utter "ouch!" The mental state here is pain. It is caused by hitting my elbow against the door frame, and it causes the utterance "ouch!" Again, we can diagram the process this way:

hitting elbow → ***pain*** → "*ouch!*"

But unlike *the belief that it is 6:00 pm*, for pain, there seems to be more to the mental state than just what causes it and what it then causes. There is also, as we said earlier, a certain kind of experience that accompanies this mental state. A robot could have a mental state that is caused by hitting its elbow on a door frame and which causes it to say "ouch!" But our intuition is that the robot isn't going to have the experience of pain, or any experience at all, for that matter.

The problem, then, for functionalism is that this theory doesn't have an obvious way of characterizing conscious experience. Philosophers and cognitive scientists have worked to correct this by modifying the theory, and interesting progress has been made. But it is still far from clear how a theory that describes the mind with boxes and arrows (e.g., as in figures 1 and 2) can explain the feeling of pain, the taste of coffee, or the experience of listening to Chopin's *Funeral March*. One conclusion that we might draw here is that functionalism just isn't equipped to explain consciousness. Nevertheless, many philosophers, psychologists, and cognitive scientists still consider functionalism a viable theory. In recent years, however, two other theories about the mind have gained momentum.

6. Reductionism: The mind is the brain

When we reject dualism, the most obvious way to develop a theory of the mind is to investigate the brain. After all, everyone who accepts monism, agrees that the mind is in some way related to the brain. Functionalism devised a way of accepting monism, while largely ignoring the brain, but maybe that was a mistake. An alternative to functionalism is a theory that claims that the mind is nothing more than activity in the brain. Because this theory is, in many ways, a response to functionalism, it's called *reductionism*. According to functionalism, the mind is a functional system that can be implemented in a human brain, a robot brain, or an alien brain. So, in a sense, functionalism placed the mind at a higher, more abstract level than the brain itself. Reductionism, then, *reduces* the mind to the brain.

Before going any further, let's review some of the resources that reductionism has at its disposal. The brain is composed of two types of cells, neurons and glial cells. Neurons are generally given the most attention because they transmit the electrical signals that carry information throughout the brain. Glial cells, which actually outnumber neurons,

perform supportive roles. There are many types of processes in the brain that underlie our cognitive abilities, but, by way of example, consider just one. One neuron excites another by releasing a neurotransmitter such as glutamate, dopamine, or serotonin into the small space between the two neurons. The neurotransmitter migrates to the second neuron and binds to receptors molecules in that cell's membrane. The binding of the neurotransmitter opens channels that allow positively charged ions—for instance, positively charged sodium ions—to flow into the neuron. If enough positive charge enters, then the neuron will generate an action potential, which allows it, by the same mechanism, to excite other neurons.

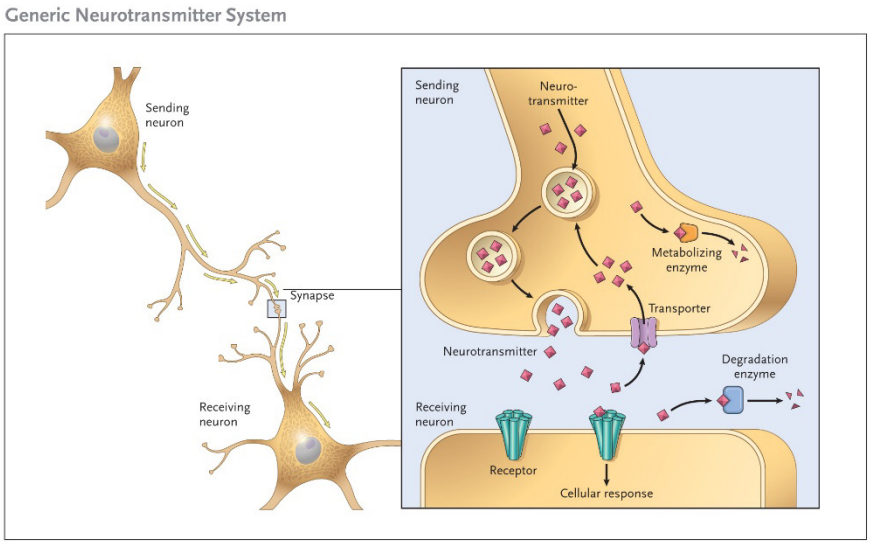

figure 3. (Left) One neuron sending a signal to a second neuron. When an action potential arrives at the pre-synaptic terminal, neurotransmitter is released and then binds to receptors on the post-synaptic neuron. This allows positively charged ions (not shown) to enter the second neuron, which will cause that neuron to generate an action potential.
(Image from Wikimedia Commons; File: Generic Neurotransmitter System.jpg; https://commons.wikimedia.org/wiki/File:Generic_Neurotransmitter_System.jpg; by NIDA(NIH); this work is in the public domain in the United States.)

Now, if dualism must be rejected, the idea that the mind should be explained as activity in the brain makes sense. After all, most processes explained by the natural sciences focus on specific physical systems. For instance, explaining the human immune system starts with the activities of white blood cells and proteins that bind to antigens. Fermentation is the process during which enzymes transform glucose or another sugar into ethanol and carbon dioxide. Photosynthesis is the process by which plants and some other organisms transform carbon dioxide, water, and photons from the sun into carbohydrates and oxygen. And there are, of course, many, many other examples.

It seems reasonable, then, to assume that the mind just is the activities of neurons, neurotransmitters, positively charged sodium ions, and the like. But when we take that step, we're giving up a lot. No longer will the mind be comprised of beliefs, desires, thoughts, intentions, sensations, and emotions. (Of course, we can still use those terms in our everyday discourse, but, if reductionism prevails, the correct scientific theory of the mind won't include those terms.) There are two or three ways in which beliefs, desires, and the rest can get replaced. It might be that specific mental states turn out to be precise activities in the brain. For instance, my *belief that I live in Mississippi* might be the activity of a particular set of neurons in the temporal lobe of my brain. Once these kinds of identifications are made, reductionism will dispense with beliefs, desires, thoughts and so forth because—so the theory claims—the neurobiological activities will more correctly and accurately explain our minds.

On the other hand, the beliefs, desires, sensations, and intentions that we think we have might not directly correlate with specific activities in the brain. In this scenario, reductionism will—if it turns out to be the correct way of explaining the mind—not only explain our minds as neurobiological activity but also reveal that everything that we thought we knew about the mind was entirely wrong. This is similar to the way in

which people once thought that witches controlled people's behavior. The concept *witch* really did, at one time, have something like scientific credibility. But eventually scientific explanations dropped *witch* and explained errant behavior in other ways. Beliefs, desires, thoughts, and so forth might be concepts like *witch* that not only get dropped from scientific discourse but are eventually understood to have been incorrect concepts. (A third possibility lies between directly locating beliefs, desires, and other mental states in the brain and eliminating them altogether.)

Reductionism—or neuroscience generally—has had a lot of success, and it has a lot of potential. But, with respect to consciousness, reductionism doesn't, at present, provide any more answers than functionalism. Neurons fire, neurotransmitters are released and migrate to nearby neurons thereby setting them into action. None of that explains consciousness, and, one might think, identical neurobiological and neurochemical processes could occur in zombies.

The reductionists response is that, unlike functionalism where we have a pretty complete idea of what this theory can offer, there is still much that we have to learn about how the brain works. Just because we don't know right now how the brain produces conscious experience, doesn't mean that we won't eventually figure it out. A useful analogy is with science's eventual ability to explain life. For millennia, it was thought that mere mechanical processes could not explain how or why certain assemblages of matter are alive. Consequently, philosophers and scientists adopted *vitalism*, the theory that living creatures contain a life force. Now we know that organisms are alive in virtue of the processes that occur inside cells, and the notion of a life force has been dropped. According to reductionists, the lesson we should take from this is that, although right now we can't picture how the brain produces conscious experience—just as it was once impossible to picture how a creature made of mere matter could be alive—we should still be confident that a more complete neuroscience will

provide an explanation. The neuroscientist Christof Koch, who has been investigating how the brain gives rise to consciousness since the early 1990s, urges us to press ahead with neuroscientific investigations:

> Many scholars have argued that the exact nature of this relationship [between the brain and conscious experience] will remain a central puzzle of human existence, without an adequate reductionistic, scientific explanation. However, similar sentiments have been expressed in the past for the problem of seeking to understand life or to determine what material the stars are made of. Thus it is best to put this question [about whether it will remain a central puzzle of human existence] aside for the moment and not be taken in by defeatist arguments.[6]

7. Dualism again

The other direction that has been taken in response to the failure—or perceived failure—of functionalism is a return to dualism, most prominently by the philosopher David Chalmers. Chalmers is motivated by two concerns. First, functionalism and reductionism can both describe processes. For instance, the example that began with the perception of a clock reading 6:00 pm and ended with entering the kitchen with the intention of getting a beer is one such process. Others are storing and retrieving information from long term memory, and neurons interacting.

Psychology, neuroscience, biology, and chemistry are equipped to explain processes. But conscious experience isn't a process. It's a feature that accompanies certain processes, and, it seems, those processes can be fully explained without explaining consciousness. The resources that we have for explaining processes appear to be useless for explaining the

[6] Koch, C., "The Neurobiology of Consciousness" in Gazzaniga (ed.), *The Cognitive Neurosciences* (Cambridge, MA: MIT Press, 2009), pp. 1137 – 1138.

conscious experience of pain, biting into a lemon, smelling coffee, or anything else.

The second idea that motivates Chalmers is that philosophical zombies—beings who are molecule-for-molecule identical to you and me but lack consciousness—are, in principle, possible. (Which is not to say that he thinks that there are any philosophical zombies; just that it is not impossible that there could be.) If there could be a being exactly like you but which lacked consciousness, it would look like you and respond exactly as you do in every situation. But, for zombie-you, everything would be dark inside. Since you and zombie-you would be molecule-for-molecule identical, Chalmers concludes that consciousness must be something extra, and we are back in the realm of dualism. But Chalmers version of dualism is very different than Descartes's.

Chalmers' response to the two issues just described is to propose that consciousness is a fundamental feature of the universe. To understand what this means, consider *temperature* for a moment. Temperature, it turns out, is mean molecular energy—that is, an average measure of how fast molecules and other particles are moving. Since temperature is explained by reference to the actions of molecules, atoms, and ions, it is not a fundamental feature of the universe. On the other hand, properties such as electromagnetic charge, mass, and space-time cannot be explained by other entities or properties. These are *fundamental properties*, and, in the end, physics just treats them as brute facts. Chalmers' proposal is that consciousness be added to this list.

Although fundamental properties are not explained in terms of other entities or processes, they are explained by laws or principles. For example, Newton's second law of motion, force = mass × acceleration, doesn't tell us what mass is or why it exists, but it does give us a rule that mass follows. Similarly, according to Chalmers, the proper way to explain consciousness

is to discover laws—he calls them *psychophysical principles*—that govern consciousness and its relationship with physical processes.

While Chalmers' theory is a version of dualism, it is different from Descartes' theory in two important ways. First, as you might have noticed, Chalmers is attempting to integrate consciousness into a scientifically informed view of the world. There's no mysterious mental substance here. As he describes it,

> This position qualifies as a variety of dualism, as it postulates basic properties over and above the properties invoked by physics. But it is an innocent version of dualism, entirely compatible with the scientific view of the world. Nothing in this approach contradicts anything in physical theory; we simply need to add further *bridging* principles to explain how experience arises from physical processes. There is nothing particularly spiritual or mystical about this theory. Its overall shape is like that of a physical theory, with a few fundamental entities connected by fundamental laws. It expands the ontology slightly, to be sure, but Maxwell did the same thing [when he postulated that electromagnetic charge and electromagnetic forces are fundamental properties of the universe]. Indeed, the overall structure of this position is entirely naturalistic, allowing that ultimately the universe comes down to a network of basic entities obeying simple laws, and allowing that there may ultimately be a theory of consciousness cast in terms of such laws. If the position is to have a name, a good choice might be *naturalistic dualism*. (1995, p. 210)[7]

[7] Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*, p. 210.

Chalmers has added to the set of fundamental features of the universe to include consciousness, which, in a way, just makes his project an expansion of physics.

He does, however, recognize that the physics we already have describes a closed system. So, while physical process may *give rise* to conscious experience (and Chalmers' psychophysical principles will explain when and how that happens), consciousness will not have causal effects on physical objects or processes; in particular, it will not have any effect on the physical processes in our brains. If Chalmers is right, this means that conscious experience doesn't have any effect on our other mental or brain states or on our behavior. Consciousness is, in a way, like a shadow. A car driving down the street has causal powers. It can, in the worst case, crash into something. But the car's shadow is inert. Once it's created, the shadow has no effect on anything else.

Perhaps, the most interesting feature of Chalmers' theory, though, is that if consciousness is a basic property like mass, charge, or space-time, then it's presumably a property that occurs throughout the universe, not just in certain parts of our brains. Chalmers suggests that consciousness occurs in physical systems that carry information. Hence, all information bearing systems—from the human brain to thermometers and sundials—could, to varying degrees, be conscious. Among many other consequences, this means that the alien and the robot with which we started are both conscious creatures.

8. Where have we ended up?

From Descartes to Chalmers, we've come full circle. Although reductionism is a promising theory, and many people haven't given up on functionalism, finishing with dualism underlines how twisty and turny the task of explaining the mind can be. We start with some

ideas and data and then do our best to bite the bullet and follow where they lead.

We shouldn't, however, overlook the progress that has been made. According to monism, as well as Chalmers' "naturalistic dualism," our minds, just like everything else in the universe, participate in mechanical processes and obey the laws of physics. That tells us a lot, and it has implications for questions about free will and whether we continue to exist after our bodies die—which are interesting questions, but ones we'll save for later.