What Is Personal Identity?

When you die, do you cease to exist? Is death The End? According to many religious traditions, no: bodily death is a mere transition to life in heaven or some other celestial realm, or perhaps to reincarnated life in another body on Earth. And even if death is the end at present, some think that future technology might defeat it. The futurist Ray Kurzweil, for instance, has predicted that by mid-century we will be able to "upload our knowledge, memories and insights into a computer," allowing us to enjoy a kind of "virtual life" inside a computer-generated virtual reality, somewhat as depicted in the movie The Matrix.¹

Keeping our eyes on the future, suppose that the Star Trek fiction of "teletransportation" (or "teleportation") becomes fact. If you step into the transporter, your brain and body are instantly scanned, and the resulting information is beamed to your chosen destination, say Mars. On Mars, the receiving station instantly reconstructs your brain and body from new matter, exactly as it was on Earth.² The process of scanning vaporizes the original brain and body, but why should you care? Here you are on Mars, stepping out of the receiving station with a new brain and body that are just as good as the original. If you're afraid of teletransportation, that's like being afraid of flying—you should just try and get over it.

Now consider another scenario. Imagine a perfect three-dimensional (3D) photocopier, which can duplicate not just physical objects such as stones and tables, but also animals. Would you like another dog, exactly like your beloved Fido, to keep him company? No problem: place Fido on top of the photocopier, press "copy," and a perfect replica of Fido comes out of a chute on the side, barking and wagging his tail. You could even photocopy yourself, getting a perfect "identical twin."

Photocopying yourself would lead to all sorts of practical problems, of course, but being photocopied doesn't harm you in any way. Suppose, though, that the 3D

¹ See www.kurzweilai.net/live-forever-uploading-the-human-braincloser-than-you-think. Kurzweil's view in this essay is somewhat similar to Derek Parfit's (see his essay in this chapter).

² This is not quite how teletransportation works in Star Trek, but for our philosophical purposes it's better to think of teletransportation this way.
photocopy develops a fault: it produces a copy just as before, but now destroys the original in the process of scanning it. Would you photocopy yourself now? Surely not—that would be suicidal! Perhaps having a replica around after your death to write your term papers and fool your family into thinking you are still alive is some compensation, but nowhere near enough.

Hold on a moment: What's the difference between teletransportation and the malfunctioning 3D photocopy? The product of teletransportation emerges at a great distance from the transporter, unlike the product of 3D photocopying, which emerges right next to the photocopy, but there doesn't seem to be any other significant difference. If so, “teletransportation” is not a means of transportation at all. Instead, the misleadingly named “transporter” is a device that destroys the person who steps into it and creates a replica at the receiving station. Instead of saying “Beam me up, Scotty,” it would be more accurate to say “Kill me, Scotty.”

The person called “James T. Kirk” in this week’s episode has not had an eventful life as a Starfleet captain after all! If he stepped out of the transporter chamber on the starship Enterprise yesterday, he is only one day old.

The issues raised above all concern our survival. What sorts of changes can we undergo and survive? Can we survive the destruction of our bodies? And does it matter if we don’t? These questions are discussed in philosophy under the heading of “personal identity,” the topic of the selections that follow.

Survival and Identity

Personal identity is a special case of a more general topic, the survival (or, as philosophers often say, the persistence) of objects over time. It is useful to have a little background in this more general topic when discussing the specific issue of our survival over time.

Commonsense opinion holds that inorganic things (e.g., rocks, laptops, and planets) and plants and animals typically come into existence at some time and cease to exist at a later time. For example, a certain cottage might come into existence when enough beams and bricks are assembled, and cease to exist a century later when it is demolished to make room for a McMansion. A mighty oak tree began life as a tiny green shoot, or perhaps an acorn, and will end its existence when it is sawn into planks.

The cottage and the oak survive a variety of events throughout their careers. The house survives a flood, say. That is, the house existed before the flood and also existed after the flood. We can put this in terms of “identity”: the house existed before the flood, and something existed after the flood that was identical to the house. Explaining what survival amounts to in terms of identity helps to clarify the notion, but it is potentially confusing. Suppose your house burns down and you build “an identical house” in its place. That is not a situation in which your house survives—it is a situation in which your house is destroyed and a replica is built in its place.
its place. Although your original house and the replica are “identical” in the sense that they are very similar, there is another sense in which they are not identical. There are two houses in the story, and in the “numerical” (or “strict”) sense of “identity” two things are never identical. We have the strict or numerical sense in mind when we say that Lady Gaga and Stefani Germanotta are identical: we don’t just mean that Gaga and Germanotta are similar, like your smart phone and your friend’s smart phone. Gaga and Germanotta are not two at all. “They” are one and the same person, with two different names (“Lady Gaga” and “Stefani Germanotta”). Similarly, when we say “$3^2 = 9,” we are not saying that 3 and 9 are two numbers. We are instead speaking about one number with two names (“3” and “9”). Survival, then, should be defined in terms of strict or numerical identity. When we ask whether you will survive some event, our question is best understood as follows: Will there be someone around after the event who is numerically identical to you?

Houses can survive repainting and the addition of a porch; they cannot survive being reduced to ashes or having their parts scattered all at once. But what if the parts of your house are scattered and replaced over a long period of time? Imagine that your ancient family home has been lovingly repaired over the years so that now not a single brick or beam from the original construction remains—the crumbling bricks and rotted beams have been slowly replaced with period bricks and beams from architectural salvage. Has the original house survived? Not an easy question, but it’s unlikely to keep you up at night. Sentiment or pure theoretical curiosity aside, it doesn’t much matter whether this is (numerically) identical to the house your grandfather lived in, as opposed to one just like it that is built on the same spot. But from your point of view, there is at least one thing whose genuine survival seems to matter a great deal, namely you. Suppose you are told that your body will undergo some ordeal—brain surgery or teletransportation or physical death and resurrection—and that there will be someone around afterward who is like you in many ways. You might reasonably say: “That’s all very well, but will that person be me?” A theory of personal identity is designed to shed light on this question.

A “Criterion of Personal Identity”

The philosophical literature on personal identity is often structured around a search for a “criterion of personal identity.” This is an idea that can be difficult to understand, so it is worth spending some effort getting clear about it. Suppose someone—call him or her “Casey”—exists at a certain time, say on Monday. Consider someone who exists the following Friday—call him or her “Drew.” What would absolutely guarantee that Casey and Drew are numerically identical? In other words: What are sufficient conditions for Casey and Drew to be numerically identical? For example, suppose that Casey looks a lot like Drew; same hair color, same eyes, and so forth. Does that guarantee that Casey = Drew?
Now let's ask a different question: What must be the case, given that Casey and Drew are numerically identical? In other words: What are necessary conditions for Casey and Drew to be numerically identical? For example, given that Casey = Drew, must it be the case that Casey and Drew look alike?

We can write out these two suggestions for, respectively, a necessary condition and a sufficient condition a little more formally. First, the sufficient condition:

It must be the case that: if Casey on Monday looks like Drew on Friday, then Casey, who exists on Monday = Drew, who exists on Friday.

Second, the necessary condition:

It must be the case that: Casey, who exists on Monday = Drew, who exists on Friday, only if Casey on Monday looks like Drew on Friday.

If we generalize these two theses to all people and all times and combine them together, we get a thesis stating necessary and sufficient conditions for personal identity, which we can express as follows:

*The Physical Appearance Criterion*: It must be the case that: \( A \), who exists at \( t_1 = B \), who exists at \( t_2 \), if and only if \( A \) at \( t_1 \) physically resembles \( B \) at \( t_2 \).

This is a *criterion of personal identity*. In general, a criterion of personal identity is a statement of the following form (leaving the “It must be the case that” implicit, and abbreviating “if and only if” as “iff”):

\[ A, \text{ who exists at } t_1 = B, \text{ who exists at } t_2 \text{ iff } \]

*The Physical Appearance Criterion* has the right form, but it is obviously wrong. Sameness of appearance is not *sufficient* for numerical identity: “identical twins” may look alike, but they are not literally one and the same person. Nor is sameness of appearance *necessary* for numerical identity. Drew on Friday may look very different from Casey on Monday; but if Casey underwent cosmetic surgery on Tuesday, they may be the same person nonetheless.

The challenge is to fill in the blank so as to render the resulting statement true. Of course, there's an easy way to do that. Just replace the blank with “\( A = B \)! Although the resulting statement is undeniably true, it is of absolutely no help in answering questions about our survival—whether we can survive bodily death, for instance. What we want is an *informative* replacement for the blank—one that does not presuppose the notion of identity that we are trying to understand. We should not assume that there is such a replacement, but that has certainly not stopped philosophers from trying.
that Casey and
conditions for
t'Casey = Drew,
sary condition
: condition:

Friday, then

who exists on

combine them
us for personal

who exists at
as B at t_p

unal identity is
that' implicit,

'sously wrong
cidual twins"
ay look very
ig surgery on
ng statement
with "A = B"
ely no help in
dly death, for
one that does
id. We should
ot stopped

One suggestion that might occur to you on reading Descartes's Meditation II (see Chapter 7 of this anthology) is this:

The Soul Criterion: A, who exists at t_i = B, who exists at t_p, iff A's immaterial soul at t_i = B's immaterial soul at t_p.

The Soul Criterion highlights something important about a criterion of personal identity. A criterion of personal identity is supposed to state how things must be if (and only if) A = B. It need not be an account of how we tell that A (say, someone we met last week) is identical to B (someone before us right now). We can often tell that A = B because A's physical appearance is the same as B's. But the corresponding criterion of identity, based on “same physical appearance,” is mistaken. Conversely, it is no strike against the Soul Criterion that we do not find out that A = B by discovering that A has the same immaterial soul as B.

The Soul Criterion is defended in the selection by Richard Swinburne (who also emphasizes the point started in the previous paragraph). An obvious objection to the Soul Criterion is that the existence of immaterial souls is extremely controversial: if there are such things, you won't be learning about them in Psychology 101. In the selection by John Locke, you can find a subtler objection.

While the existence of our souls is debatable, the existence of our bodies seems plain enough. So a natural replacement for the Soul Criterion is this:

The Bodily Criterion: A, who exists at t_i = B, who exists at t_p, iff A's body at t_i = B's body at t_p.

Is this right? Imagine that you swap brains with someone else—call him or her “Emerson.” (The selection by Swinburne discusses an example of this sort.) Emerson's brain is transplanted into your body, and your brain is transplanted into Emerson's. Suppose that medical technology is sufficiently well advanced so that after the operation there are two people, alive and well: one with Emerson's old body, and one with yours. Where are you? Where your original body is, or where Emerson's body is?

Many people think that in cases of this sort, you follow your brain and not your body. So perhaps a better suggestion is:

The Brain Criterion: A, who exists at t_i = B, who exists at t_p, iff A's brain at t_i = B's brain at t_p.

But again, there are objections. Imagine that you have some brain disease that will eventually destroy all your brain cells if left unchecked. Suppose that medical technology has advanced to the point where we can gradually replace each brain
cell with a prosthetic artificial cell (a tiny device containing a silicon chip). At the end of the process, your brain has completely vanished, replaced by a prosthetic brain. If the artificial cells are sufficiently good, won't you still be around, thankful that the new technology has saved your life? But if you can survive the loss of your original brain, then the Brain Criterion is incorrect.

A quite different idea, proposed by Locke, is that our survival does not consist in the survival of a thing, like a soul, body, or brain, but rather in psychological connections across time. And Locke had a specific suggestion for what sorts of psychological connections are important; namely, those provided by memory. "[A]s far as this consciousness can be extended backwards to any past action or thought," he writes in his Essay Concerning Human Understanding, "so far reaches the identity of that person" (see section 9 of Locke's essay on page 507). This suggests the following:

*The Memory Criterion:* A, who exists at $t_a = B$, who exists at $t_b$ if B can remember 'at $t_a$ (some of) the experiences of A at $t_a$.

The Memory Criterion implies that amnesia—perfect and total amnesia—amounts to death, and this may seem implausible. Faced with a grim choice between death and amnesia, a self-interested person might well choose amnesia as the lesser of two evils. "At least I'll still be around to start again," he might think. His friends and family might have a similar thought. And if this is right, then the Memory Criterion is unacceptable.

Many philosophers have tried to develop Locke's basic idea in a way that avoids this objection, among others. The most famous neo-Lockean theory of personal identity comes from Derek Parfit in his 1984 book *Reasons and Persons.* In the selection from that book given in this chapter, Parfit argues for a "psychological criterion" of personal identity—one that emphasizes various forms of psychological continuity over time, not just memory. But he also argues, astonishingly, that "personal identity is not what matters." According to Parfit, granted that the Psychological Criterion is correct, in some circumstances your survival should be of no concern to you! Compare the case in which the teletransporter functions properly and the case in which it malfunctions, creating two duplicates of the original body instead of one. In the latter case, Parfit argues, the original person does not survive. But from the point of view of the person himself, this case is not relevantly different from the case in which the machine works properly. In both cases, there is someone around after the event who is psychologically (though not physically) continuous with the original. The only difference is the *number* of such people. Parfit argues that no one should care very much about this numerical fact, and hence that no one should care very much about whether he will exist in the future. All that matters is that people will exist in the future who are psychologically continuous with us as we now are. Since genuine identity and psychological continuity normally go together in our experience, we confuse them and mistakenly think that personal identity is what matters when we worry about our own survival. According to Parfit, a correct account of the nature of personal identity can disabuse us of this error.
As the reader will have noticed, philosophical discussions of personal identity often invoke wild science-fiction thought experiments. We are asked to imagine a bizarre scenario involving brain swapping or teleportation and to consult our "intuitions" about survival and identity. It is not hard to see why this method should be necessary. A philosophical theory of personal identity is meant to apply to every possible case. So we must consider far-out cases in order to assess our theories. And yet the method has its pitfalls. The problem is not that our judgments about far-out cases are uncertain, though that may be so. As Bernard Williams argues in his essay in this chapter, the problem is that a single case may elicit different intuitions when presented in different ways. In particular, cases that appear to refute the "bodily" criterion of personal identity can be reformulated so as to confirm it. If this is right, then the method of cases must be deployed with care. It is not useless, but it is fallible. As all of the authors in this chapter stress, the method must be supplemented by reflection on the deepest question in the area: "Why exactly does it matter whether I survive, and what could personal identity be that it should matter in this way?"

John Locke (1632–1704)

Locke was an English philosopher and medical doctor. His greatest work is An Essay Concerning Human Understanding (1689), which is about the limits of human knowledge. His Two Treatises of Government (1689) and Letter Concerning Toleration (1689), both published anonymously, made important contributions to political philosophy. The second Treatise gives a theory of legitimate government in terms of natural rights and the social contract. Locke's political views influenced the Founders of the United States, in particular Thomas Jefferson.

OF IDENTITY AND DIVERSITY
from An Essay Concerning Human Understanding

3. Let us suppose an atom... existing in a determined time and place; it is evident, that, considered in any instant of its existence, it is in that instant the same with itself. For, being at that instant what it is, and nothing else, it is the same, and so must continue as long as its existence is continued. In like manner, if two or more atoms be joined together into the same mass, every one of those atoms will be the same, by the foregoing rule: and whilst they exist united together, the mass, consisting of the same atoms, must be the same mass; or the same body, let the parts be ever so differently jumbled. But if one of these atoms be taken away, or one new one added, it is no longer the same mass or the same body. In the state of living creatures, their identity depends not on a mass of the same particles, but on something else. For in them the variation