Introduction to Philosophy Reader

Contents

Beliefs and other mental states

Gregory Johnson

The mind is a frequent topic of discussion in philosophy, and since it is so important, philosophers have developed a relatively precise way of talking about it. This is sometimes a little bit different than how people ordinarily think about the mind, so here is a quick introduction.

The mind is composed of mental states (or psychological states), and the most commonly invoked types of mental states are beliefs and desires. In everyday discourse, we often think of our beliefs as being intimately linked with our values, for instance, as with our religious or political beliefs. Philosophers agree that we have those kinds of beliefs, but beliefs are much more pervasive (and often mundane) than just our so-called "deeply held beliefs." Also, in everyday discourse (and perhaps in contrast to what I just said), people generally put beliefs on a continuum with knowledge—beliefs are less certain and knowledge is more certain. That's not, however, how philosophers treat the relationship between belief and knowledge.

First, a belief is any sentence (or as philosophers say, any proposition) that a person takes to be true. That person might be wrong, and the belief may be false, but that doesn't matter for its status as a belief. A standard example of a belief, then, is this:

The belief that I live in Mississippi.

As far as I am aware, I live in Mississippi, and so that is one of my beliefs. (And the proper way to express any belief is with "the belief that" followed by the proposition.) I might also, incorrectly, have this belief:

The belief that Mississippi is next to Georgia.

Although this belief is false, it's just as much a belief as my *belief that I live in Mississippi*, which is true.

We can, of course, qualify our beliefs, and so sometimes instead of having,

the belief that it is raining,

I might have

the belief that it is probably raining.

But let's leave aside beliefs that are about probable states of affairs. Any of the thousands (or millions) of sentences that you take to be true are beliefs that you have. Everything from beliefs about your parents' names, where you are right now, and the weather to beliefs about your purpose in life and whether Jesus of Nazareth rose from the dead.

Next, how are beliefs related to knowledge? The belief is the mental state, but sometimes we can also give a belief the title "knowledge." There are different theories about what is required to have knowledge; the traditional one requires that (a) the person who has the belief also has justification for holding it and (b) the belief is actually true.

For instance, let's say that someone posts on Facebook that he just saw three aliens leaving a Walmart in Springview, Nebraska. Nicole sees the post, and she immediately acquires this belief:

the belief that there are, or recently were, three aliens in Springview, Nebraska.

But although she has this belief, she does not have justification for it. After all, she is basing her belief about what would be an astounding event on one Facebook post without seeking corroboration from a more reliable source. (She, no doubt, thinks that her belief is justified. She wouldn't have acquired it if she didn't. But nevertheless, objectively speaking, she doesn't have justification for her belief.)

Here's another example. Dr. Cook's office is next door to mine, and right now, I have this belief:

The belief that Dr. Cook is in her office.

My justification for this belief is that I hear her voice coming from her office—and that, objectively speaking, is sufficient justification. But "she is in her office" still has to be true for my belief to count as knowledge. How justification and truth are different is a little too complicated to get into here, but let's just say that it is true that she is in her office. Then, according to this definition of knowledge, I know that Dr. Cook is her office.

On the other hand, let's imagine that, for some reason, someone is playing a recording of her voice in her office and she is somewhere else. I still have justification for my belief (that is, I still hear her voice coming from her office), but now my belief is false. Thus, now I do not *know* that Dr. Cook is in her office. So, beliefs can be justified or not justified and they can be true or false. But however that goes, the mental state is always just a belief. Then, sometimes the sentence that the belief is about will also be something that the person knows.

Desires are the other most commonly referred to type of mental state—and no, we do not mean just sexual desires. Basically, a desire is anything that a person wants. So, for example, I might have either (or both) of these desires:

the desire that it snow

or

the desire for a burrito.

Here, the content of the mental state is not a full sentence. It's just an outcome (or state of affairs) that I want: for it to snow or for a burrito.

Beliefs and desires are front and center, but there are many more types of mental states. For instance, there are intentions, thoughts and ideas (although thoughts and ideas may be the same as beliefs), emotions, doubts, and memories. All of these, like beliefs and desires, are mental states with some sort of content—that is, there is something, expressible in words, that the mental state is about. For instance, I *intend to send the email*

or I *hope that it doesn't rain on Saturday*. Other mental states, for example, sensations such as pain or hunger, don't have content. There are different kinds of pain (some of which can be described very vividly) and there are causes of pain, but the pain itself is just pain. It doesn't have content, at least not content that is expressed with words.

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.

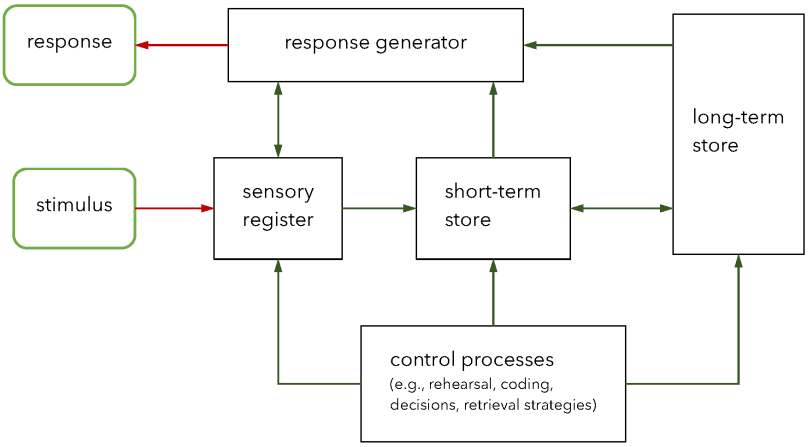


figure 1. Atkinson and Shiffrin's memory model (1968). Each component in this model is defined in terms of its role or function in this system.

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.)

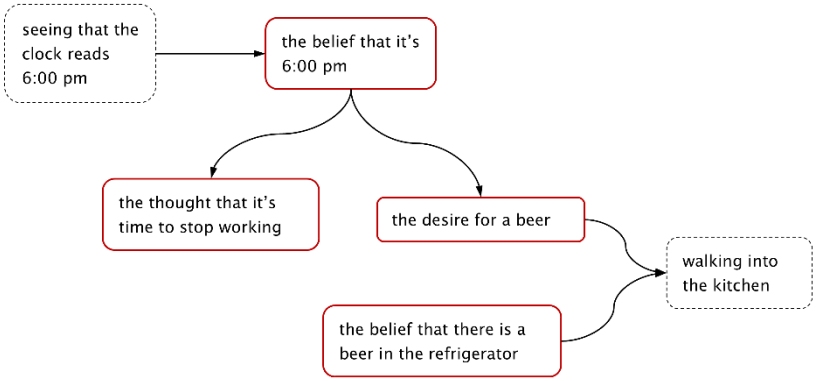


figure 2. Since what really matter according to functionalism are the causal interactions between stimuli in the environment, the mental states, and behaviors, we can create a diagram much like Atkinson and Shiffrin's.

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.

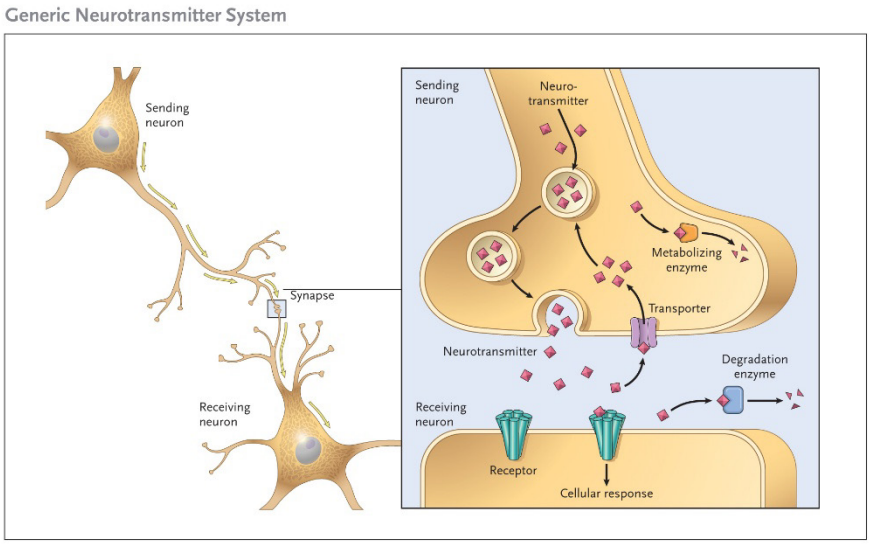


figure 3. (Left) One neuron sending a signal to a second neuron. When an action potential arrives at the pre-synaptic terminal, neurotransmitter is released and then binds to receptors on the post-synaptic neuron. This allows positively charged ions (not shown) to enter the second neuron, which will 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.

Could I have taken the other road?

Libertarianism versus Determinism

Gregory Johnson

Robert Frost's "The Road Not Taken" begins,

> Two roads diverged in a yellow wood,
> And sorry I could not travel both
> And be one traveler, long I stood

After a bit of pondering, the narrator finishes his account this way:

> Oh, I kept the first for another day!
> Yet knowing how way leads on to way,
> I doubted if I should ever come back.
>
> I shall be telling this with a sigh
> Somewhere ages and ages hence:
> Two roads diverged in a wood, and I—
> I took the one less traveled by,
> And that has made all the difference.

The intervening lines complicate the interpretation of the poem a bit, but we can all relate to the situation described here. Two options—important or not so important—present themselves to us, we deliberate, select one of them, and then act. That we could have chosen the other option seems obvious. But could we really have done so?

The free will debate seeks to answer this question. *Will*, in the sense that it is used here, is the part of us that directs our deliberate actions. And *free* means that, when give two or more options, either one can be selected. (A broader sense of *free*—the one that we use to describe, for example, not being imprisoned or tied to a chair—is not the issue here.) If our actions are not free, then either (1) they are determined by prior events, which means that, in each instance, they could not have been done differently, or (2) they

are random, which means that, although they are not determined, they also aren't guided by our wills. (While we are clarifying terminology, we can also note that, usually—although not always—*an action is caused* and *an action is determined* have the same meaning, and, when they do, they the result is the same: the action could not have been done differently.)

Returning to the question 'Could we really have chosen the other option?', on initial reflection, it surely seems to you that you could have. As we will see, however, matters are not that simple. The response that probably seems unbelievable, that we couldn't have, actually turns out to be the stronger position. Let's see how we get there.

1. Determinism and libertarianism

The two central theories about the will are *determinism* and *libertarianism*. According to determinism, we do not have free wills. The central idea that underwrites this theory is *the principle of universal causality*, which states that every event, including every human action, is caused by an earlier event or events in accordance with the laws of physics. These earlier events can include brain or mental activity, and so the immediate causes of our actions will normally be neurobiological or psychological. But those neurobiological and psychological events will themselves have been caused by the events that surround us as we go through life or, perhaps, by genetic or other biological factors. One way or another, however, these various events determine our actions. As a consequence, if we had a complete knowledge about a person and his or her environment, as well as a complete knowledge of the relevant laws of physics, genetics, biology, neuroscience, and psychology, then, according to determinism, we would know with certainty which actions this person would take. As Henry Thomas Buckle put it in the 19th century,

> If, for example, I am intimately acquainted with the character of any person, I can frequently tell how he will act under some given

> circumstances. Should I fail in this prediction, I must ascribe my error not to the arbitrary and capricious freedom of his will, nor to any supernatural pre-arrangement, for of neither of these things have we the slightest proof; but I must be content to suppose either that I had been misinformed as to some of the circumstances in which he was placed, or else that I had not sufficiently studied the ordinary operations of his mind. If, however, I were capable of correct reasoning, and if, at the same time, I had a complete knowledge both of his disposition and of all the events by which he was surrounded, I should be able to foresee the line of conduct which, in consequence of those events, he would adopt.[1]

Of course, we never have this kind of complete knowledge of another person, and we don't have a complete enough understanding of how the human mind works. (Although if you have known someone really well for a long time, you might notice that it is often possible to predict his or her behavior). But not being able to predict another person's behavior perfectly doesn't detract from Buckle's assertion that, in principle, these predictions can be accurately made because every action is caused by a person's "disposition" and the "events by which he was surrounded."

On the other side of the debate, libertarianism is the theory that we do have free wills.[2] This theory maintains that some of the time—although not always—we act freely. It can still be that sometimes, or maybe even often,

---

[1] Pp. 18 – 19 in Buckle, H. T. (1872). *History of Civilization in England*, vol. 1.

[2] A possible point of confusion is the name *libertarianism*, which this theory shares with the political movement and party. Both have adopted the name because it is derived from the Latin word for free, but otherwise they have nothing in common and shouldn't be confused or conflated.

our actions are determined by our biology, or our habits, or our environment, but *some of the time* our actions are not determined. In those cases, at a certain moment in time, and with all prior conditions remaining the same, a person can do either action *A* or action *B*.

The narrator in Frost's poem chose one of the two roads. According to determinism, this person, at that moment in time, could not have taken the other road. Some aspect of his mind—an intention, a desire, an urge—caused him to select the road that he did. Hence, given that he had that intention, desire, or urge, and not a different one, his action could not have been different. Of course, if the narrator returns to that fork, he may very well take the other road, but at this later time he will, in a variety of ways, be a different person. In contrast, libertarianism maintains that the narrator could have, at that moment, taken the other road. Hence, although the narrator has certain beliefs, desires, and urges, they don't cause or determine one specific action.

2. The evidence

Obviously, the reason why most people believe that they have a free will is because, often, when we are faced with two or more options, we feel as though we can do either one. We consider, choose, and act, but as we do, it seems to be within our power to have acted differently. This, as compelling as it might seem at first glance, is not a very strong argument for libertarianism. As Ledger Wood explains, it just amounts to this:

P1. I feel myself free.

C. Therefore, I am free.[3]

But we can feel lots of things that don't mesh with reality. I may feel that I am an NBA-level basketball player, but that feeling, obviously, doesn't

[3] P. 388 in Wood, L. (1941). "The Free-Will Controversy." *Philosophy*, 16: 386-397.

make me an NBA-level basketball player. What I need is other, independent evidence to corroborate my feeling.

Looking for evidence to support my feeling, however, quickly takes us to determinism. We can't know for sure if the principle of universal causality holds everywhere in the universe, but all of the evidence that we have points to it being true. Right now, I am seated at a desk in my office. When I look around this room, I am certain that every object was placed—that is, *caused to be*—in its present location. Similarly, when I look out the window, I am confident that every tree, building, car, and so forth got to where it is by way of a causal process, and those causal processes all obeyed the laws of physics. Nothing appeared uncaused, and everything is exactly where it should be according to the laws of physics.

As the philosopher Louis Pojman aptly puts it, "We cannot easily imagine an uncaused event taking place in ordinary life" (p. 399). Consider, for instance, this account of a car accident:

> One day you read a news headline about a one car crash that occurred not far from where you live. You read on. The car was totaled, although luckily no one was hurt. In an unusual twist, however, the state and local police report that the crash had no cause. The driver did not do anything to cause it. The car itself didn't malfunction in any way. And it wasn't caused by road conditions, the weather, or any of the other vehicles on the road at the time. It just happened. According to the official police statement, "This is one of those rare cases in which a vehicular accident has no cause. Despite the seriousness of the crash, there was no event preceding the accident that caused it."

Pause for a moment to consider whether you can imagine a car crash that doesn't have a cause. One that just spontaneous happens.

The story continues.

> Despite the conclusion of the state and local police, however, the insurance claims adjustor assigned to the case insists that the investigation isn't over yet. "I am convinced," she says, "that this accident had a cause—the driver made a miscalculation, perhaps, or some part of the vehicle failed at an inopportune moment." And little bit later in the article, she states emphatically, "there isn't going to be an insurance settlement until the cause of the crash is found." That makes sense, you think. The crash must have had a cause. The police just haven't figured out what it is yet.

This story illustrates that we experience and understand the world through the lens of universal causality. And it's not just that some of the time we expect an event to have a causal explanation—for example, when there is a car crash or a new book is sitting on my desk. *All* of the time, we expect events to have causes. While discussing Immanuel Kant's explanation of why we believe the principle of the universal causality, Pojman comments,

> Our mental construction demands that we read all experience in the light of universal causation. . . . [W]e cannot understand experience except by means of causal explanation. (p. 401)

This is, perhaps, a more sophisticated point than the libertarian's argument that 'I feel free; therefore, I am free,' but needing the principle of universal causality to understand the world doesn't thereby make the principle true. Moreover, our belief that every event has a cause conflicts with our belief that we have free wills. Both cannot be true, yet almost all of us readily accept both.

Moving beyond this stalemate brings us back to the observation that every event or state of affairs that each of us—scientists and non-scientists alike—has encountered has had a cause. We haven't observed every event in the universe, but collectively, we've observed quite a number of individual events. And every single one, or at least every one reported by a reliable source, has had a cause. Thus, from observing *this event has a cause*,

this event has a cause, *this event has a cause*, *this event has a cause*, and so on, trillions and trillions of times, we conclude that *therefore, every event has a cause*. We cannot be certain that this conclusion is true, but it's as close to certain as can be.

3. Libertarianism and actions

Since it seems to us that we usually do have the ability to choose among multiple options right before we act, we might think that an analysis of how we choose our actions would demonstrate the strengths of the libertarianism position. It is surprisingly difficult, however, to give an account of choosing actions that is consistent with this theory. To begin, let's consider what the libertarian does not want in a description of an allegedly free action. First, the decision to perform the action cannot be determined by prior events, including the agent's other mental states. It has to be possible that the action could have been done differently, and so the selection of the action can't have a cause that determines what it will be. Second, although the action cannot be determined by prior events, it also should not be random or arbitrary. When a person has the option to do action A or action B, whichever one she does can't be decided by a coin flip or some similar random procedure inside her head.

3.1 Uncaused events

So then, how does the libertarian describe the process that produces free actions? One possibility is that the process that produces an action begins with an uncaused event. Let's say that I am considering two options for the Thanksgiving holiday: (1) I can visit my sister in North Carolina and spend Thanksgiving with her family or (2) I can go to London with some friends. And let's also say that, in the end, I visit my sister in North Carolina. This version of libertarianism would maintain that my decision to go to my

sister's home in North Carolina was uncaused. It is what some libertarians call a *basic mental action*.

This explanation satisfies our first criterion: the decision, being uncaused, was not determined by any earlier events. Libertarians don't deny that I have many beliefs about my two options: what the trip will be like if I go to North Carolina and what it will be like if I go to London; how much I want to make each trip; how important travel time and costs are to me; and so forth. According to this account, however, none of these beliefs and other mental states cause—or force or push or tip—my decision. After all, the decision was uncaused.

At the same time, this account fails to satisfy our second criterion. If the decision just happens, if it's spontaneous, then we can't point to any reason why I am going to North Carolina instead of London. Of course, in this case it might seem that, even if I am randomly assigned one of these two options, either one will still appear to make sense. But if the decision really is spontaneous—and unmoored from my beliefs, desires, and other mental states—then, apparently, I could arrive at any decision. According to this account, I could just as well end up deciding to travel to Winnipeg or Santiago. Hence, we have to conclude, that according to this version of libertarianism, my decision would be random.

3.2 Caused by the agent

So far, we have used *caused* and *determined* interchangeably. *Caused* in this sense means *caused while following the laws of physics* (or any other laws of nature that we might want to invoke). If one billiard ball hits another and sends the second one into the corner pocket, it's clear that, given the laws of physics, the second billiard ball's location in the corner pocket was caused and it was determined. In other words, as soon as the pool cue hit the first billiard ball, the final location of the second one was set. The philosopher Roderick Chisholm, however, proposes *non-deterministically*

caused events. If an action is caused, but caused non-deterministically, then (unlike in the billiard ball example) it could have, with the same cause, turned out differently.

Using an example from Aristotle's *Physics* about a man moving a stone with a stick, Chisholm explains,

> We may say that the hand was moved by the man, but we may also say that the motion of the hand was caused by the motion of certain muscles; and we may say that the motion of the muscles was caused by certain events that took place within the brain. But some event, and presumably one of those that took place within the brain, was caused by the agent and not by any other events.[4]

The precise event that Chisolm is unsure about, however—the one that was "caused by the agent" somewhere in the brain—is the very one that needs an explanation. Although we don't know everything about how the brain works, we know a lot, and we know that there is no little person in there somewhere pulling levers and turning dials: at one moment, pressing these neurons into service, and at another moment, coaxing other neurons into action. Pondering how we can make sense of an agent—or what we might call *the self*—causing events in the brain can get muddled quickly. Instead, let's turn to our two criteria.

The agent in Chisholm's account causes activity in the brain, but nothing causes the agent to act one way or another. Hence, although the activity in the brain is caused (by the agent), it is not determined. Let's say that it is *brain activity A* that causes the man to move his hand so that stick moves the stone, while *brain activity B* would cause the man to drop the stick and pick up a beer. Since nothing forces the agent to initiate *brain*

[4] P. 8 in Chisholm, R. (1964). "Human Freedom and the Self."

activity A instead of *brain activity B*, either one could happen. Hence, the criterion that the action not be determined is satisfied.

At the same time, as you might have foreseen, this account won't be able to satisfy the second criterion: that the action isn't random. Before directly addressing that issue, we might wonder if there is really a difference between this account and the previous one that invoked uncaused basic mental events. Chisholm, anticipating this objection, offers the following explanation—where A refers to the brain activity that causes the hand to move the stick.

> The only answer, I think, can be this: that the difference between the man's causing A, on the one hand, and the event A just happening, on the other, lies in the fact that, in the first case but not the second, the event A *was* caused and was caused by the man. There was a brain event A; the agent did, in fact, cause the brain event; but there was nothing that he did to cause it. (1964, p. 10)

Chisholm takes this tactic to preserve the idea that the event has a cause. It is not supposed to be *in*deterministic—that is, "happening so to speak out of the blue" (p. 7).

But whether it's the case that nothing caused the brain activity or it's the case that nothing caused the agent to initiate the brain activity, our concern is why one motion was made with the hand instead of another. And to that end, the same problem that we discussed for uncaused basic mental events applies here as well. If nothing causes the agent to initiate *brain event A* instead of *brain event B* (or, if we want to put it in terms of mental states, if nothing causes the agent to initiate the decision to move the stone with the stick versus the decision to pick up a beer), then the agent does not have any reason for doing one or the other. Putting the same point in a different way, let's assume that there are reasons for doing both actions: moving the stone with the stick and dropping the stick and

picking up a beer. If, however, these reasons have no influence or impact on whichever chain of events the agent sets in motion, then whatever the agent does has to be initiated by a mental flip of a coin.

4. Determinism and actions

You might realize at this point that there is an inherent tension in the libertarian position. On the one hand, this theory holds that some of our actions are not caused by earlier events (including mental states). But on the other hand, if the decision to do an action is not caused, then it's spontaneous and random, which is not how anyone—libertarians or non-libertarians—wants to explain our actions. Determinism fairs much better here. Consider this example.

> I have a class to teach at 10:00 am. I have the belief that the class starts at 10:00 am, I have the desire to be in the classroom a few minutes before it starts, and I have the belief that I will be there on time if I leave my home by 9:25 am. These mental states cause my action: leaving at 9:23 and heading to campus.

We can say that I *chose* to go to campus or *decided* to go to campus, but, given that I had those mental states and not other ones (and given that there were no other extenuating circumstances), it doesn't seem that I could have acted differently. If I had the belief that my class began at 10:00 am and the desire to be there, but, yet, I stayed home or went somewhere else, we wouldn't say that I was acting freely. We would say that I was acting oddly or, perhaps, psychotically. Hence, counterintuitively perhaps, for our actions to make sense and be meaningful, we need them to be determined by our mental states.

We might also consider a case where we are very aware of two competing options, and we have reasons for doing both.

> Let's say that I have the option to visit my sister in Virginia or my sister in North Carolina. I would like to do both, but that's not possible. So, what causes my action? I have beliefs about when I last saw each sister, when, if not on this trip, I will be able to see each one, how much time and effort it will take to get to each of their homes, and so forth. I also know how much I want to spend a few days with each one and her family and how important travel time and costs are to me. Let's say that, ultimately all of these beliefs and desires weigh in favor of going to Virginia, and so I travel there.

In this case, I might just decide to flip a coin, but if I don't, then my beliefs have to be considered and my desires have to be weighed. But my mental states—the strongest desires and the best reasons—will still cause my action.

When we first encounter it, determinism seems cold and impersonal, but the world would be much colder and more impersonal if all of my beliefs and desires pointed me toward one action, but somehow, I found myself doing the other one.

5. Can you believe that the wall is red?

To complete the picture for determinism, it must also be the case that we do not choose our mental states. If we can, then, although they cause our actions, our actions could still be free. This may seem like an opening for libertarianism, but, in fact, it's generally agreed, by both determinists and libertarians, that we don't choose our beliefs, desires, and other mental states. Beliefs, for each of us, simply record what we take to be true. You can see this by trying a simple experiment. Assuming that you are inside, you can see the color of the nearest wall. In my case, I can see that it is light blue, and that perception causes my *belief that the wall is light blue*. Can I just choose to believe that the wall is some other color, say, dark green? I can

utter the sentence, "I believe that the wall is dark green," but I can't actually have that belief because that's not the way that the world presents itself to me.

Of course, there are more complex cases, but they seem to follow the same rule. There are also instances when people change their beliefs, but those, as well, appear to follow the rule that beliefs must track the way that we think the world is.[5] Take a belief that might seem to be one that you did choose: either (*a*) the belief that Jesus of Nazareth rose from the dead, or (*b*) the belief that Jesus of Nazareth did not rise from the dead. Whichever belief you hold, you didn't acquire it in the same way that you acquired your belief about the color of the wall. Nonetheless, it's just as clear in this case that if you believe (*a*), then you can't just choose to believe (*b*), or vice versa. People do, occasionally, switch between (*a*) and (*b*), but when they do, it's because they've read or heard something relevant that causes the change. It's not because they just decided to switch beliefs. (Or if it ever were simply a switch without the person being exposed to new ideas or points of view, then, again, it would seem odd or, perhaps, a sign of psychosis. We don't actually want our beliefs to change without reasons for them to do so.)

In addition to beliefs, our actions are caused by our desires, emotions, character, habits, determination, and, perhaps, other types of mental states. These sorts of mental states don't represent information in the same way as beliefs do, but they do, in a variety of ways, push us toward one action or

[5] There are also beliefs for which, because our information is incomplete, we only have a certain degree of confidence. For instance, I might have *the belief that I probably have a meeting next week*. That doesn't really change anything, though. If my confidence level that I have a meeting next week is around 70 percent, I can't choose to believe either that I definitely do have a meeting next week or I definitely don't have a meeting next week.

another. We can see that we do not choose our desires, emotions, character, and so on, with the same test that we used for beliefs. Let's just take desires. Think of a food that some people really find appetizing, but you don't—for instance, maybe deviled eggs. Can you, all of a sudden, choose to want or desire deviled eggs? You can say that you want them. You can force yourself to eat them. And maybe if you eat them long enough, you'll find that they aren't as disgusting as you thought. But you can't just flip a switch and desire them. We have the desires that we have, apparently, because of some mixture of our experiences, upbringing, and genetics.

Hard determinism, soft determinism, and moral responsibility

Gregory Johnson

1. The argument from moral responsibility

What many people believe is the strongest argument for libertarianism doesn't try to explain how free actions are possible but instead focuses on moral responsibility. For our purposes, *being morally responsible* will just mean that we can be praised or blamed for our actions. Libertarians maintain that we can only be praised or blamed when it is the case that we could have acted differently. For instance, let's say that I am standing by a pool, see a child who appears to be drowning, but do nothing. If nothing is preventing me from jumping into the water to save the child, then, it seems, I deserve to be blamed for not acting. On the other hand, if I am, for some reason, tied to a chair and cannot move, then I do not deserve blame. The difference between the two scenarios is not hard to grasp. In the second case, although I was present while the child was drowning, I simply couldn't save him, and so if he does drown, I shouldn't be blamed for the tragic outcome.

In the same way, if determinism is true, then, in every circumstance, we couldn't have acted differently, and so we, apparently, do not ever deserve praise or blame. The question, then, is do we, in fact, sometimes deserve praise or blame? According to libertarianism, yes. If this is correct, then we have a compelling argument that determinism is false:

P1. We can only be morally responsible in those situations when we could have acted differently.

P2. According to determinism, in every situation, we could *not* have acted differently.

P3. We are morally responsible for at least some of our actions.

C1. Therefore, determinism is false.

And, then, using this conclusion as a premise, we have an argument that libertarianism is true:

P4. Determinism is false.

P5. If determinism is false, then libertarianism is true.

C2. Therefore, libertarianism is true.

Both arguments are valid, and so if premises 1 - 3 are true, then the first conclusion, *determinism is false* has to be true, and if premise 5 is true, then the final conclusion, *libertarianism is true*, has to be true. The issue, however, is whether the premises are, in fact, true. Both determinists and libertarians accept the first two premises, and so the question is whether premise 3 is true. It certainly seems as though we are morally responsible for at least some of our actions, and it's probably best to live our lives as though we are. But there isn't any evidence that we are, and there's no apparent way of generating such evidence. After all, there is no investigation that we can undertake that will demonstrate that we are creatures with "moral responsibilities."

2. Punishment

It can be difficult to know what to make of the debate at this stage. The arguments that we examined in the previous chapter favor determinism. The argument for moral responsibility appears to support libertarianism, but it really only shifts the question to whether or not we have moral responsibilities. In the second half of this chapter, we will look at a new theory, one that rejects the first premise in the moral responsibility argument. First, however, let's think about an issue that's related to moral responsibilities, namely, the relationship between libertarianism, determinism, and punishment.

Consider the following crimes committed by two elderly women in the late 1990s and early 2000s. In 1997, Helen Golay, who was 67 at the time, and Olga Rutterschmidt, who was 64, began taking out life insurance policies on a homeless man, Paul Vados. Two years later, Vados was found dead in an alley, and Golay and Ruttershchmidt recieved the payouts from the life insurance policies. In 2002 and 2003, Golay and Ruttershchmidt took out life insurance policies on another homeless man, Kenneth McDavid. He was hit by a Mercury Sable station wagon in 2005, and again, Golay and Ruttershchmidt collected on the life insurance policies.

In 2008, Golay and Rutterschmidt's killing spree came to an end when they were convicted of murdering the two men. Both women were given life sentences without the possibility of parole.

This penalty is no surprise, and there are myriad other penalties for the various infractions that people commit every day. But what exactly justifies Golay's, Rutterschmidt's, and everyone else's punishments? The government has to be able to justify the punishments that it imposes, and so how might it do so? To answer this question, we will look briefly at the two main theories of punishment: retributivism and deterrence.

2.1 Theories of punishment

Retributivism is the idea that a punishment is justified because it gives the offender what he or she deserves; in other words, the punishment is retribution for the crime. *What someone deserves* might be a little vague, but the basic idea is that the offender has committed an offense, and this event justifies a proportional punishment. Golay and Rutterschmidt might have been given a different penalty—for instance, the death penalty or, in a different time, they might have been banished from society—but a lifetime imprisonment without the possibility of being released is a penalty that they deserve.

Once it is sketched out, many people are sympathetic to retributivism, but if they are just asked what justifies punishment, more people will probably invoke something along the lines of the *deterrence theory of punishment*. According to this theory, punishments are justified because they deter or discourage future crime, either by the offender or by others who might commit similar crimes. We can also justify Golay and Rutterschmidt's punishment with this theory. Being in prison for the rest of their lives will prevent Golay and Rutterschmidt from committing any crimes in the future, and it will make other citizens who might be inclined to murder someone for insurance money think twice about it.

Those are the two most prominent theories of punishment, but there are others. One is *rehabilitation*, which is a justification for punishment that requires that the punishment be set up in such a way that the offender's behavior is reformed. (This, however, is not a justification that could be given for Golay and Rutterschmidt's punishment. They are not being locked up for the rest of their lives so that they can be rehabilitated.) Other, less central, although still important, justifications for punishment are satisfying the victims' desire for punishment, preventing vigilante action, and, in cases of imprisonment, keeping the rest of the community safe from the offender.

2.2 Determinism and punishment

A naïve view of determinism holds that, if this theory is true, it would make punishment impossible. This is clearly false. Determinism very well may be true, and punishment exists. Trying again, we might say that if determinism is true, then *justified* punishment is impossible. This is also false. If determinism is true, then we cannot use retributivism to justify punishments. If we could not have acted otherwise, then, as mentioned earlier, determinists and libertarians agree that we deserve neither praise nor blame for our actions. Taking this idea a step further, if determinism is

true, then, we not only don't deserve blame, we don't deserve punishment. But if determinism is true, we can justify punishment with the deterrence theory, as well as with the rehabilitation model, satisfying victims' desire for punishment, preventing vigilante action, or keeping society safe.

But let's focus on deterrence. Locking up Golay and Rutterschmidt will ***determine*** what their prospects for committing crimes will be in the future. Moreover, just the belief that committing this kind of crime will bring about a severe punishment—and then seeing the state follow through on that threat—will ***cause*** many other individuals to refrain from murdering anyone. (Which is not to say that other beliefs, such as *the belief that murder is wrong*, won't also cause people to refrain from committing such an offense. On the other hand, some beliefs—say, *the belief that I won't get caught*—will sometimes cause people to kill others for the insurance money despite the intended deterrence.)

The moral is that, if determinism is true, we have to give up one justification for punishment, retributivism. But determinism is perfectly consistent with deterrence, as well as with the other justifications for punishment. So, if we decide that determinism is true, we are just as justified as we ever were in locking up Golay and Rutterschmidt.

3. Determinism *and* moral responsibility?

So far, we have treated determinism as a single theory. There are, however, two different versions of it in the philosophical literature. The one that we have so far encountered is *hard determinism*. Let's now turn to *soft determinism* or what is often called *compatibilism* because it maintains that determinism and moral responsibility are compatible. Compatibilism doesn't give up the central commitment of determinism. Just as with hard determinism, according to compatibilism, every event has a cause, and so every event, including every human action, could not have been done

differently. But compatibilism attempts to make a distinction between these two categories of actions:

(*a*) Actions that, according to the theory, are free (or "free"), even though they are determined.

(*b*) Actions that, according to the theory, are determined and not free.

At the beginning of the previous chapter, I said that the issue before us was whether, when confronted with two options, we had the freedom to do either one—take the other road, visit the other sister, drop the stick and pick up a beer, or what have you. I also explained that a broader sense of *freedom*—the one that we use to describe, for example, not being imprisoned—is not the issue here. Now, however, we want to contrast not being free in the way that we have so far thought about it (i.e., we don't have free wills) with the other sense of not being free (i.e., being locked in a prison cell). The British philosopher A.J. Ayer who is credited with formulating the contemporary statement of compatibilism, asks us to consider cases where we are compelled to act a certain way.[1] There are several.

(1) Someone hypnotized me and is now directing my actions.

(2) Someone—for instance, a parent, spouse, or boss—has managed to psychologically manipulate me to such an extent that it is "physically impossible for me to go against his will."

(3) Someone is pointing a gun at me and telling me how to act. (In this case, it is conceivable that I won't follow the instructions, but, assuming that I do, we would say that I was compelled to act as I did.)

(4) I have a psychological disorder like kleptomania that causes my actions.

---

[1] Ayer, A.J., (1954). "Freedom and necessity" in *Philosophical Essays*.

Now, remember, according to Ayer, *all* of our actions are determined, but he wants to distinguish between our "normal" determined actions and ones like (1) - (4) where we are compelled to act a certain way either by some other agent or by a psychological disorder.

4. *Voluntary* (or "*voluntary*") actions and moral responsibility

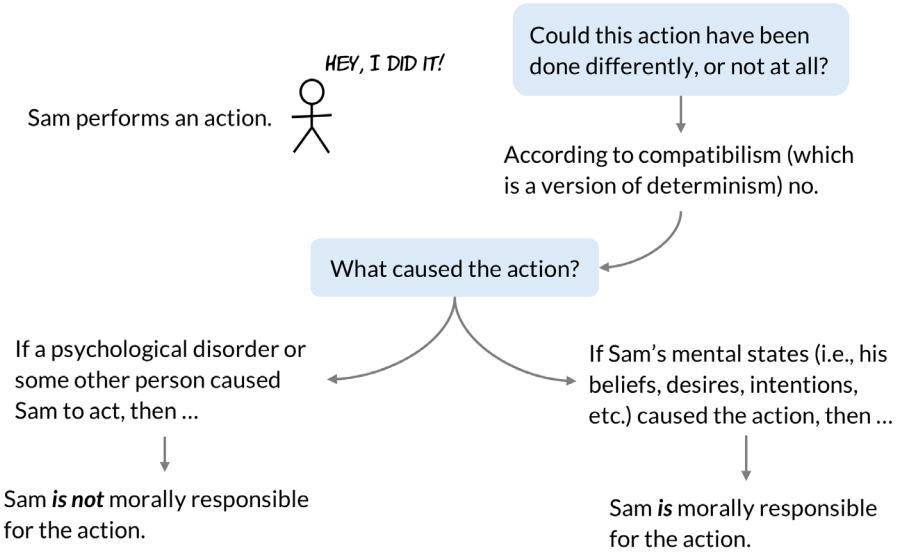
Again, when examining compatibilism, it is important to remind ourselves that this theory is a version of determinism. As such, just as with hard determinism, this theory maintains that every event, including every human action, is caused. That being said, using Ayer's term, we can call actions that are caused by our mental states *voluntary* (or "*voluntary*" if you like). Given that someone has the beliefs, desires, intentions, memories, and emotions that he has—and not different ones—his action could not have been done differently. But nonetheless, insofar as the person is not compelled to act by someone pointing a gun at him or because he has been hypnotized, the compatibilist maintains that we should call the action *voluntary*. For instance, if, as I said in the previous chapter, my beliefs and desires cause me to head to campus at 9:23 am, then, although my action is determined by my mental states, it is a voluntary action in this sense.

On the other hand, Ayer calls actions that are either caused by a person being compelled by someone else or caused by a psychological disorder *involuntary*. So, imagine that, instead of my mental states determining where I am headed when I leave home, a fugitive from justice is pointing a gun at me and directing me to drive him to the next state. In this case, my action is involuntary.

Taking this a step further, according to compatibilism, we are morally responsible for our voluntary actions. And then, naturally, we are not morally responsible for the involuntary ones. Although they could not have been done differently, since voluntary actions are caused by our mental states, we take ownership of them in ways that we don't for

involuntary actions. When I tell a lie with the desire to deceive someone, the lie is caused by that desire plus *the belief that I won't be caught* and *the belief that the lie will benefit me*. In other words, the lie is caused by my own mental states. Given that I have those mental states (and perhaps other relevant ones as well), I couldn't have not told the lie. But, at the same time, no one was holding a gun to my head or compelling me in some other way. Hence, according to the compatibilist, I am morally responsible for the lie.

In contrast, as you might anticipate, if the fugitive from justice who is pointing a gun at me tells me to lie and I do so, then I am not morally responsible for the lie.



5. The debate over compatibilism

The debate between hard determinism and libertarianism was a debate about the very nature of the universe. Namely, whether there are, or can be, events that come into existence without a cause. As far as compatibilism is concerned, that debate is settled, and determinism is correct. The compatibilist's task is to think about how we should understand ourselves

and the world once we have accepted determinism. About this, there can be much to say.

The central issue is whether we can really be justified calling some actions *voluntary* (and *free*) if determinism is correct. William James, for one, makes his distaste for compatibilism clear:

> Old-fashioned determinism was what we may call *hard* determinism. It did not shrink from such words as fatality, bondage of the will, necessitation, and the like. Nowadays, we have a *soft* determinism which abhors harsh words, and, repudiating fatality, necessity, and even predetermination, says that its real name is freedom; for freedom is only necessity understood, and bondage to the highest is identical with true freedom. Even a writer as little used to making capital out of soft words as Mr. Hodgson hesitates not to call himself a "free-will determinist." Now, all this is a quagmire of evasion under which the real issue of fact has been entirely smothered.[2]

James's lack of patience with compatibilism is understandable. If any theory requires biting the bullet and accepting where the evidence leads, it seems to be determinism. Arguing that we act freely, even though we could not have acted differently, can seem like a poor attempt to paper over the reality in which we find ourselves.

This criticism is fair, but we should also keep in mind what motivated the development of compatibilism. As we saw in the previous chapter, although libertarianism is intuitively very compelling, determinism is the much stronger position. At the same time, although there are justifications for punishment that are consistent with determinism, most people don't want to give up the basic idea of moral responsibility.

---

[2] James, W. "The dilemma of determinism."

We can reject compatibilism on the grounds that it is simply based on redefining the terms *free* and *voluntary*. But for all of that uncompromising integrity, we are left with libertarianism, which—besides having the weaknesses discussed in the previous chapter—may be just as incompatible with moral responsibility as hard determinism. Imagine that I tell a lie, but my action was not caused by *the desire to deceive someone*, *the belief that I won't be caught*, and *the belief that the lie will benefit me*. Instead, telling the lie just somehow happened spontaneously. In this case, it's far from clear how I am morally responsible for it. As Ayer explains,

> If it is a matter of pure chance that a man should act in one way rather than another, he may be free but can hardly be responsible. And indeed when a man's actions seem to us quite unpredictable, when, as we say, there is no knowing what he will do, we do not look upon him as a moral agent. We look upon him as a lunatic. (1954)

In the face of this problem, holding me morally responsible because the action was caused by my mental states, doesn't look so unappealing.

This brings us to a second problem with which the compatibilist must reckon. It is possible to make a distinction between kleptomaniacs and the typical bank robber. The former will steal objects that they don't need or value. The bank robber, meanwhile, steals something that almost every values, namely, money. But for all of our mental states (and the actions that are caused by them), we need a clear line demarcating the so-called normal and healthy ones from the ones that are not. Unfortunately, there is no such line. Hence, whereas, it is always obvious when my action is involuntary because someone is pointing a gun at me or has hypnotized me, it is not always obvious if my action is caused by my "normal" mental states or a psychological disorder. This, however, may be a problem that we are

used to encountering. It's not uncommon to find ourselves wondering if someone's action was caused by a trauma or an abusive environment. If we decide that it was, then we often don't hold the person morally responsible in the way that we otherwise might.

What is ethics anyway?

Gregory Johnson

1.

For something that pervades our lives, most people cannot give a very clear definition of ethics. That's easy to fix, however. Ethics is the study of morality. Of course, that only shifts the question to What is morality? Again, the definition is easily supplied. Actions—and maybe other things, but we'll focus on actions—are morally right or morally wrong. (Or neither, some actions are morally neutral.) If an action is morally right, then this means that we ought to do it. We might not, and we might not be punished if we don't, but we *should*. On the other hand, if the action is morally wrong, this means we should not do it. Still, when faced with a moral dilemma, knowing the definition of morality isn't really much help. Which action should I choose? And how do I know it's the right one?

One answer might be that religion and the associated religious texts tell us which actions are morally right and which ones are morally wrong. But using religion as a guide faces some problems. First, there's no consensus about which religion should be consulted. Christians will say Christianity, Hindus will say Hinduism, Jews will say Judaism, Muslims will say Islam, Buddhists will Buddhism, and on it goes; and that's not even getting into the sects that divide every religion. But even if we select one religious text—let's take the Christian Old Testament—upon examination, we find moral guidelines that don't seem quite right. There is slavery, the questionable treatment of women and children, and the command, given in *Exodus*, that "Six days shall work be done, but on the seventh day you shall have a holy Sabbath of solemn rest to the Lord; whoever does any work on it shall be put to death."

Making sense of the moral guidelines in the Old Testament can be set aside, however, because there is a deeper problem. One that was first articulated by the ancient Greek philosopher Plato in the fourth century B.C.E. Almost all of Plato's writings that we have today are dialogues, and in each the main character is Plato's slightly older contemporary Socrates. In a dialogue titled *Euthyphro*, Socrates and a priest named Euthyphro (who may or may not have been a real person) try to work out the definition of *piety*. About mid-way through the dialogue, Euthyphro proposes this definition "I would certainly say that the pious is what all the gods love, and the opposite, what all the gods hate, is the impious." Socrates, pressing for clarification, asks, "Is the pious being loved by the gods because it is pious, or is it pious because it is being loved by the gods?"

If we replace *the pious* with *morally correct actions* and *gods* with *God*, then Socrates is asking which one of these is correct:

(1) Morally correct actions are loved by God because they are morally correct.

(2) Morally correct actions are morally correct because they are loved by God.

If you can see the difference between these two, that's a good start. But let's pause for a moment to consider what's at stake here. The idea that the Old Testament (or any other religious text) tells us what is morally right and morally wrong is called the *divine command theory*. That's what's being invoked when someone says, "such-and-such is morally right because God (or the Bible) says that it is," and that's what Euthyphro has in mind with his definition. Plato is showing us that this idea can be understood in two ways. Today, when this dilemma is discussed, the two options are usually phrased this way:

(1b) God commands an action because it is moral.

(2b) An action is moral because God commands it.

The first option means that there are some morally correct actions, and God tells us to do those actions. But God didn't do anything to make those actions morally correct, and so they would still be morally correct even if God never existed. What's significant about this option is that actions are not right or wrong because God says that they are. Rather, there is some other reason why the actions are moral.

The second option states that God does make certain actions morally correct. By telling us to do certain things and not others, God thereby determines which actions are morally correct and which ones are morally wrong. Although at first glance, this looks like a powerful option, it turns out to be pretty unattractive. This is the scenario: imagine that all possible actions exist (killing, lying, telling the truth, helping the needy, and so forth), but none of them are morally right or morally wrong yet. Then God comes along and randomly (yes, randomly) chooses some that he is going to tell humans to do and some that he is going to tell humans not to do. If the Ten Commandments are to be believed, he settled on, among other things, not killing, not stealing, and not committing adultery. But he could just as well have well gone with kill, steal, and commit adultery. If he had, then killing, stealing, and adultery would all be morally correct. Of course, we might want to say that God wouldn't do that. God had reasons for choosing some actions to be moral and others to be immoral. But if God had reasons, then this option turns into the first one. With the second option, there aren't reasons; he just picked the actions.

Since it includes this element of randomness, the second option is not considered viable. But the first option tells us that the reason an action is morally correct is independent of what God says. So, morality has to have some justification other than the religious texts or God's word. The divine command theory falls short.

Still, although understanding why actions are morally right and morally wrong is an important part of doing ethics, we might think that we can still salvage something here by noting that, even if we don't have the reasons for why some actions are morally right and others are morally wrong, we can take comfort in the idea that God, with some justification or other, put various edicts in the Old Testament (or any other religious text) for us to find. We can simply proceed with his list. Unfortunately, even that isn't so simple, and the requirement that we kill people who work on the Sabbath isn't the only problem. In fact, let's set that aside, and take the commandment "thou shall not kill." Simple enough, but consider this situation described by the philosopher Bernard Williams:

> Jim finds himself in the central square of a small South American town. Tied up against the wall are a row of twenty Indians, most terrified, a few defiant, in front of them several armed men in uniform. A heavy man in a sweat-stained khaki shirt turns out to be the captain in charge and, after a good deal of questioning of Jim which establishes that he got there by accident while on a botanical expedition, explains that the Indians are a random group of the inhabitants who, after recent acts of protest against the government, are just about to be killed to remind other possible protesters of the advantages of not protesting. However, since Jim is an honored visitor from another land, the captain is happy to offer him a guest's privilege of killing one of the Indians himself. If Jim accepts, then as a special mark of the occasion, the other Indians will be let off. Of course, if Jim refuses, then there is no special occasion, and Pedro here will do what he was about to do when Jim arrived, and kill them all. Jim, with some desperate recollection of schoolboy fiction, wonders whether if he got hold of a gun, he could hold the captain, Pedro, and the rest of the soldiers to threat, but it is quite clear from the set-up that nothing

> of that kind is going to work: any attempt at that sort of thing will mean that all the Indians will be killed, and himself. The men against the wall, and the other villagers, understand the situation, and are obviously begging him to accept. What should he do?

Many people, although not all, have the strong intuition that Jim should kill one of the captives. We might recognize why Jim doesn't want to do this. We might even realize that we would have great difficulty doing it ourselves. But, nonetheless, it is what Jim *should* do. Killing in this instance is the morally correct action. Not taking the captain's offer would be morally wrong.

Now we are in the thick of it. We have a moral dilemma, and, perhaps, an idea of what the morally correct action is. But *why* is killing a person the right thing to do in this situation? Selecting an action is part of what we must do when faced with a moral dilemma, but it's only part of it. If we want to be able to say that we did the right thing, then we also have to justify our action. That is, we have to give the reasons why it was the correct action in that situation. The principle 'God commands an action because it is moral' does not provide those reasons, and so we turn to the first of the two main theories in ethics.

2. Utilitarianism

Utilitarianism is most commonly associated with the 19th century philosopher and politician John Stuart Mill. A statement of the theory is short and to the point: *the morally correct action is the one that produces the greatest amount of happiness for all involved, or all who will be affected*. Applying this theory just amounts to doing a calculation. First, we identify the different actions that can be taken and who will be affected by those actions. In the example above, Jim can either elect to kill one person or he can refuse. Jim's decision will affect, at least, the twenty captives, the other villagers, the captain, and Jim himself. Perhaps the circle can be expanded

further, but this will do. If Jim kills the one captive, then the other nineteen live. If Jim refuses, then Pedro will kill all twenty.

So, if Jim kills one person, there is much less loss of life. But that's not the only factor, there are also the villagers who, let's say, all have family and friends among the twenty captives. For some of them, Jim killing one person will cause a great deal of unhappiness. But those people will also be unhappy if Jim refuses and Pedro kills all twenty. Hence, if only one person is killed, the villagers, on the whole, will be much happier than if all twenty are executed. The captain seems as though he'll be equally content either way. That just leaves Jim.

The effect that killing or refusing to kill has on Jim counts, but it's no more or less important than the effect of these actions on anyone else. Maybe, if he kills the one person, Jim will eventually feel proud for doing something difficult and saving the lives that he could. Or maybe he'll be traumatized or feel guilty for the rest of his life. Maybe he'll be happier if he decides not to get involved. Or maybe he'll be wracked with guilt for the rest of his life if he refuses to kill. Whatever it might be is a factor, but it's not going to tip the balance, and so it really doesn't matter. If we add up all of the happiness and unhappiness that is created if Jim kills one captive and compare that to the amount of happiness and unhappiness created if Jim refuses, we find that, overall, Jim killing one person creates more happiness than if he refuses. Therefore, according to utilitarianism, killing the one person is the morally correct action.

Utilitarianism tells us what the morally correct action is, and it also tells us *why* that action is morally correct. But before thinking about the why, let's clarify one part of the theory: maximizing happiness. Sometimes simply adding up the number of people who are made happy and the number made unhappy will suffice, but that's not quite what the theory says. It's the amount of happiness or unhappiness that each action creates. Consider a different example.

> Imagine that you are a heart surgeon at a large hospital. One upcoming evening, you are not scheduled to work and so you make plans to meet six friends for dinner. They haven't seen you for many months, and you promise them that you will be there. The day of the dinner, however, every other heart surgeon at your hospital comes down with the flu and is sent home. Just as you are preparing to leave, a six-year-old girl who needs emergency heart surgery is admitted to the hospital. There is no one else who can perform this surgery, and so you must decide whether to break the promise that you made to your friends or to do the surgery. (For the sake of the example, let's say that the hospital won't compel you to do the surgery, and, for whatever reason, your friends won't ever know why you missed the dinner. Maybe it's a top-secret surgery.) The girl's family is limited to her and her parents, and they have no close friends.

If you perform the surgery, just the girl and her parents will be happy. Your six friends, meanwhile, will be unhappy. And to keep it simple, let's say that you will be equally happy either way.

Importantly, however, by performing the surgery and saving the girl's life, you will make her and her parents incredibly happy. Each one of your friend's unhappiness is far outweighed by the happiness of the girl or either one of her parents. Hence, even though fewer people are made happy than unhappy, performing the surgery will create more happiness than unhappiness. Conversely, if you join your friends for dinner, don't perform the surgery, and the girl dies, then the unhappiness that this creates, for her parents and briefly for the girl, will be enormous. Each of your friends' happiness will be tiny by comparison. Now, the girl's and her parents' unhappiness outweigh your six friends' happiness. Thus, even though performing the surgery will only make three people happy, according to utilitarianism, it is the morally correct action.

An issue that might occur to some people at this point is measurement. How do we measure and then add up happiness? There are methods for measuring happiness or, at least, for measuring the strength of people's preferences, although philosophers have noted certain problems with these methods. But often the specifics don't really matter. Everyone can see that more happiness is created if Jim kills one person than if he refrains and lets Pedro kill all twenty. Similarly, the heart surgery case is pretty straightforward. The only added twist is that if some people are made happy and others are made unhappy and, for each person, the strength the happiness or unhappiness varies, then we need to think a bit more about which sentiment outweighs the other to arrive at an answer.

Now, let's turn to the primary purpose of utilitarianism: providing a reason why Jim should kill the one captive. Happiness is valuable. In and of itself, it is a good thing. Unhappiness, pain, and suffering are not. Hence, according to utilitarianism, creating happiness and minimizing unhappiness, pain, and suffering are what give actions their moral worth. Utilitarianism takes that idea and gives us a formula for acting on it. Now, one need not accept this. There is no experiment or proof that tells us that actions that maximize happiness and minimize suffering are the only ones that have moral worth. But, nonetheless, utilitarianism supplies one justification for why some actions are morally correct and others are morally wrong. And, importantly, it seems like maximizing happiness is the right justification some of the time. The main problem for this theory is that, sometimes, utilitarianism tells us that an action is morally correct, when, to many people, it seems not to be. This is the classic example:

> A rape and murder, perhaps racially motivated, are committed in a town long beset by racial tension. You are the chief of police, and you've spent a lifetime working to make your community safe. Now this has happened, and an outbreak of violence, in which many people will probably be killed, looks likely. In turns out,

> however, that there is a homeless man in one of your jail cells. If you frame this man for the crime, there will be a quick trial, he will be found guilty, and the violence will be avoided. Besides this homeless man, no one but you and the real criminal—who will presumably remain silent—will know what you have done. What should you do?

In this case, the people affected by your decision are the homeless man, the real criminal, the people in the town, some of whom may die depending on your decision, and you, the chief of police. An outline of the calculation goes as follows. If you *don't frame* the innocent man, he won't be found guilty, but, probably, there will be an outbreak of violence and a number of people will die; there will also be longer lasting negative effects for the town. On the other hand, if you *frame* this man, he will go to prison and he may be put to death. He cares about whether this happens, but he may not have many friends or family who do. Everyone else in this community, meanwhile, will be relatively content and unharmed if he is quickly found guilty of the crimes. Either way, there's not going to be much, if any, happiness created. But, if you frame the homeless man, there will be much less unhappiness. Thus, according to utilitarianism, framing the innocent man is the morally correct action. It is what you, the chief of police, should do. The problem, though, is that, for many people, framing an innocent person seems wrong, not right, even given the consequences.

3. Kantian ethics

Our other major ethical theory is one that was developed by the 18th century philosopher Immanuel Kant. The theory is basically just a statement of what Kant called the *categorical imperative*. He formulated four versions of the categorical imperative, which he claimed were different ways of saying the same thing. The four are related, although it's not obvious that they are exactly equivalent. At any rate, this is the first

version: *act only in accordance with that maxim through which you can at the same time will that it become a universal law*.[1]

A maxim is a rule, and when using the categorical imperative, the first step is to identify the rule that you would be following if you performed some particular action. In the previous example, if you frame the homeless man, you are following this maxim: frame an innocent person. Next, with that maxim in hand, consider this: if you had the power to make this maxim a universal law, could you reasonably do it? The "reasonably" part is going to be important, but first, a universal law is, for instance, *the speed of light is 186,000 miles per second* or *energy equals mass times the speed of light squared* (i.e., $E = mc^2$). A little more informally, this is also a universal law: On earth or in any environment with a gravity similar to earth's, if you step off a ledge and nothing else interferes, you will drop down to the nearest surface. This isn't just a guideline, and it's not a law that you have the option of following or not. If you step off a ledge, you are going to drop. Similarly, (although only hypothetically) if you make *frame an innocent person* a universal law, every time you or anyone else has a chance to frame someone who is innocent, that's what will happen.

Kant thought that our ability to reason dictated morality. So the question is, if you could, would it be reasonable to make *frame an innocent person* a universal law? Kant's answer would be no, and that has nothing to do with whatever mayhem might be created or avoided by framing innocent people. Unlike utilitarianism, Kantian ethics puts no weight on the consequences of an action. Rather, it would not be reasonable to make *frame an innocent person* a universal law because someday you might be that

---

[1] The second version of the categorical imperative is equally important in contemporary ethics, but it will not be our focus. It states: "Act in such a way that you always treat humanity, whether in your own person or in the person of any other, never simply as a means, but always at the same time as an end."

innocent person who gets framed, and, naturally, you wouldn't want that to happen. Since you cannot wish for this maxim to be a universal law, it is morally wrong for you to frame the homeless man.

Conversely, both of the maxims: *search for criminals* and *do not frame innocent people* are ones that anyone could, reasonably, want to become universal laws. Hence, according to Kantian ethics, the morally correct actions are the ones that follow these maxims.

Again, we have (maybe) figured out which action is morally correct and which one is morally wrong. This also comes with a justification. All of this business about maxims and turning them into universal laws may seem odd, but it is a very clever formula for getting at an important idea. What you do in any particular situation should be what anyone else is also entitled to do. If no one else should do that action that you are considering, then you shouldn't either. Morality is holding yourself to the same rules that you want everyone else to follow.

But while the categorical imperative seems to get things right some of the time, just as with utilitarianism, there is a point at which it will conflict with many people's intuitions. This is famously illustrated by an objection to Kant's theory that was posed by the political writer Benjamin Constant in 1797. The example concerns a murderer "who has asked whether our friend, who is pursued by him, had taken refuge in our house." It seems obvious, at least to many people including Constant, that if a murderer is pursuing your friend—or pursuing anyone for that matter—you should direct the murderer away from, not toward, wherever his intended victim is hiding. But that's not what the categorical imperative tells us to do.

The maxim is *tell a lie*. Why this cannot, reasonably, be turned into a universal law is interesting. We might think that if it did become a universal law, then, just as in the example about framing an innocent person, it would put us at a disadvantage. Others would lie to us. That fits, but according to Kant, there is an even more fundamental reason why it

would not be reasonable to turn the maxim *tell a lie* into a universal law. If lying became a universal law, then it would no longer be possible to lie. No one would take anyone at their word, and so it would be impossible to effectively convey something that wasn't true. Making it a universal law would make it impossible to do what you are, right now, attempting to do. Because of the contradiction inherent in making it a universal law and, at the same time, acting on the maxim, we can't (reasonably) wish for that maxim to become a universal law. Hence, telling a lie is morally wrong. The maxim *tell the truth*, meanwhile, runs into no such problem, and so, according to the categorical imperative, telling the truth is always the morally correct action.

Nonetheless, lying to the murderer in this situation seems to be the only morally acceptable option. Even if we have a strong distaste for lying, one might think that we should overcome that distaste, and if we didn't, then we would be doing something morally wrong. But that's not what Kant thought. He stuck by the categorical imperative. Imagine, he said, you lie to the murder. But just at that moment, your friend, who had been hiding in your house, sneaks out and goes to where you, by lying, led the murder. Your well-intentioned lie only ensures that the murderer finds his victim.

Kant's response isn't too satisfying, however. If there is one thing that has to be true about ethics, it is that we make the best decisions that we can based on the facts that we have. Acting morally doesn't require us to be able to control the future, but it does require us to make reasonable inferences about what is going to happen and then act.

4. Rights

The one ethical concept that almost everyone has heard of and many people often invoke is *rights*. A natural response to the case of the sheriff thinking about framing an innocent man is to insist that the man has a right

not to be framed. Framing him infringes on his rights. As familiar as rights might be, though, stating the content of this theory, or perhaps theories, is not as simple as it is for utilitarianism or Kantian ethics. To even get started, we need to back up a bit. Earlier, I said that if an action is morally right, then it is something that you should do. You have an obligation—a moral obligation—to do it. Or, flipped around, if the action is morally wrong, then you have an obligation to refrain from doing it. Utilitarianism and Kantian ethics tell us which actions we are obligated to do and which we are obligated to refrain from doing. Rights are slightly different. If the homeless man has a right not to be framed, that doesn't impose an obligation on him. Having the right doesn't mean that *he* should keep himself from being framed. Rather, if he has the right not to be framed, then that imposes an obligation on others. It is because the man has this right that the police chief should not frame him, imprison him, and maybe let him be executed. In virtue of the man's right, it would be wrong for the police chief to act that way.

There are also different ways of talking about rights. There are *legal rights*—rights that people are granted, one way or another, by the state. There are also, maybe, *human* or *natural rights*. These are rights that people have simply because they are people. (Others besides human beings—some animals, for instance—may also have natural rights, but we'll have human beings in mind.) Thus, if the right to free speech is a human right, then all humans, everywhere and at all times, have this right. At some times in the past and in some places today, this right may not be recognized and trying to exercise the right is tricky or dangerous. But, still, everyone has it. If, on the other hand, the right to free speech is only a legal right, then in some places, people have that right, in other places they do not. In the United States, with some exceptions, we have it. In Turkmenistan, well, they're not so lucky.

One advantage of legal rights is that it is clear which ones people have. If it's written down and the government recognizes it, then the citizens have that right. But when rights are invoked with regard to whether an action is morally right or morally wrong, it's almost always human rights that we have in mind. It's a little bit of a mystery, though, how we confirm that people do, in fact, have human rights. And then, if they have them, which ones they have.

Let's assume that there are some human rights and look at another distinction: *negative rights* and *positive rights*. Earlier I said that rights impose an obligation on others. This obligation can work in two ways. On the one hand, it might just be an obligation not to interfere. If this is all that is required by the right, then it is a negative right. The standard example is the right to free speech. If I have that right, then everyone else has an obligation not to interfere with me as I speak. But that's all that others have to do. No one has to provide me with a podium or listen to what I'm saying. On the other hand, for some rights, having the right does impose an obligation on others to do something. Here the standard example is the right to an education. If children have this right, that doesn't only mean that no one should interfere with them as they get an education. It means that someone—maybe parents, maybe the community, maybe the state—has an obligation to provide them with an education. Or take the right to medical care, which is also a positive right. Despite all of the disagreements over this in the United States, we almost unanimously agree that people do have a right to some level of medical care. We don't want people to lie dying on park benches or to set their own broken bones at home. We agree that they are entitled to some amount of medical care, provided by someone.

With the distinction between positive and negative rights in mind, we can see that rights may give us an insight into what makes some actions morally correct and others morally incorrect. Take the claim that there are

human rights, but those rights are only negative rights. This is libertarianism, and, fundamentally, it is based on the idea that people should have as much freedom as possible. Our obligation is to let people do as they like, as long as they don't interfere with others. A competing idea is that there are, not only negative rights, but also some positive rights. This is based on the idea that there are some things—for instance, education and medical care—that are essential for a person to have a chance to flourish. Without those things, we would still, biologically, be human beings, but we would be mere animals or automatons. We wouldn't, fully, be persons—that is, rational, moral, autonomous agents. Hence, people have a right to those things, and someone or some group has an obligation to provide them.

The tension between the two positions is easy to see. Consider Judith Jarvis Thomson's example,

> In a small community, a child is suddenly struck with a deadly infection. There is a medicine that will fight the infection and save the child's life, but the only supply of it in this town is owned by and is in a locked box on the back porch of a woman who is away and cannot be contacted. There isn't time to obtain the medicine from somewhere else. What should be done?

The right to property is a negative right: as long as I acquire my property legally, everyone—including the state—has an obligation not to interfere with it. The right to some basic level of medical care is, as we said, a positive right: someone has an obligation to provide that care when it is needed. So here the woman's right to her property bumps up against the child's right to medical care. If there are only negative rights, then the child's right to medical care doesn't exist and it's morally wrong to take the medicine from the woman without her consent. If there are both positive and negative rights, then the child's right to medical care most likely exists, but there is still the question of whose right takes priority. In the end, all

we can say is that rights give us a means of addressing this dilemma, but if our starting point is just that there are some positive and some negative rights, then that alone will not produce a solution.

~ ≈ ~

Utilitarianism, Kantian ethics, and rights are three of the main players in the study of morality, but they are only part of it. There are variations of those theories and other theories altogether, including some that focus on the person we ought to be, not just the actions that we should or should not take. Virtue ethics, which was originally developed by the ancient Greek philosopher Aristotle, focuses on our character. According to Aristotle, our moral obligation is to cultivate virtues such as generosity, honesty, compassion, prudence, and courage. Once that's accomplished, hopefully, morally correct actions will follow. In a somewhat similar vein, in the second half of the twentieth century, a number of female philosophers developed the ethics of care, which orients morality around a more maternal perspective. Instead of rules that dictate what we should and should not do, the focus is on relationships, caring for and nurturing others, and making sacrifices for one's family or community.

5. Making moral decisions

On January 12, 2010, a 7.0 magnitude earthquake, followed almost immediately by 6.0 and 5.7 magnitude aftershocks, hit southern Haiti. The earthquake killed over 200,000 people and displaced more than a million. One of those affected was a thirty-eight-year-old Port-au-Prince resident named Nathalie LeBrun. Her house collapsed in the earthquake killing most of her extended family, but, as her luck would have it, she survived because she had checked into a hospital earlier that day. The earthquake didn't help, of course, but Nathalie's most significant medical problems were chronic conditions, severe heart failure and a related lung condition. In the week that followed the earthquake, Nathalie ended up at an

American-run field hospital where she was given oxygen, which made it possible for her to breathe normally and kept the oxygen level in her blood from falling too low.

The field hospital, however, had a very limited supply of bottled oxygen. The night after she arrived, the tank that Nathalie had been given ran out and she came close to dying. That outcome was averted when more bottled oxygen was found the next morning, and later, Nathalie was moved onto an oxygen concentrator—a device that could remove oxygen from the air and deliver it to a patient. But the situation was tenuous. The field hospital wasn't getting more bottled oxygen, and it also didn't have enough diesel, which was needed to run the generators that powered the oxygen concentrators.

After a couple of days, the field hospital's liaison officer, in consultation with the head doctor, decided that Nathalie would no longer be given oxygen, even though taking it from her meant that she would likely die. Other patients needed the oxygen during surgery, and they would then recover from their injuries. Nathalie, meanwhile, needed a constant supply of oxygen, and since she had a chronic condition, she might never be well enough to leave the hospital. The liaison officer was a captain and nurse practitioner named Patrick Kadilak. When Sheri Fink, who reported this story, asked him about Nathalie, he said,

> We're running out of oxygen. The country itself doesn't have oxygen. So, I have to make the decision, 'no, she can't have the oxygen; turn it off.' I have to look at the greater good that we can provide with the limited resources we have.

Since the American field hospital was no longer providing her with care, Nathalie had to be transferred to a Haitian hospital. The hospital to which she was being sent was unlikely to have oxygen, and so it was expected that she would probably die there.

When she arrived, she was in severe distress, but an American physician who was volunteering at the hospital, Dr. Paul Auerbach, improvised and treated her with the resources that he had: diuretics to remove the fluid from her lungs and a tank with a little bit of oxygen left in it. That stabilized her. Shortly thereafter, more fuel unexpectedly became available and Nathalie was put back on an oxygen concentrator. Several months later, with Sheri Fink's help, Nathalie traveled to the United States for surgery to correct her heart condition. Given its severity, however, that turned out not to be possible, and when she couldn't get a transplant, she died.

Treating Nathalie in the aftermath of the earthquake illustrates the tension between Kantian ethics and utilitarianism. We can reasonably wish for the maxim *save a life that is in danger* to become a universal law, and we cannot do the same for the maxim *don't save a life that is in danger*. Therefore, the categorical imperative tells us that the morally correct action is to do what we can to save Nathalie. We do the same for each patient who follows her, and, even if at some point we exhaust our resources, we've done what we could for each person as he or she was presented to us. Sheri Fink clearly leans toward Kantian ethics. When asked later about what we should take away from Nathalie's story. She answered,

> Let's not give up. The conclusion is let's not give up. It turned out there were options for this woman. It turns out that somebody was able to extend her life. Now you could very well argue that she should have died in that moment because look at all the resources that were spent. But I just feel like there was some value in her existence. There was so much value.

But at the same time, it was reasonable to believe—and likely true—that other patients would be saved if the oxygen was available for them. Given the crisis and the limited supply of oxygen, the utilitarian justification for withholding care from one person so that multiple other patients can

benefit (and in the end, benefit more than Nathalie would have) is straightforward. To maximize happiness, or at least to maximize positive outcomes, we withhold the oxygen from Nathalie and use it on the other patients who don't have chronic conditions. Now, that's the morally correct action.

So where does that leave us? These moral theories give us a procedure for determining which actions are morally correct and which are morally incorrect. It's not so easy, however, to figure out which theory is the right one. One option is to think long and hard about what, ultimately, gives actions their moral value. Maximizing happiness and minimizing suffering? Or holding ourselves to the same rules that we want everyone else to follow? Rights, virtue ethics, and the ethics of care provide us with more ideas and principles, and, for those who aren't faint of heart, there are many more ethical theories. If you come to the conclusion that one of these theories has gotten it right, you bite the bullet and stick with the theory no matter what.

Another option is to lean on our intuitions. In the example of Jim wandering into a South American village, killing the one captive seems to be the morally correct action. In the example of the sheriff in the town where a riot is imminent, not framing the innocent man seems to be the morally correct action. Or maybe not. People have different intuitions, which is one reason why we hope that an ethical theory will help us decide how to act. In the end, however, we probably don't have a perfect procedure for making moral decisions, and we haven't turned ethics into a science. But that shouldn't be taken to mean that whatever we feel like doing is thereby acceptable. The ethical theories give us insight into our moral decisions and the means to think about, discuss, and justify those decisions. Even if we selectively apply the theories, at least when we do, we understand why we think that the action we have taken is, in that situation,

the right one. And that is very different than just doing what we "feel" and refusing to give, or being unable to give, a clear justification for our action.

Recall that our original task was simply to understand what the terms *ethics* and *morality* mean. We've covered that much, and if it has turned out that understanding a little bit has only made things more complicated, then, well, we've learned that ethics isn't easy.