How Can Determinists Believe in Free Will?

Some people think that we can’t be held responsible for what we do, given that our actions are the inevitable consequence of the laws of nature. They’re only half right.

By Nikhil Krishnan November 6, 2023

According to the Stanford neurobiologist Robert M. Sapolsky, determinism means that human beings don’t really make choices. Moral judgments like blame and praise are based on an illusion.Illustration by Till Lauer

You’re walking fast, late for work. The line into the subway is barely moving. A man is walking very slowly, holding up everyone behind him. You’re annoyed. And then you catch a glimpse of him. He’s walking with the shuffle of the very old. You’re inclined to be a little more tolerant; after all, he can’t walk any faster. You look again—no, he’s not old, just drunk. It’s too late for him to sober up, but, it occurs to you, it was once up to him not to be drunk. And now you’re annoyed again.

But why stop there? There are bars everywhere, and billboards advertising the pleasures of spirits. The days are getting colder, and you live in a cold country—a cold country and a decadent one. Everyone drinks; how could he do otherwise? But, again, why stop there? Generous soul that you are, you wonder if he had a bad day, or week, or year, or life—one marked by the kind of suffering from which the bottle promises respite. Can you be sure that he doesn't come from a long line of alcoholics, helpless in the grip of their compulsion?

You might go further. Perhaps all this was simply meant to be. Recall that old French polymath Pierre-Simon Laplace and his omniscient "demon." If the demon knew where every particle in the universe was at a given moment, he could predict with perfect accuracy every moment in the future—which is another way of saying that the future is wholly "determined" by the past. The demon, of course, merely illustrates a thesis that can be stated in more sombre terms: everything that happens is the inevitable consequence of the laws of nature and what the universe was like once upon a time. We're bound to do what we in fact do.

"Causal determinism," the philosopher's unlovely term for that unsettling hypothesis, is the default assumption of most modern science. It matters a good deal if the idea implies that none of our actions are what we call "free." If science tells us to be determinists, and determinism is incompatible with freedom, shouldn't we give up on judging people for doing what they were destined to do?

That's what the Stanford neurobiologist Robert M. Sapolsky urges. He thinks the time has come to accept the truth about determinism and acknowledge that "we have *no* free will at all." What follows? Early in his book "Determined: A Science of Life Without Free Will" (Penguin Press), Sapolsky lists, with the morbid relish of a man daring to think the unthinkable, the implications of his heresy: no one is ever blameworthy—or, equally, praiseworthy—for doing anything. No one, Sapolsky writes, "has *earned* or is *entitled* to being treated better or worse than anyone else." Ordinary human sentiments—resentment and gratitude, love and hate—are pretty much irrational in their normal forms: "It makes as little sense to hate someone as to hate a tornado because it supposedly decided to level your house." One practical implication is that, since nobody's to blame for anything, criminal justice shouldn't be about retribution. Accordingly, he tries to view human beings without judgment. Free-will skeptics are, he suggests, "less punitive and more forgiving."

The steps of his reasoning are familiar: if everything is determined, there is no freedom; if there is no freedom, there is no moral responsibility. Science tells us that everything is indeed determined. Ergo, no freedom; ergo, no responsibility. Are we bound to agree?

Much of "Determined," appropriately for a book by a neurobiologist, is an attempt to defend the part of the argument which sits safely in the scientist’s wheelhouse: that our best contemporary science has established the truth of determinism. Sapolsky draws on explanations at every level—from atoms to culture—to make the case that everything we think and do is caused by something other than free will. We can’t control, say, the conditions of the crucial first nine months of our existence. “Lots of glucocorticoids from Mom marinating your fetal brain, thanks to maternal stress, and there’s increased vulnerability to depression and anxiety in your adulthood,” he writes. You thought your failing marriage was to blame for your depression. But it turns out it was all bound to happen long ago. It isn’t so much that our blame was aimed in the wrong direction as it is that blame, in the strict sense, never truly made sense.

Biochemistry only confirms Sapolsky in his conviction, as does physics (though he treads more warily through the mysteries of a discipline at least as demanding as his own). History does, too: details from our species’ past, he argues, explain why some cultures are peaceable and others martial, some monotheistic and others polytheistic.

Why do we resist the truth? In part because we’re skilled at concocting stories in which we’re in charge. One experiment Sapolsky mentions suggests that people had cooler attitudes about gay marriage when considering it in a room with a disgusting smell. The effect was connected to the way certain odors activate the brain's insular cortex, which is what makes rancid food nauseating to us. But did the subjects know what was going on in their heads? "Ask a subject, Hey, in last week's questionnaire you were fine with behavior A, but now (in this smelly room) you're not," Sapolsky writes. "They'll claim some recent insight caused them, bogus free will and conscious intent ablaze, to decide that behavior A isn't okay after all." They were being played, but were desperate to regard themselves as the players.

Sapolsky's summaries are pithy and pacy, his spoiling-for-a-fight tone enjoyably provocative. Still, he doesn't assert that any one piece of research demonstrates that there is no free will. Rather, he says, "All these disciplines collectively negate free will because they are all interlinked, constituting the same ultimate body of knowledge." In the end, "there's not a single crack of daylight to shoehorn in free will."

To his exasperation, though, many sophisticates are skeptical about free-will skepticism. In a 2020 survey of academic philosophers, some sixty per cent of them—a strikingly large majority in a profession allergic to consensus—didn't agree that determinism ruled out freedom. Where Sapolsky, who says that he has been a free-will skeptic since adolescence, sees an excitingly counterintuitive conclusion, philosophers often see a reductio ad absurdum.

Determinism, they tend to hold, is compatible with freedom, and therefore with moral responsibility, and therefore with blame, gratitude, and so on.

The technical term for that happy reconciliation is "compatibilism." A compatibilist, that is, agrees that our actions are determined but denies that this truth casts doubt on anything of significance about human practices. Sapolsky will have none of it. He devotes some of his most trenchant writing to these quislings, whose arguments, he claims, "boil down to three sentences":

> a. Wow, there've been all these cool advances in neuroscience, all reinforcing the conclusion that ours is a deterministic world.
>
> b. Some of those neuroscience findings challenge our notions of agency, moral responsibility, and deservedness so deeply that one must conclude that there is no free will.
>
> c. Nah, it still exists.

How fair is that précis? Sapolsky clearly thinks that philosophers should start with the results of the sciences (objective, external, impartial) and ask what they imply for our naïve prescientific self-conceptions. If science, coming from outside, tells us that our self-conceptions are confused, so much the worse for them. Yet there’s another way of approaching the question: from the inside out.

Of those who take such an approach, the most influential is Peter Strawson, a compatibilist philosopher whose arguments make no appearance in “Determined.” (Sapolsky mentions him once in a list of his allies, evidently thinking of his son Galen.) What we label “free will,” as Strawson saw it, was something that we could grasp solely in relation to the role it plays in our lives, in our daily practices of judging others and ourselves.

Start from the fact that, in our dealings with other people, we are sometimes angry or resentful, or, in happier moments, grateful. Does the fact that we’re material beings situated within vast chains of causation make these responses—Strawson calls them “reactive attitudes”—irrational? In any case, could we live an intelligible human life while giving up such reactions entirely?

The really troubling forms of philosophical skepticism—say, about the authority of moral demands, the possibility of altruism, the legitimacy of governments—are troubling in part because it’s possible to live without a belief in such things. Can we live without a belief in the possibility of holding one another and ourselves responsible for the things we do? Try it. We are always a “sorry,” a “thank you,” or a “how dare you” away from slipping back into the bad old ways. Sapolsky's long-standing convictions on the subject notwithstanding, he admits to being a normal guy with a normal guy's feelings. "It's been a moral imperative for me to view humans without judgment or the belief that anyone deserves anything special, to live without a capacity for hatred or entitlement," he writes. "And I just can't do it." He's in permanent misalignment with his theory of the world: "Sure, sometimes I can sort of get there, but it is rare that my immediate response to events aligns with what I think is the only acceptable way to understand human behavior; instead, I usually fail dismally." If even the archest of skeptics cannot live out his skepticism, how serious an alternative is it?

For that matter, if Sapolsky has no choice in how he thinks, what would it even be to try to think any particular way? It's unclear how he could subscribe to "hard incompatibilism," as he avows, and believe in such ancien-régime vestiges as "moral imperatives." The very idea that someone "ought" to do something makes sense only if that person is capable of making choices.

For Strawson (like Kant before him), the difference between a story of causation and a story of choice-making is one of viewpoint. We'll never purge ourselves of those "reactive attitudes," nor should we try to. But that's not a rejection of determinism; as Strawson says, we can sometimes "achieve a kind of detachment from the whole range of natural attitudes and reactions . . . and view another person (and even, perhaps, though this must surely be more difficult, oneself) in a purely objective light—to see another or others simply as natural creatures whose behavior, whose actions and reactions, we may seek to understand, predict and perhaps control.”

The parents of teen-age boys are always oscillating between regarding their impossible offspring as creatures moved by hormones to do stupid things and regarding them as rational human beings who should know better. In time, those boys will learn to give Mom a pass for her possibly perimenopausal outbursts and understand that Dad, with his atherosclerosis, occasionally forgets things. On a night out in college, they’ll learn to spot when it might still be possible to persuade an increasingly drunk friend to order an Uber home and the point at which all that remains to do is to pick him up off the ground and bundle him into one as if he were an unwieldy mattress.

What Sapolsky considers “the only acceptable way to understand human behavior,” then, is one that we adopt *some* of the time. Is it really one we should adopt *all* of the time? Physics tells us a story about ourselves in which we are as much matter and energy as that mattress. Biochemistry tells us another story. So does neuroscience. So do history, psychology, and anthropology. All these modes of explanation tell us, in various ways, that people are causal systems, enmeshed in larger ones. But, from a first-person perspective, we're seldom inclined to wait around to see what the system does when we're faced with a decision. Instead, we do what even Sapolsky finds himself compelled to do—we get on with things and open the kitchen cupboard to decide (or, anyway, "decide") what kind of tea we're going to have.

A few minutes later, a kettle boils. Why? Consider these answers: because I want a cup of Earl Grey; because it's teatime in England and that's what we do around here; because the water molecules have reached a certain energy state. Must we choose among these statements? Can't they all be true?

Sapolsky, to his credit, resists the fallacy of moving from ontological reductionism (the correct view that we're all physical stuff, governed by the laws of physics) to methodological reductionism (the error of concluding that physics is the only discipline that explains anything). Wisely, he doesn't think that history, psychology, or anthropology are just window dressing on the hard truth of our material, molecular existence. Mental states and neurobiological physical states, Sapolsky thinks, "can't be separated—they're just two different conceptual entry points to considering the same processes." Different, but apparently equal, neither privileged over the other. Why, then, does he resist the first-person register of explanation, in which we believe, want, intend, and decide?

Probably because his notion of "free will" isn't the one we actually make use of. Every now and then, it occurs to Sapolsky that the parties to the debate may be talking past each other, and he makes an effort, a little grudgingly, to clarify his terms. "What is free will?" he asks early in the book. "Groan, we have to start with that."

What follows is not a definition but a challenge. A man, Sapolsky invites us to imagine, "pulls the trigger of a gun." That's one description. Another is that "the muscles in his index finger contracted." Why? "Because they were stimulated by a neuron," which was in turn "stimulated by the neuron just upstream. . . . And so on." Then he throws down the gauntlet: "Show me a neuron (or brain) whose generation of a behavior is independent of the sum of its biological past, and for the purposes of this book, you've demonstrated free will."

Sapolsky's qualifying phrase—"for the purposes of this book"—suggests he recognizes that there are other things "free will" might mean to other people. The real question is whether the thing he thinks has been disproved will be much missed—whether Sapolsky's concept of free will is "our" concept, what we take it to be when, uncorrupted by the specialists, we go about our days blaming and thanking, loving and hating.

His definition of free will, or the test he proposes in lieu of one, exemplifies an approach advocated in the nineteen-fifties and sixties by the German-born philosopher Rudolf Carnap, one of the great exponents of logical positivism. Carnap, who thought that philosophy should be a handmaiden to the sciences, found ordinary language squishy and unreliable; philosophers, he argued, should reconstruct, or “explicate,” its terms in more precise ways. Take a term of ordinary language such as “warm,” a word we understand as relating to our sensations, and a term of science, “temperature,” which, unlike “warm,” is a quantitative notion, employable in scientific formulas. Carnap thought that we would do well to replace the first with the second, or, anyway, to use the second to define the first. Something is warmer when its temperature increases, and the temperature of a room, unlike its mere “warmth,” can be precisely and objectively defined with reference to a thermometer.

In the positivist spirit, a philosopher or a scientist may replace “We have free will” with some determinate claim that might be supported or undermined by experimental evidence. But if we’re going to allow the experimental results, and the ambitious thesis that they support, to overthrow our everyday practices of praising, blaming, and punishing, the scientific explication had better be close enough to the everyday concept that underlies these practices. Otherwise, as Strawson—a leading critic of Carnap, and of logical positivism—objected at the time, we’re not solving a problem; we’re just changing the subject.

Sapolsky may want us to replace our everyday concept, complex and muddled as it must be, with a new, improved concept. But he needs to offer us a good argument for doing so. In other words, if this isn’t how we already use “free will,” is it how we *should* use it? We don’t ask this question often enough. Maybe that’s because “free will” has become, for better or for worse, a term of ordinary language. As with other such large, abstract questions, our professed views on the free-will problem rarely do justice to the complexity of our real thoughts on the subject.

There are sane and humane reasons for reminding ourselves that, as Sapolsky puts it, “*some* people have much less self-control and capacity to freely choose their actions than average, and that at times, we *all* have much less than we imagine.” But Sapolsky and skeptical fellow-travellers such as Sam Harris want to blow up our everyday practices; they’re convinced that the innumerable subtle distinctions we mark in our ordinary talk and practices don’t matter. Whatever it is we think we’ve got, we haven’t got it.

Once we give up on free will, after all, we have to be willing to say that nobody has ever done anything intentionally (or voluntarily or deliberately or on purpose). That no line can truly be drawn between the malicious and the accidental, the voluntary and the forced. That no event has really been an *action*, except in the undemanding sense in which the puppet Punch killing Judy was an action. We are not, except in this minimal sense, "agents." Bits of our bodies move; things happen. That's it.

Yet we can wonder whether this ideal of the "causeless cause" is what we really mean in our careless everyday talk of free will. If it isn't, then insisting on that as the appropriate standard is, as Strawson once remarked, "like offering a textbook on physiology to someone who says (with a sigh) that he wished he understood the workings of the human heart."

As it happens, Carnap's own definition, or explication, of "free choice" was devised to be compatible with the truth of determinism. Free choice, he wrote, "is a decision made by someone capable of foreseeing the consequences of alternate courses of action and choosing that which he prefers." He saw no contradiction between free choice thus understood and even the strongest form of determinism. More than that, he thought that without determinism—that is to say, without reliable relations of cause and effect—there *is* no free will. The point of making a choice, he noted, is that it has consequences. Indeterminacy—should the kind that exists at the quantum level encroach upon the meso-level scale of human lives—would replace our agency with sheer randomness, which is not anyone's idea of freedom.

Still, if Sapolsky's radical revisionism threatens to eliminate our moral vocabulary altogether, the traditional compatibilist approaches can err in the other direction: they can deny us grounds for reforming our practices of moral judgment. So we might want to ask—in the philosophical tradition not of positivism but of pragmatism—about the "cash value," the practical uses, of invoking free will. Ignore the many things we say about free will and focus on what we do. By attending to our practices, we can conduct a more profitable line of inquiry: What do we need free will to be?

Consider these questions. How does my response to a student who has failed to meet a deadline change when I discover that she sprained her wrist the day before? How do I respond to a piece of fruit chucked at me when I discover that the chucker is two years old? How vehemently do I press for a tough prison sentence when I learn that the defendant was abused as a child? Is there a difference between someone jumping into a pool and being pushed in? Between falling in after he stepped on a banana peel and falling in because he was drunk?

In each of these situations, we can ask whether the person in question had or lacked something called "free will." But look back at the factors that seemed to undermine the exercise of free will: rotten luck, immaturity, circumstance, coercion, accident, and incapacity. When the terms on the obverse side of the contrast are so disparate, it's hard to be confident that there really is a single thing called "free will" whose presence or absence we can meaningfully debate.

The philosopher J. L. Austin, in a paper titled "A Plea for Excuses," observed that, though it's tempting to view "freedom" as a positive term requiring elucidation, we tend to use it just to rule out one of these miscellaneous antitheses. Freedom's just another word for no excuses. In Austin's spirit, you can wonder how much would be lost if we ceased talking about free will altogether and spoke more specifically of what we meant. "He was not acting of his own free will," you tell me. "What do you mean?" I ask. "Was he being held at gunpoint? Sleepwalking? High on narcotics?" "Oh, no," you say, "I just mean he's a toddler." Maybe you should have just said that. Maybe, in other words, there is no *one* thing that we need "free will" to be.

Sapolsky's most persuasive passages remind us of the many modest changes that we have made to our practices in light of advances in our understanding of certain facts about ourselves. Things got better for people with epilepsy once we learned that the condition was not a form of demonic possession that was “brought on by someone’s own freely chosen evil.” Things got better for people with schizophrenia (and for their families) once it was recognized that the condition was at base a biochemical disorder, not a product of faulty mothering. Some of our current moral conceptions and presumptions may come to seem just as confused. We may currently be blaming people whom it would be better to treat, tolerate, or simply avoid.

Or, perhaps, better to forgive. That was, after all, part of Sapolsky’s case for his new morality: free-will skeptics are “less punitive and more forgiving.” To understand all, he might have said, is to forgive all. But he can’t have really meant that. Forgiveness is, as much as vengeance, a concept that can be applied only from within the first-person point of view. If free-will skepticism means never having to say you’re sorry, then it also means never being forgiven. Sapolsky’s ethic of forgiveness demands that we retain something of our old-fashioned belief in holding one another responsible.

The traditional project of compatibilist philosophers has been to treat determinism, free will, and moral responsibility as fixed parts of a triad and to search for ways to reconcile them. The skeptics, seeing an irresolvable contradiction, have concluded that the whole idea of moral responsibility has to go. But there are other ways of reconciling scientific and moral inquiry. It may be that we need a suppler and more humane approach to holding one another responsible, an approach that takes more seriously what our best scientific accounts tell us about ourselves. We needn’t follow the skeptics to the conclusion that the best morality would be no morality at all to recognize that our current morality remains a work in progress. ♦