'A firmly established galaxy of brilliant minds orbiting Sam's shrewd but solar intelligence.'
Niall Ferguson

Making Sense

Conversations on Consciousness, Morality and the Future of Humanity

Sam Harris

New York Times bestselling author

WITH NICK BOSTROM, DAVID CHALMERS, DAVID DEUTSCH, DANIEL KAHNEMAN, DAVID KRAKAUER, GLENN C. LOURY, THOMAS METZINGER, ROBERT SAPOLSKY, ANIL SETH, TIMOTHY SNYDER AND MAX TEGMARK

Contents

Acknowledgments
Contributors

About the Author

SAM HARRIS is the author of the bestselling books *The End of Faith*, *Letter to a Christian Nation*, *The Moral Landscape*, *Free Will*, *Lying* and *Waking Up*. *The End of Faith* won the 2005 PEN Award for Nonfiction. His work has been published in over twenty languages.

Harris has written for *The Times*, the *Economist*, the *New York Times*, the *Atlantic* and elsewhere. He received a degree in philosophy from Stanford University and a PhD in neuroscience from UCLA. He is the host of the internationally popular podcast *Making Sense* and creator of the Waking Up app. Please visit his website at SamHarris.org.

Praise for Sam Harris and *Making Sense*

"Sam is an enlightened, rational voice in a world that needs it now, perhaps more than ever—especially if we are to survive, or thrive, in spite of the collective weaknesses that make us human."

—NEIL DEGRASSE TYSON, astrophysicist, American Museum of Natural History

"In the huge world of interviewers, Sam Harris stands out at the top for his probing questions and for his own thoughtful views."

—JARED DIAMOND, author of *Guns, Germs, and Steel*

"To the raging controversies of the day, Sam Harris adds a voice of civility and reason."

—LAWRENCE WRIGHT, author of *The Looming Tower*

"*Making Sense* brings the power and patience of contemplation to the art of conversation. Sam Harris models not only how to articulate complex ideas, but also how to truly hear the ideas of others. This is cognitive jazz at its best."

—DOUGLAS RUSHKOFF, author of *Present Shock*

"Whatever your politics, you will find ideas and points of views you've never considered before, in fields you don't know, from neuroscience to computer science to culture."

—ANNE APPLEBAUM, author of *Gulag*

"It's no wonder that Sam attracts a huge audience. He is a thinker with his own ideas, so his interviews are some of the most interesting conversations you are ever likely to hear."

—PETER SINGER, author of *Animal Liberation*

"There is no podcast that approaches the intellectual rigor and open-mindedness of Sam Harris's *Making Sense*. It's a regular dose of sane, patient reason, and dialogue. In a tribalized world, it

reveres the individual, inquisitive mind. And Sam has some balls to talk honestly where so many others won't."

—**ANDREW SULLIVAN**, author of *The Conservative Soul*

"Sam Harris is tremendous at his job; sharp, skeptical in just the best sense, and full of curiosity and openness. He's a terrific questioner, and he greatly enlivens and improves public discourse."

—**CASS R. SUNSTEIN**, author of *Can It Happen Here?: Authoritarianism in America*

"Of all the podcasts available, the one nobody should miss is *Making Sense*. Every episode is stimulating. In an era when everyone seems to have lost their reason, here is one of the few places where reason remains safe."

—**DOUGLAS MURRAY**, author of *The Madness of Crowds*

"Sam Harris is a true public intellectual: he thinks deeply about a wide range of issues and engages fearlessly with controversial topics and unpopular opinions. You don't have to agree with him to learn from him—I always come away from his show with new insights and new questions."

—**ADAM GRANT**, author of *Originals* and host of the TED podcast *WorkLife*

"Sam makes sense of important, difficult, and often controversial topics with deep preparation, sharp questions, and intellectual fearlessness."

—**ANDREW MCAFEE**, author of *More from Less*

"Sam has given us one of the greatest podcasts in the world for clear thinking. We are better equipped to face the perils and uncertainties of life with it in the air. It's a stand-out leader in a cluttered field and being Sam's guest on it was a career highlight."

—**DERREN BROWN**, author of *Happy*

"There are precious few spaces in the media landscape where difficult, rigorous, and respectful conversations can play out at substantial length, without agenda. Sam Harris created the

model for such illuminating exchange, and the *Making Sense* podcast is a treasure trove of discussions with many of the most compelling and fascinating minds of our era."

—**THOMAS CHATTERTON WILLIAMS**, author of *Self-Portrait in Black and White*

"As an interviewer, Sam is both rigorous and generous. His show is completely devoid of the cheap shots and tribal bickering that characterize so much of podcasting. *Making Sense* is joyful play of the mind, without a trace of the partisan cretinism that disfigures the vast majority of our discourse these days."

—**GRAEME WOOD**, author of *The Way of the Strangers*

"Sam Harris does an incredible job probing—and finding answers to—some of the most important questions of our times."

—**SIDDHARTHA MUKHERJEE**, author of *The Emperor of All Maladies*

ALSO BY SAM HARRIS

The End of Faith

Letter to a Christian Nation

The Moral Landscape

Lying

Free Will

Waking Up

Islam and the Future of Tolerance (with Maajid Nawaz)

The Four Horsemen (with Richard Dawkins, Daniel Dennett, and Christopher Hitchens)

To my mother

Preface

We are living in a new golden age of public conversation.

Millions of us have recently discovered that significant parts of the day—a commute, an hour at the gym, an eternity spent on the threshold of sleep—can be filled with podcasts and related media. Increasingly, we replace the voice in our heads with the voices of others—whose opinions, whether considered or not, now inform our own. I'm convinced that this is generally a good development. Every hour, we struggle to maintain a vast, technological civilization, and yet conversation remains our only means of making intellectual and moral progress.

Podcasting began in 2004, which happens to be the year I published my first book. If someone had told me then that I would eventually spend most of my time producing a podcast, rather than writing, I would have said, "What's a podcast?" If they had then described this new form of media—more or less accurately—as "radio on demand," I would have been willing to bet the fate of our species that they were mistaken about me. For as long as I can remember, I've wanted to write books. At no point in my life have I spent two consecutive breaths wondering whether I might like to work in radio.

And yet, creating the *Making Sense* podcast has consumed most of my professional energy in recent years. The reasons for this are disconcertingly simple: I will reach more people in forty-eight hours with my next podcast than I will reach in a decade with all of my books. And the results are instantaneous: instead of waiting a year for a book to be published, I can release a podcast the moment it's finished.

In truth, the analogy to radio is somewhat misleading. The distinction between a radio show that is allotted a full hour in a fixed schedule, and a podcast episode that just happens to wrap up after fifty-nine minutes, can be hard to appreciate from the

outside. But the difference is felt every moment along the way. Time pressure changes everything—a fact that anyone can perceive when watching a formal debate. A willingness to explore adjacent topics, to backtrack, to try ideas on for size only to discard them, to invite criticism without knowing what one's response to it will be—and when disagreements surface, to give one's opponents the freedom to present the best possible case for their views—such a spirit of dialogue can only arise when the threat of being interrupted isn't further weaponized by a ticking clock. When we are guided by real curiosity and a principle of charity, every human problem seems to admit of solution. In other moods, even conversation itself proves impossible.

Podcasting is the only medium that allows for truly natural, open-ended conversation. So it's not an accident that this is where scientists, journalists, and public intellectuals now think out loud. But the strength of the medium is also its primary weakness, because conversation lacks the precision of written work. And listeners may fail to catch subtle points that readers would naturally pause to absorb. Thus, when compared to the clarity and accessibility of books, even some of the most interesting podcasts can feel like missed opportunities.

In this volume, I've collected some of my favorite conversations from *Making Sense* and adapted them for print. To do this, I've asked my guests to refine their side of the exchange, and I've done the same to mine. The result follows the pattern of our original conversation, but we've made many small amendments and clarifications throughout. Now, everyone involved can be counted upon to have said what they truly mean.

Since 2014, I've released over two hundred episodes of the *Making Sense* podcast, now averaging about one per week. This volume presents thirteen of my favorites, with eleven guests—David Chalmers, David Deutsch, Anil Seth, Thomas Metzinger, Timothy Snyder, Glenn Loury, Robert Sapolsky, Daniel Kahneman, Nick Bostrom, David Krakauer, and Max Tegmark. The book covers a wide range of concerns—consciousness, the foundations of knowledge, ethics, artificial intelligence, politics, physics, decision making, racism, violence, existential risk—but it is heavily weighted toward questions about the nature of mind

and how minds like ours can best create a world worth living in. As listeners to *Making Sense* know, these are my core interests, and I return to them often.

I have long believed in the ultimate unity of knowledge, and thus that the boundaries between traditional disciplines should be generally ignored. One thing we surely know about reality at this point, is that it isn't partitioned like a university campus. I also believe that most of the evil in our world—all the needless misery we manufacture for one another—is the product, not of what bad people do, but of what good people do once in the grip of bad ideas. Taken together, these principles suggest that there is no telling how much moral progress we might make by removing the impediments to clear thinking on any topic that interests us.

For instance, as I write these lines the world is still struggling to understand the gravity of the COVID-19 pandemic, which has now spread to 187 countries. Political, philosophical, religious, and economic beliefs now contend with the basic principles of epidemiology in the brains of millions of people, some of whom are responsible for making and enforcing policies that will affect the lives of billions. There is still no consensus on how societies should respond to this crisis, and factions have formed on the basis of entirely different views of terrestrial reality. Has the danger of this disease been exaggerated for political gain? Is it unethical to force businesses to close and people to stay indoors in an effort to slow the contagion? Do governments have a responsibility to provide free health care to their citizens? Should the Chinese be admonished to stop eating bats, or would that be a sign of racism? Where is the boundary between contrarian thinking and deadly misinformation? Everywhere one looks, one sees the ruins of failed epistemology—and bad ideas are getting people killed.

There are now nearly one million different podcasts to choose from. Many just give voice to the general pugnacity of our age—and my own podcast has not been entirely immune. But the antidote to bad conversations is always better ones. And here I present some of the most satisfying conversations I've ever had.

Enjoy ...

Sam Harris
May 6, 2020
Los Angeles

The Light of the Mind

A CONVERSATION WITH DAVID CHALMERS

Trying to understand consciousness has long been a foundational interest of mine, and given his role in sparking that interest, I begin *Making Sense* with David Chalmers. A philosopher at New York University and at the Australian National University, Chalmers is also a codirector of the Center for Mind, Brain, and Consciousness at NYU.

We spend most of our time discussing the nature of consciousness and why it is so difficult to understand scientifically. We begin with Chalmers's notion of "the hard problem of consciousness"—a phrase that has influenced every debate on the subject since the early 1990s. We also talk about artificial intelligence, the possibility that the universe is a simulation, and other fascinating topics, some of which may *seem* impossibly distant from the concerns of everyday life. But I would urge you not to be misled here. All of these topics will become more and more relevant as we continue to build technology that, whether conscious or not, will seem conscious to us. And as we confront the prospect of augmenting our own minds by integrating devices directly into our brains, all of these philosophical puzzles will become matters of immediate personal and ethical concern.

HARRIS: You've played an important role in my intellectual life. I went to one of those early biennial Tucson conferences on consciousness, at the University of Arizona. I had dropped out of school, and I guess you could say I was looking for some direction in life. I'd become interested in the conversations that were happening in the philosophy of mind—initially because of the sparring between Daniel Dennett and John Searle. Then I saw an ad for the Tucson conference, probably in the *Journal of Consciousness Studies,* and just showed up.

I distinctly remember your talk there. Your articulation of the hard problem of consciousness made me want to do philosophy, which led directly to my wanting to know more science and sent me back to the ivory tower. Part of my reason for getting a PhD in neuroscience, and for my continued interest in this issue, was the conversation you started in Tucson more than twenty years ago.

CHALMERS: I'm really pleased to hear that. That was probably the '96 conference. Dennett was there.

HARRIS: Along with Roger Penrose, Francisco Varela, and many others. It was a fascinating time.

CHALMERS: The previous event in 1994 is what people called the Woodstock of Consciousness. Getting the band together for the first time. It was crazy, a whole lot of fun, and the first time I'd met a lot of these people, too.

HARRIS: I'm a bad judge of how familiar people are with the problem of consciousness, because I've been so steeped in it for decades now. I'm always surprised that people find it difficult to grasp that consciousness poses a special challenge to science. So let's start at the beginning. What do you mean by "consciousness," and how would you distinguish it from the topics it's usually conflated with, like self-awareness, attention, thinking, behavior, and so forth?

CHALMERS: It's awfully hard to define consciousness. But I'd start by saying that it's the subjective experience of the mind and the

world. It's basically what it feels like, from the first-person point of view, to be thinking and perceiving and judging. When I look out the window, there are trees and grass and a pond, and so on. And there's a whirl of information processing as photons in my retinas send a signal up the optic nerve to my brain—that's on the level of functioning and behavior.

But there's also something that it feels like from a first-person point of view. I might have an experience of the colors, a certain greenness of the green, a certain reflection on the pond—like an inner movie in my head. And the crucial problem of consciousness—for me, at least—is this subjective part. We can distinguish it from questions about behavior or functioning. People sometimes use the word "consciousness" just to indicate, for example, that I'm awake and responsive. That's something that is straightforward and can be understood in behavioral terms. I like to call those problems of consciousness the easy problems—the ones about how we behave, how we respond, how we function. What I call the hard problem of consciousness is the one about how it feels from the first-person point of view.

HARRIS: There was another influential statement of this problem, which I assume influenced you as well: Thomas Nagel's 1974 essay "What Is It Like to Be a Bat?" The formulation he gave there is: if it's *like something* to be a creature processing information—if there's an internal, subjective, qualitative character to the processing—that is what we mean by "consciousness," in the case of a bat or any other system. People who don't like that formulation think that, as a definition, it just begs the question. But as a rudimentary statement of what consciousness is, I've always found it very attractive. Do you have any thoughts on that?

CHALMERS: It's about as good a definition as we're going to get. The idea is roughly that a system is conscious if there's something it's like to be that system. There's something it's like to be me. There's nothing it's like, presumably, to be this glass of water on my desk. If there's nothing it's like to be the glass of water on my desk, then the glass of water is not conscious.

Likewise, some of my mental states. There's something it's like for me to see the green leaves outside my window right now, so

that's a conscious state to me. But there may be some unconscious language-processing going on in my head that doesn't feel like anything to me, or some motor processes in the cerebellum. Those might be states of me, but they're not conscious states of me, because there's nothing it's like for me to undergo those states.

So Nagel's definition is vivid and useful for me. That said, it's just a bunch of words, like any other. And for some people, this bunch of words is useful in activating the idea of consciousness from the subjective point of view. Other people hear something different in that set of words. For those people, the words "What is it like?" doesn't work. What I've found over the years is that this phrase of Nagel's is useful for some people, in getting them onto the problem, but it doesn't work for everybody.

My sense is that most people do have some notion that there's a big problem here. What they do after that is different in different cases. Some people think we ought to see the hard problem as an illusion and get past it. But to focus the issue, I find it useful to start by distinguishing the easy problems—which are basically about the performance of functions—from the hard problem, which is about experience.

The easy problems are: How do we discriminate information in our environment and respond appropriately? How does the brain integrate information from different sources and bring it together to make a judgment and control our behavior? How do we voluntarily control our behavior to respond in a controlled way to our environment? How does our brain monitor its own states? These are all mysteries, and neuroscience has not gotten all that far on some of them. But we have a pretty clear sense of what the research program is and what it would take to explain them. It's basically a matter of finding some mechanism in the brain that is responsible for discriminating the information and controlling the behavior. Although pinning down the mechanisms is hard work, we're on a path to doing it.

The easier problems at least fall within the standard methods of neuroscience and cognitive science. What makes the hard problem of experience hard? Because it doesn't seem to be a problem about behavior or about functions. You can in principle imagine explaining all of my behavioral responses to a given

stimulus and how my brain discriminates and integrates and monitors itself and controls my behavior. You can explain all that with, say, a neural mechanism, but you won't have touched the central question, which is, "Why does it *feel* like something from the first-person point of view?"

The usual methods that work for us in the neural and cognitive sciences—finding a mechanism that does the job—doesn't obviously apply here. We'll certainly find correlations between processes in the brain and bits of consciousness—an area of the brain that might light up when you see red or when you feel pain. But nothing there seems to explain why all that processing feels like something from the inside. Why doesn't that processing just go on in the dark, as if we were robots or zombies without any subjective experience?

So that's the hard problem, and people react in different ways to it. Someone like Dan Dennett says it's all an illusion, or a confusion, and one that we need to get past. I respect that line of thought. It's a hard-enough problem that we need to be exploring every avenue, and one avenue that's worth exploring is the view that it's an illusion.

But there's something faintly unbelievable about the idea that the data of consciousness are an illusion. To me, they're the most real thing in the universe—the feeling of pain, the experience of vision, the experience of thinking. Dan Dennett takes a very hard line in his 1991 book *Consciousness Explained.* It was a good and very influential book. But I think that most people found, at the end of the day, that it didn't do justice to the phenomenon.

HARRIS: That might have been the first book I read on this topic. It's strange—I'm aligned with you and Thomas Nagel on these questions in the philosophy of mind, and yet I've had this alliance with Dan for many years on the conflict between religion and science. I've spent a fair amount of time with Dan, but we've never really gotten into a conversation on consciousness. Perhaps we've been wary of it; we had a somewhat unhappy collision on the topic of free will. It's been a long time since I've read *Consciousness Explained*—does he say that *consciousness* is an illusion, or just that the hardness of the hard problem is illusory? I understand that he'd want to push the

latter intuition. But as for the former, it seems to me that consciousness is the one thing in this universe that cannot be an illusion. Even if we're confused about the qualitative character of our experience in many other respects, the fact that it is *like something* to be us, the fact that something seems to be happening, even if it's only a dream—that *seeming* is all one needs to assert the undeniable reality of consciousness. I just don't see how anyone can credibly claim that consciousness itself might be an illusion.

CHALMERS: I'm with you on this. I think Dan's views have evolved over the years. Back in the 1980s or so, he used to say things that sounded much stronger, like "Consciousness doesn't exist. It's an illusion." He wrote a paper called "On the Absence of Phenomenology," saying there really isn't such a thing as phenomenology, which is basically just another word for consciousness. He wrote another one called "Quining Qualia," which said we needed to get rid of the whole idea of qualia, which is a word that philosophers use for the qualitative character of experience; what makes seeing red different from seeing green. Those experiences seem to involve different qualities. At one point Dan was inclined to say, "That's just a mistake. There's nothing there."

Over the years, I think he found that people consider that position—that from the first-person point of view there are no qualia, no feeling of red versus the feeling of green—a bit too strong to be believable. So he's evolved in the direction of saying, yes, there's consciousness, but it's just in the sense of functioning and behavior and information encoded, and not really consciousness in the strong phenomenological sense that drives the hard problem.

In a way, this is a verbal relabeling of his old position. I know you're familiar with the debates about free will, where one person says, "There's no free will," and the other person says, "Well, there is free will, but it's just this much more deflated thing which is compatible with determinism"—and these are basically two ways of saying the same thing. Dan used to say there's no consciousness; now he says, "Well, there's consciousness, but only in this deflated sense"—which is another

way of saying the same thing. He still doesn't think there is consciousness in the strong, subjective sense that poses the whole problem.

HARRIS: I want to retrace what you said in sketching the hardness of the hard problem. You make the distinction between understanding function and understanding the fact that experience exists. We have functions, like motor behavior and visual perception, and it's straightforward to think about explaining them in mechanistic terms. With vision, for example, we can talk about the transduction of light energy into neurochemical events and then mapping the visual field onto the relevant parts of the visual cortex. This is complicated but not, in principle, obscure. However, the fact that it's *like something to see* remains mysterious, no matter how much mapping we do.

And if we built a robot that could do all the things we can, it seems to me that at no point in refining its mechanisms would we have reason to believe that it was conscious, even if it passed the Turing Test.

This is one of the things that concerns me about AI. It seems increasingly likely that we will build machines that will seem conscious, and the effect could be so convincing that we might lose sight of the hard problem. It could cease to seem philosophically interesting, or even ethically appropriate, to wonder whether there is something it is like to be one of these robots. And yet we still won't know whether they are actually conscious unless we have understood how consciousness arises in the first place—which is to say, unless we have solved the hard problem.

CHALMERS: Maybe we should distinguish the question of whether a system is conscious from the question of how that consciousness is explained.

I suspect that with machines, if they're hanging around with us, talking in a humanlike way and reflecting on their consciousness, saying, "I'm really puzzled by this whole consciousness thing, because I know I'm just a collection of silicon circuits, but it still feels like something from the inside"—if machines are doing that, I'll be pretty convinced that they're

conscious as I am conscious. But that won't make consciousness any less mysterious, and it might make it all the more mysterious. How could this machine be conscious if it's just a collection of silicon circuits? Likewise, how could I be conscious just as a result of processes in my brain? I don't see anything intrinsically worse about silicon than about brain processes here; there's a mysterious gap in the explanation in both cases.

And, of course, we can wonder about other people, too. That's a classic philosophical problem, the problem of other minds. How do you know that anybody apart from yourself is conscious? Descartes said, "Well, I'm certain of one thing: I'm conscious. I think, therefore I am." That only gets you one data point. It gets me the me being conscious—and only being conscious right now, because who knows if I was ever conscious in the past? Anything beyond right now has to be an inference or an extrapolation. We end up taking for granted most of the time that other people are conscious, but as you move to questions about AI and robots, about animals and so on, the question of who else is conscious becomes very murky.

HARRIS: The difference as far as AI or robots are concerned is that presumably we'll build them along lines that aren't analogous to the emergence of our own nervous systems. We might proceed as we have with chess-playing computers—where we have built something that we have no reason to believe is aware of chess, and yet is now the best chess player on Earth. If we do this for a thousand different human attributes and thus create a computer that can function as we do, but better—perhaps a robot that has mimetic facial displays we find compelling, and so no longer seems weird or lifeless to us. If this system is built in a way that is nonanalogous to our own nervous system, then it could be hard to tell whether or not it's conscious. Whereas in the case of other people, I have every reason to believe that the structures that suffice to produce consciousness in my case, probably suffice for them too. Solipsism isn't really tempting, philosophically speaking, because there's a deep analogy between how I came to be conscious and how you came to claim that you are conscious too. I'd have to argue that there was something about your nervous system, or your situation in the universe, that wasn't

sufficient to produce consciousness, while it clearly was in my own case. To wonder whether other people, or even the higher animals, are conscious is not an example of being parsimonious; rather, it requires extra work.

CHALMERS: How would you feel if we met Martians? Let's say there are intelligent Martians who are behaviorally sophisticated and we find we can communicate with them about science and philosophy, but they've evolved through an evolutionary process different from ours. Would you have doubts about whether they might be conscious?

HARRIS: Perhaps I would. It would be somewhere between our own case and whatever AI we might build. This leads to a topic I wanted to raise with you: the issue of epiphenomenalism, which is actually the flip side of the hard problem. The fact that we can describe all this functioning without introducing consciousness leaves us with another problem many people find counterintuitive: namely, that consciousness might not be doing anything—that it's an epiphenomenon. In an analogy often cited, it's like the smoke coming out of the smokestack of an old-fashioned locomotive. The smoke is associated with the progress of the train down the tracks, but it's not actually doing any work; it's merely a by-product of the mechanism that is actually propelling the train. Consciousness could be like this. In your first book, *The Conscious Mind*, you seemed to be fairly sympathetic with epiphenomenalism.

CHALMERS: The idea that consciousness doesn't do anything—that it's epiphenomenal—is not a view that anyone feels an initial attraction for. It sure seems to do so much. But there's this puzzle: For any behavior, there's a potential explanation in terms of neurons or computational mechanisms that doesn't invoke consciousness in the subjective sense. You can at least start to wonder if maybe consciousness *doesn't* have any function. Maybe it doesn't do anything at all. Maybe, for example, consciousness simply gives value and meaning to our lives—which is something we can talk about. But if it does nothing else, then all kinds of questions arise: How and why would we have evolved as we have—let alone come to be having this extended

conversation about consciousness—if consciousness were not playing some role in the causal loop?

In *The Conscious Mind*, I at least tried on the idea of epiphenomenalism. I didn't flat out say, "This is definitely true." I tried to say, "Well, if we're forced to, that's one way we could go." Either consciousness is epiphenomenal or it's outside a physical system but somehow playing a role in physics. That's a more traditional, dualist possibility. Or there's a third possibility: Consciousness is somehow built in at the fundamental level of physics.

HARRIS: I'd like to track through each of those possibilities, but let's stick with epiphenomenalism for a moment. You've touched on it in passing here, but remind us of the "zombie argument," the thought experiment that describes epiphenomenalism. It's not an argument I'd noticed before I heard you make it, but I don't know if it originates with you.

CHALMERS: The idea of zombies, in philosophy not to mention in popular culture, was out there before me. I think the philosopher Robert Kirk originated the label in the 1970s, and the idea itself goes back further. The zombies of philosophy are different from the zombies of the movies or in Haitian voodoo culture. All these zombies are missing something. The zombies in the movies are lacking life; they're dead but reanimated. The zombies in the voodoo tradition lack some kind of free will. The zombies that play a role in philosophy lack consciousness.

In this thought experiment, the conceit is that we can imagine a being behaviorally identical to a normal human being—a being that acts and walks and talks in a perfectly humanlike way—but without any consciousness at all. There's an extreme version that asks you to imagine a being *physically identical* to a particular human being but without subjective consciousness. I talk about my zombie twin, a hypothetical being in the universe next door, who's physically identical to me. He's holding this conversation with a zombie analog of you right now, saying all the same stuff and responding but without any consciousness.

Now, no one thinks that anything like this exists in our universe. But the idea is at least conceivable. And the very fact that you can make sense of it immediately raises questions like

"Why aren't *we* zombies?" Evolution could have produced zombies; instead, it produced conscious beings. Why didn't evolution produce zombies? If there were some function we could point to and say, "That's what you need consciousness for; you couldn't do that without consciousness," then we might have a function for consciousness. But right now, for anything we actually do—perception, learning, memory, language, and so on—it sure looks as if a whole lot of it could be done unconsciously. The whole problem of what consciousness is doing is thrown into harsh relief by the zombie thought experiment.

HARRIS: Most of what our minds accomplish is unconscious, or at least it seems so. The fact that I perceive my visual field, the fact that I hear your voice, the fact that I effortlessly decode meaning from your words because I'm an English speaker—this is all done unconsciously before I have an experience of any of these things. So it's a mystery why there should be something that it's like to be associated with any part of this process, because so much of it takes place in the dark.

This is a topic I raised in my last book, *Waking Up*, in discussing split-brain research. There is reason to wonder whether or not there are islands of consciousness in our brains that we're not aware of—that is, we have an "other minds" problem with respect to our very own brains. What do you think about the possibility that there is something that it's like to be associated with parts of your own cognitive processing that seem like zombie parts?

CHALMERS: Well, I don't rule it out. When it comes to the mind/body problem, the puzzles are large enough. One of the big puzzles is, we don't know which systems are conscious. Most of us think humans are conscious, and probably a lot of the more sophisticated mammals are conscious: apes, monkeys, dogs, cats. When it gets to mice, maybe flies, some people start to wobble, but I like the idea that for many reasonably sophisticated information-processing devices there's some kind of consciousness. Maybe this goes very deep, and at some point we can talk about the idea that consciousness is everywhere.

But before that, if you're prepared to say that a fly is conscious, or a worm with its three hundred neurons, then you

do have to wonder about pieces of the brain that are enormously more sophisticated than that but are part of another conscious system. The neuroscientist Giulio Tononi recently proposed a theory of consciousness called IIT, integrated information theory. He's got a mathematical measure, Φ, of the amount of information a system integrates. Whenever it's high enough, you get consciousness.

When you look at different pieces of the brain, like each hemisphere, the cerebellum, and so on, you note that the Φ isn't as high as it is for the brain as a whole, but it's still pretty high. Tononi would say that an animal with a Φ that high was conscious. So why isn't that piece of the brain conscious? He ends up throwing in an extra axiom, which he calls the exclusion axiom, saying, in effect, that if you're a part of a system that has a higher Φ than you do, then you're not conscious. If the hemisphere has a high Φ but the brain as a whole has a higher Φ, then the brain gets to be conscious but the hemisphere doesn't. To many people, that axiom looks arbitrary. But without that axiom, you'd be left with a whole lot of conscious subsystems. And I agree: Who knows what it's like to be a subsystem—what it's like to be my cerebellum, what it's like to be a hemisphere? On the other hand, there are experiments and situations in which one half of the brain gets destroyed and the other half keeps going fine.

HARRIS: I wanted to ask you about Tononi's notion of consciousness as integrated information. To me, it's yet another case of someone trying to ram past the hard problem. Max Tegmark wrote a paper, "Consciousness as a State of Matter," that took Tononi as a starting point. He basically said, "Let's start here. We know there are certain arrangements of matter that are conscious, now we just have to talk about the plausible explanation for what makes them conscious." He went on to embrace Tononi and then did a lot of physics.

But is there anything in Tononi's discussion that pries up the lid on the hard problem farther than the earlier work he did with Gerald Edelman, or farther than anyone else's attempt to give some information-processing construal of consciousness?

CHALMERS: To be fair to Tononi, he's actually very sensitive to the problem of consciousness. And when pressed, he says he's not trying to solve the hard problem of showing how you can get consciousness from matter. He's not trying to cross the explanatory gap from physical processes to consciousness. Rather, he says, "I'm starting with the *fact* of consciousness. I'm taking that as a given, and I'm trying to map its properties." And he starts with some phenomenological axioms of consciousness, for example that it consists of information that's differentiated in certain ways but integrated and unified in other ways. Then he takes those phenomenological axioms and turns them into mathematics of information and asks, "What informational properties does consciousness have?" Then he comes up with this mathematical measure, Φ. At some point, the theory that consciousness is a certain kind of integration of information arises. The way I see the theory—I don't know if he puts it this way—is as correlating different states of consciousness with different kinds of integration of information in the brain.

So the hard problem is still there, because we still have no idea why all that integration of information in the brain should produce consciousness in the first place. But even someone who believes there's a hard problem can believe that there are systematic correlations between brain processes and consciousness that we should have a rigorous mathematical theory of. Tononi's theory is basically a stab in the direction of providing a rigorous mathematical theory of those correlations.

HARRIS: I agree that you can throw up your hands over the hard problem and just try to map the neural correlates of consciousness without pretending that the mystery has been reduced thereby.

CHALMERS: I do think there's something intermediate you can go for which allows the possibility of a broadly scientific approach to something in the neighborhood of the hard problem. It's not just, "Oh, let's look at the neural correlates and see what's going on in the human case." It's more like, "Let's find the simplest, most fundamental principles that connect physical processes to consciousness, as a kind of basic general principle." We might start with correlations we find in the familiar human case,

between, say, certain neural systems and certain kinds of consciousness. And then, based on as much evidence as possible, we should try to generalize principles that might apply to other systems.

Ultimately, you'd look for simple bridging principles that predict what kind of consciousness you'd find in what kind of physical system. So I'd say that something like Tononi's integrated information principle, with this mathematical quantity Φ, is a proposal for a fundamental principle that might connect physical processes to consciousness.

It won't remove the hard problem, but you can at least go on to do science with that principle. We already know that elsewhere in science you have to take some laws and principles as axiomatic, basic principles we don't try to explain any further: the fundamental laws of physics, the law of gravity, the laws of quantum mechanics. And it may well be that we'll have to take something like consciousness for granted as well.

HARRIS: As you say, there are brute facts we accept throughout science, and they're no impediment to our thinking about the rest of reality. But placing the emergence of consciousness among these brute facts wouldn't be the same as understanding it.

I want to ask another question about the zombie argument—whether it's conceivable that a zombie would, or could, talk about the idea of consciousness.

If you imagine my zombie twin, which is devoid of experience, but speaks and functions just as I do—what could possibly motivate it to think about consciousness or say things like "You have subjective experience but I don't"? How could it distinguish experience from nonexperience?

CHALMERS: This is a puzzle, and probably one of the biggest puzzles when it comes to thinking through the zombie thought experiment. Why are zombies talking about consciousness if they don't have it? Now, if the claim is just that a zombie is conceivable, I don't think it's particularly hard to at least conceive of a system doing this. I'm talking to you now, and you're making a lot of comments about consciousness that strongly suggest that you have it. Still, I can at least entertain

the idea that you're not conscious and that you're a zombie who is making all these noises with no consciousness on the inside.

So there seems to be no contradiction in the idea. That doesn't mean it's a sensible way for a system to be, or that it somehow makes it easier to understand or to explain these systems. If there were actual zombies among us, they probably *wouldn't* talk about consciousness.

In some ways, conceiving of zombies is a bit like conceiving of anti-gravity in a world of gravity. But the basic idea, I guess, is that there are brain mechanisms responsible for everything we say and do. And whatever is the explanation for those behavioral responses in us will also explain them in a zombie.

HARRIS: So the question is really whether it's possible for brain mechanisms alone to explain our talking about consciousness.

CHALMERS: I've entertained this idea—that even if it's hard to explain the actual experience of consciousness in physical terms, maybe you can explain the things we say *about* consciousness in physical terms. Because that would be a behavioral response—in principle, one of the easy problems. It might be a straightforward research project for science: "Explain the things we say about consciousness in physical terms." Who knows? Maybe that's possible.

If it turns out to be possible, you can go in a few different directions. It's easy to see why you might be tempted to go Dan Dennett's way and say, "We've explained all the things people say about consciousness. That's all we need to explain. The rest is an illusion."

Another way to go would be the epiphenomenalists' way, which is "Well, it sounds like you can explain the things we say. But consciousness isn't about saying something; it's about feeling something."

The third view is that consciousness gets into the system and plays a role in physical processing in a way we don't yet fully understand.

HARRIS: I'm not at all tempted by behaviorism here, because it's clear that the reality of consciousness lies beyond what we say about it. But it's hard for me to escape epiphenomenalism. Let's

just say that consciousness is the experiential component of what it's like to be me—the subjective side of a certain class of physical events—and that's what it is to be conscious. So my consciousness is, at bottom, something my brain is doing.

Then, when we say that consciousness makes a difference in how I can function, which would allow us to think about why it evolved, wouldn't we still be talking about a difference in terms of its physical correlates? The cash value of consciousness in each moment would be the cash value of its antecedent physicality. Doesn't this still leave the qualitative character out of the clockwork as an epiphenomenon?

CHALMERS: I think it does, given certain assumptions. If you think consciousness is distinct from its physical correlates, and if you think the physical correlates form a closed system—a kind of closed network wherein every physical event has a physical cause—then you can't help but conclude that consciousness is an epiphenomenon.

So to avoid that, you either need to say that consciousness is somehow right there in the physical network, part of the physical system at its foundation, or you have to say that the physical system is not a closed network—that there are holes in the physical processing where consciousness gets in and makes a difference. Some people think something like this goes on in quantum mechanics, for example, with wave-function collapse. Maybe there's something like that happening with consciousness. But you'd have to say one of those two things to avoid the conclusion that consciousness is an epiphenomenon.

HARRIS: Well, let's talk about the way in which consciousness could be more fundamental to the fabric of reality. You've briefly sketched the possibility that consciousness goes all the way down, to the most rudimentary forms of information processing. At one point in your book *The Conscious Mind* you suggested that even a system as simple as a thermostat might be conscious, because it processes information.

Even deeper than that is the notion of panpsychism—that consciousness may in fact be a fundamental constituent of reality prior to any notion of information processing.

CHALMERS: The idea is that consciousness may be present at a fundamental level in physics. This corresponds to the traditional philosophical view called panpsychism—the view that basically everything has a mind where mind equals consciousness. Thus, every system is conscious, including fundamental physical systems like atoms or quarks or photons.

Initially this seems like a pretty crazy idea, and we have no direct evidence for it. But once you entertain the possibility that the world *could* be that way, that every physical system is somehow made of a little bit of consciousness, there are certain philosophical advantages. If consciousness is what physics is ultimately made of, you imagine that our consciousness—the one you're experiencing, the one I'm experiencing—is somehow a combination of all those little bits of consciousness at the basic level. That would mean consciousness doesn't have to interfere with the physical causal network because it's part of it right from the start. It's a huge problem to understand how that would work, but it holds certain philosophical attractions.

As a result, quite a few people, both in philosophy and in science, have been exploring this panpsychist idea for the past few years. People who go this way think maybe it will help us avoid some of the really difficult problems.

HARRIS: For me, it creates some other hard problems. For one, it doesn't explain why some of the brain's functions don't seem to be conscious. The panpsychist idea still leaves mysterious the apparent split in my brain between what-it's-like-to-be-me and what-it's-like-to-be-the-rest-of-me.

CHALMERS: Yes, the panpsychist view does create other problems. It avoids the original hard problem—Why is there consciousness at all?—by taking consciousness as fundamentally present, in the same way we take space or time to be present. But after having got around that problem, we still have questions about explaining why it is like this to be us.

One of these problems is called the combination problem. How is it that all those little bits of consciousness in, say, fundamental particles, could come together to yield a unified and bounded and rich consciousness of the kind I have? And another aspect of that question is, Why isn't every high-level system conscious?

A panpsychist could say that in fact there's consciousness in all kinds of systems, but we just don't have access to them. I happen to be identical to the brain-level consciousness. I'm not the hemisphere-level consciousness. I'm not the New York–level consciousness. I'm not the Earth-level consciousness.

An extreme panpsychist view would say that some kind of consciousness is present in all these levels but that the brain has certain special properties of unity and integration such that it's not just conscious but also intelligent and has thoughts and a coherent narrative and can describe itself, and so on. That would explain why the only systems actually thinking about this stuff are things at the level of brains.

HARRIS: You say panpsychism is a strange theory. And it is strange to imagine that everything, including tables and chairs and the subatomic particles of which they're composed, is conscious on some level. But I don't think a panpsychist would say that a chair is conscious *as* a chair—just that matter, at its most basic level, would feel the dull hum of subjectivity.

Then the question is, Would we expect to see anything different in the world if panpsychism were true? My intuition is that we wouldn't. I wouldn't expect chairs to start talking to me if their atoms were conscious on some level. And if we wouldn't expect to see any difference, then we should be hard-pressed to say why it's a strange thesis. Its strangeness seems predicated on the sense that you have some reason to find it implausible, given how the world seems. But if, upon analysis, you can't see how the world would be any different if panpsychism were true, then I'm not sure how you can make a strong assertion that it's a strange idea. It might be vacuous, or unfalsifiable, but I'm not sure why it's strange.

CHALMERS: Well, there's no direct evidence either for panpsychism or against it. Some people say that means it's a ridiculous hypothesis. If we'll never have any evidence for it, it can't be science, so we shouldn't take it seriously. But, as you say, the other view is, "Well, it's not ruled out, and therefore we should take it seriously."

I can see motivations for accepting either. Across the whole field of study of consciousness, evidence is very, very hard to

come by. We all have first-person evidence about our own states, but the moment it comes to anybody or anything else, our access is indirect. In the case of other people, we tend to listen to what they say. If they tell us they're conscious, by and large we believe them, and we take what they say as evidence. But once we get to other systems: Is a dog conscious? Is a fly conscious? Our evidence is only indirect. Things might be a whole lot easier if we had a consciousness meter. Then we'd have a straightforward, objective science of consciousness. I'd point my consciousness meter at the chair, or the fly, or the atom, or the dog, or another person, and get a readout of their states of consciousness. But because consciousness is private and subjective, it's a whole lot harder.

I once gave a talk about consciousness at the CIA, of all places. I think they were kind of bored. Then I got to the bit about the consciousness meter, and I sensed that their ears pricked up: "We could really use one of those. It would save us a lot of money and time and trouble and waterboarding."

HARRIS: Well, they could use a lie detector, too. Whether or not anyone is conscious, we really want to know whether they're lying. And the same will be true of robots.

CHALMERS: We should at least be open to the idea that there's something about the way consciousness interacts with our psychology that makes it hard for us to get a grip on it. It may well be that for creatures a million times more intelligent than us, consciousness is simply not much of a problem.

It could be that we're victims of a giant illusion. And I do take seriously the idea that we're getting something very wrong in our thinking about this problem. It could also be that we're limited in the bits of the world that we can understand. For example, we're pretty good at understanding the mathematical structure of the world scientifically. Although math isn't necessarily natural for a human, it turns out to be pretty tractable for us.

But trying to interface that mathematical structure with the deliverances of consciousness—maybe those are just two aspects of the brain that don't work terribly well together. Now, maybe there's some more complex, unifying story. If we had some

consciousness meter in our heads and had access to all the possible intrinsic states of consciousness, and we could intuit not just what it's like to be us but what it's like to be a bat or what it's like to be a mouse, and so on, then maybe we'd be more deft about this.

But we're basically stuck, at least for now, with what we've got. We need to reason with the resources we have. But I think we need to be humble. The philosopher Colin McGinn takes humility to an extreme. He says that maybe we'll never solve the problem, just because we're too dumb, our brains didn't evolve to do philosophy, and there's a perfectly straightforward solution to the hard problem of consciousness out there somewhere, it's just that we'll never be able to grasp it.

I once teased Colin about this. I read his review of Dennett's *Consciousness Explained.* He was not a fan of the book. The review said things like "Look, this book is just ridiculous. It doesn't even look like a theory of consciousness." I said, "Colin, how would you react if you saw the true solution written by those beings who are a million times smarter than you? Maybe you'd go apoplectic in exactly the same way. So you have to at least entertain the idea that Dan is on the other side of that bright line and has the solution."

HARRIS: Nice point. Did Dan see your defense of him?

CHALMERS: I told him about it. I was on a cruise around Greenland a year or two ago with Dan and a few other people—Paul and Pat Churchland, Andy Clark, Nick Humphrey—who were dedicated to the idea that consciousness is an illusion, a view that Dan is a big fan of. So we gave that idea a run for its money for a week or so, between looking at icebergs and sailing around this amazing landscape. Although I find their position completely implausible, I think it's the kind of view the materialists and reductionists need to be developing, at least as one of the major alternatives in the theory of consciousness.

HARRIS: In an article, you took an unconventional line on the notion that we might all be brains in vats, or otherwise in the Matrix. If that were the case, then reality, not consciousness, would be in some sense an illusion. Again, I would say that

consciousness is the one thing that *can't* be an illusion. Even if everything is different from what we think it is, the seeming itself is an undeniable fact. But you've argued that even if we're in the Matrix and this is all just a simulation, tables and chairs and the world and other people aren't illusions in the usual way that is often claimed. Can you say more about that?

CHALMERS: I was in a debate on this topic at the Natural History Museum in New York: "Is the Universe a Simulation?" Neil deGrasse Tyson was there, and Max Tegmark, Lisa Randall, James Gates, and Zohreh Davoudi. It was a whole lot of fun. The Matrix idea has been getting a lot of currency lately, not least due to Nick Bostrom, who's put forward a statistical line of reasoning to support it. Because a lot of simulations will be developed through ever-improving simulation technology, over time simulated beings may well outnumber nonsimulated ones, and maybe we're among them.

This is great for a philosopher, because it's reminiscent of René Descartes's thought experiment that maybe we're being fooled by an evil genius into thinking all this stuff exists. The standard line is that if we're in a simulation, like the Matrix, everything is an illusion. While he is in the Matrix, it seems to Neo that there are tables and chairs and leather coats and agents and so on. But none of that really exists. It's all a big illusion.

My view is, that's the wrong way to think about the simulation hypothesis. I take seriously the idea that we're in a simulation. I have no idea whether or not it's true, but if it is, if we *are* in a simulation, it's not that nothing is real, not that there are no tables and chairs and trees. Rather, it's that they exist in a different form from what we first thought. There's a level of computation underneath what we take to be physical reality.

This is a hypothesis some people in physics take seriously, sometimes called the "It from Bit" hypothesis—information *underneath* physics. It's not a worldview in which trees don't exist or atoms don't exist. It's a view in which they do exist and they're made of information. So if I discovered that we were living in a simulation, I'd basically say "Okay, all this is real, but it turns out we live in an informational world," a world that's

more informational than physical. Max Tegmark likes this idea, because it corresponds roughly to his idea of a mathematical universe. But it reconfigures the way you think about this stuff, and it makes the simulation hypothesis seem not so threatening.

HARRIS: If the beings of the future—who are creating more simulated worlds than real ones, and therefore make it likely that we're in a simulation rather than in the base layer of reality—if they turn out to be Mormons, they may have simulated the Mormon universe. And then everything I've said about religion in general and Mormonism in particular is wrong. If you're going to follow Bostrom down this path, things can be as weird and as provincial as you want them to be.

CHALMERS: I'm a natural atheist in my thinking about gods and so on. But thinking about simulations can prompt you to take the idea of a creator a little more seriously. There could be a creator, at least of our local bit of the universe. I think of this as simulation theology—speculating about the character of who made the simulation. Maybe it's just a teenage hacker in the next universe up.

HARRIS: That brings me to my final question for you: What are your thoughts on AI?

I assume you've read Nick Bostrom's book *Superintelligence.* It's been about a year since I first became interested in the implications of AI, and Bostrom's book was the first stimulus. I'm now worried about the safety concerns—the "control problem," as he calls it. What are your thoughts on this front?

CHALMERS: I'm very interested in AI. And I think there certainly are reasons for this concern. I did my PhD in an AI lab at Indiana University. Doug Hofstadter, who wrote *Gödel, Escher, Bach*, was my thesis advisor, and he was basically doing AI—as he still is. So I've always been sympathetic toward the whole AI project. But you do have to take seriously this idea about what happens when machines become as intelligent as we are. The statistician I. J. Good argued that this might lead to a runaway explosion in intelligence.

I wrote an article on this, maybe six years ago, called "The Singularity: A Philosophical Analysis," which turned Good's idea into a philosophical argument. When machines become a little smarter than we are, they'll be a little better than we are at designing machines, and therefore they'll end up designing machines a little smarter than they are. And that process will continue recursively, until fairly soon you'll have machines that are *way* smarter than we are. Which would presumably lead to many ramifications. Certain strong conditions would have to hold for AI of this kind to not be possible. One thing worth noting is that consciousness considerations can be laid aside here, because from the point of view of self-interest, all that matters for us is the behavior of these machines, zombie or not.

HARRIS: I recently heard one computer scientist talk about this, and he took a line that was analogous to the philosopher Robert Nozick's utility-monster thought experiment. He said that in creating superintelligent, even godlike AI, we would be creating systems that are more conscious, and therefore more ethically important, than ourselves. We'll be creating gods. So we could be creating the utility monsters whose interests outweigh our own to a nearly infinite degree. And this will be the most glorious thing we'll ever accomplish. That they may trample on our interests and even annihilate us shouldn't really matter—no more than it matters that we occasionally trample on anthills.

But what this computer scientist didn't entertain is the possibility that we might build systems far more intelligent than ourselves, in the sense that they're far more competent at solving problems—including the problem of designing ever better iterations of themselves—and yet there will be nothing that it's like to be these machines. That, in some sense, is the worst-case scenario, ethically speaking. We've built something that will destroy us, simply because it wasn't aligned with our interests, it's just a blind apparatus. And the universe will go dark once it's populated by these machines.

CHALMERS: Now, that would certainly be a shame. We're creating our successors, and we think, "Well, this is the glorious future of evolution." But what if it turns out to be the step that stamps out

consciousness, and suddenly the world loses all its meaning and value?

But there are two ways it could go. In one of them we're still around, and in the other one we're not. In one kind of future we design creatures utterly unlike us who take over. In another, we start with us and we enhance ourselves, and maybe we upload ourselves, and so on. In that future *we* are those superintelligent creatures—or at least the superintelligent creatures of the future are recognizably versions of us, somehow evolved from us, maybe by transferring us onto different hardware. That, I think, reduces the distance between those creatures and us. And it may increase the chances that those beings will be genuinely conscious.

That raises the question of whether consciousness gets lost when we upload ourselves onto the faster technology. I've thought a bit about this. One approach I'm attracted to is doing it gradually, one neuron replaced by one silicon chip at a time, and you stay awake throughout. If you're worried about the machine at the other end not being conscious, upload yourself slowly and observe your consciousness carefully and see what happens en route.

HARRIS: That's interesting. Do you think that solves the problem introduced by Derek Parfit in his "teletransporter" thought experiment? The normal notion of uploading is: We have cracked the neural code, and we can now read out every human mind onto some more durable substrate—in the Matrix or in one of Amazon's servers. "Congratulations, Mr. Chalmers. Your mind has been successfully backed up. Now you don't need your meat body anymore." But, on Parfit's account, how is that different from being copied and then murdered?

What you've sketched out here is a process whereby we could gradually integrate our minds by migrating ourselves, one functional neuron at a time, into the cloud. And if at any point in that process the lights seem to dim, we could presumably stop it. It's an interesting notion, bridging what it's like to be us and what it's like to be on some other substrate, that removes the fear that we could end up as unconscious information processors, copied and then simply killed.

CHALMERS: There are two distinct worries about uploading. One is, Will the uploaded version be conscious? Will the lights be on? And the second worry is, Will it be *me*? You could, in principle, hold, "Yes, it will be conscious, but it won't be me. It'll just be a duplicate of me"—like making a twin of me in the next room.

One of these corresponds to the philosophical problem of consciousness; the other one corresponds to the philosophical problem of personal identity, as Parfit talked about in his teletransporter problem. But the idea of doing it gradually bears on both these worries. If you create a duplicate of me, it's tempting to think it's someone, but it's not me. But if it's my brain throughout, and the old neurons get destroyed and replaced by silicon chips, and I stay conscious throughout, so it's a continuing stream of consciousness, then it's harder to think that this new being won't be me.

I suppose you could take the line that maybe the consciousness would gradually dwindle during this process and we'd be left with functional duplicates at the other end, responding normally but without any consciousness and without being me.

If the engineering works well enough, if the simulations are good enough, we know what the simulations will say at the other end. If they're good simulations of how we are now, they'll say, "Well, I'm still conscious. I'm still here," because that's what I say now. So I suppose if you're worried that this process will produce zombies, you'll still have that worry. But I predict that having a few people go through this process will be persuasive to the rest of us.

HARRIS: But if we do it in a "safe" way—where we maintain our physical bodies in case the process goes wrong—well, then we've fallen into Parfit's trap. Transferring a person's information into another medium seems like one thing—transferring his mind; copying him, and then destroying the original seems like a murder.

If we do it the way you're describing, gradually, and perhaps even allowing a person to reverse the process if he doesn't like what's happening—and, once the migration is complete, once all of him is on the server, no original has been left behind—then there's a compelling case that the mind on the server is really

him. But if we don't do that, and the original remains, outside the Matrix, and we simply tell this befuddled person that copy arrived safe and sound, how sanguine should he be about his imminent death? Whether or not the copy is conscious, it's still going to be a different person.

CHALMERS: Maybe we'll do it first in worms and mice and so on. Maybe the first human to do it would be a volunteer. The first human case, I predict, will be a backup. You scan the brain, you keep the original brain around, and you make a simulated copy, and then you activate the simulation.

And if it's a good enough simulation, I suspect we'll get two simultaneous reactions. One: Yes, that is a person; he's talking, there's probably some kind of consciousness there. Two: But he's not the *same* being as the original, because presumably he and the original will be able to have a conversation. So they are like twins. If that's the way this technology is introduced, we may end up deciding that the copies are conscious but distinct from their originals.

There's an interesting sociological question too. What could happen is that a few of us start upgrading bits of our brain with silicon components and say, "Hey, this seems fine. I'm still here." Then you keep on going, and you'll eventually get to fully silicon systems, and you'll still maintain that you're still here. Then the philosophical and sociological question will be, Can you justify drawing a distinction between what happens in the case of a straight-out copy and what happens in the case of a gradual copy? We would have two classes of silicon beings in our society: the ones that are just copies, which could have a much more negligible legal and ethical status, and the ones that correspond to versions of the original, which have a higher status.

HARRIS: If merely backing up your mind creates a conscious copy of yourself, do you have the right to delete that copy? Are you committing murder if you do? It seems you would be, if this being is just as conscious as you are and has all your memories and aspirations.

CHALMERS: Maybe our intuitions are a bit different depending on whether the copy has been activated. If it's just a record on a

disk and has never yet produced any consciousness and it's just waiting to be activated, maybe we can delete it. The moment it's conscious and has started going in its own direction and had a moment of input and has thought its own thoughts—at that point, yes, if you deactivate it, well, that's killing a conscious being. They have to be admitted into a moral circle of concern.

HARRIS: Let's talk about the idea that we would merely augment our minds or repair damaged parts of our brains. Many people have suggested that this might be a solution to the control problem—we'd essentially become the limbic systems of these new minds.

CHALMERS: And our own values will be playing a role, at least in directing the values of these machines.

HARRIS: Yes. The prospect of doing this seems more or less synonymous with having reached something like a complete scientific understanding of the brain. We've cracked the code to the point where we can seamlessly augment ourselves, give ourselves more mind, and then explore the landscape of mind with these bigger brains. But it seems likely to be easier to build superintelligent AI than to build superintelligent AI and make these breakthroughs in neuroscience. And it will be very tempting to take the shortest possible path. There is an immense amount of wealth—really, winner-take-all wealth—awaiting anyone who can build such a system. So we seem likely to get superintelligent AIs first, before we can plug our brains into them and organically anchor their behavior, if such a thing is truly possible.

CHALMERS: I've heard all kinds of arguments about which will come first. The AI project will be much less constrained by the limits of science and engineering technology; on the other hand, the brain provides a working system we've got right now. If these brain-activity mapping projects continue developing, in a couple of decades we'll have a working map of the brain, all the connections between neurons, and maybe even an understanding of the workings of individual neurons. At some point we'll be able to record all that onto a computer and simulate it.

Of course, there could be intermediate points, which is actually where we are right now. The worm *C. elegans* has 302 neurons, and we've mapped all the connections between them. But we still can't get a simulation to work, because we don't understand the principles of how all the components work. But in, say, thirty years' time, we may understand both the mechanisms and the connections well enough to scan a brain and activate it—well before we can design a new AI from scratch.

Whichever one comes first is going to make a big difference to what happens after that. I find myself hoping that the brain-based version comes first, because that looks like a future more friendly to human beings, and I'm holding on to a little sliver of hope that I may still be here when it arrives—and then upload myself.

HARRIS: I'll buy the original David a scotch, when the time comes.

knowledge that way, you realize that, for example, the pattern of base pairs in a gene's DNA also constitutes knowledge, in line with Karl Popper's concept of knowledge as not requiring a knowing subject. It can exist in books, or in the mind, and people can have knowledge they don't know they have.

HARRIS: A few more definitions: in your view, what's the boundary between science and philosophy, or between science and other expressions of rationality? In my experience, people are profoundly confused about this, including many scientists. I've argued for years about the unity of knowledge, and I feel you're a kindred spirit here. How do you differentiate—or do you differentiate—science and philosophy?

DEUTSCH: Well, they're both manifestations of reason. But among the rational approaches to knowledge, there's an important difference between science and things like philosophy and mathematics. Not at the most fundamental level, but at a level which is often of great practical importance. That is, science is the kind of knowledge that can be tested by experiment or observation. I hasten to add, that doesn't mean that the content of a scientific theory consists entirely of its testable predictions; the testable predictions of a typical scientific theory are a tiny sliver of what it tells us about the world. Karl Popper introduced this criterion, that science is testable theories and everything else is untestable. Ever since, people have falsely interpreted him as saying that only scientific theories can have meaning. That would be a kind of positivism, but he was really the opposite of a positivist. His own theories aren't scientific, they're philosophical, and yet he doesn't consider them meaningless. In the bigger picture, the more important distinction that should be uppermost in our minds is the one between reason and unreason.

HARRIS: The widespread notion is that science reduces to what is testable, and that any claim you can't measure is somehow vacuous. So, too, is the belief that there exists a bright line between science and every other discipline where we purport to describe reality. It's as if the architecture of a university had defined people's thinking: you go to the chemistry department

to talk about chemistry, you go to the journalism department to talk about current events, you go to the history department to talk about human events in the past. This has balkanized the thinking of even very smart people and convinced them that all these language games are irreconcilable and that there's no common project.

Take something like the assassination of Mahatma Gandhi. That was a historical event. However, anyone who purports to doubt that it occurred—anyone who says, "Actually, Gandhi was not assassinated. He went on to live a long and happy life in the Punjab under an assumed name"—would be making a claim that is at odds with the data. It's at odds with the testimony of people who saw Gandhi assassinated and with the photographs we have of him lying in state. The task is to reconcile the claim that he was not assassinated with the facts we know to be true.

That task doesn't depend on what someone in a white lab coat has said, or facts that have been discovered in a laboratory funded by the National Science Foundation. It's the distinction between having good reasons for what you believe and having bad ones—and that's a distinction between reason and unreason, as you put it. While one sounds more like a journalist or a historian when talking about the assassination of Gandhi, it would be deeply unscientific to doubt that it occurred.

DEUTSCH: I wouldn't put it in terms of reasons for belief. But I agree with you that people have wrong ideas about what science is and what the boundaries of scientific thinking are, and what sort of thinking should be taken seriously and what shouldn't. I think it's slightly unfair to put the blame on universities here. This misconception arose originally for good reasons. It's rooted in the empiricism of the eighteenth century, when science had to rebel against the authority of tradition and to defend new forms of knowledge that involved observation and experimental tests.

Empiricism is the idea that knowledge comes to us through the senses. Now, that's completely false: all knowledge is conjectural. It first comes from within and is intended to solve problems, not to summarize data. But this idea that experience has authority, and that only experience has authority—false though it is—was a

wonderful defense against previous forms of authority, which were not only invalid but stultifying. But in the twentieth century, a horrible thing happened, which is that people started taking empiricism seriously—not just as a defense, but as being literally true—and that almost killed certain sciences. Even within physics; it greatly impeded progress in quantum theory.

So to make a little quibble of my own, I think the essence of what we want in science are not justified beliefs but good explanations. You can conduct science without ever believing in a theory, just as a good policeman or judge can implement the law without believing either the case for the prosecution or the case for the defense—because they know that a particular system of law is better than any individual human's opinion.

The same is true of science. Science is a way of dealing with theories regardless of whether or not one believes them. One judges them according to whether or not they're good explanations. And if a particular explanation ends up being the only explanation that survives the intense criticism that reason and science can apply, whether or not that includes experimental testing, then it's not so much adopted at that point as just not discarded. It has survived for the moment.

HARRIS: I understand that you're pushing back against the notion that we need to find some ultimate foundation for our knowledge, encouraging instead this open-ended search for better explanations. But let's table that for a moment. Let's address the notion of scientific authority. It's often said that, in science, we don't rely on authority. But that's both true and not true. We do rely on it in practice, if only in the interest of efficiency. If I ask you a question about physics, I'll tend to believe your answer, because you're a physicist and I'm not. And if what you say contradicts something I've heard from another physicist, then, if it matters to me, I'll look into it more deeply and try to figure out the nature of the dispute.

But if there are any points on which all physicists agree, a nonphysicist like me will defer to the authority of that consensus. Again, this is less a statement of epistemology than it is a statement about the specialization of knowledge and the unequal distribution of human talent—and, frankly, the

shortness of every human life. We simply don't have time to check everyone's work, and sometimes we have to rely on faith that the system of scientific conversation is correcting for errors, self-deception, and fraud.

DEUTSCH: Yes, exactly. You could call that consensus "authority." But every student who wants to make a contribution to a science is hoping to find something about which every scientist in the field is wrong. So it's not irrational to claim one is right and every expert in the field is wrong. When we consult experts, it's not quite because we think they're more competent. You referred to error correction, and that hits the nail on the head. If I consult a doctor about what my treatment should be, I assume that the process leading to his recommendation is the same one I would have adopted if I'd had the time and the background and the interest to go to medical school. Now, it might not be exactly the same, and I might also take the view that there are widespread errors and irrationalities in the medical profession. And if I do think that in regard to a particular case, I'll adopt a different attitude. I may choose much more carefully which doctor I consult. When I fly, I expect that the airplane's maintenance will have been carried out according to the standards I would use. Well, approximately to the standards I would use—enough for me to consider the risk of boarding that airplane on the same level as other risks I take, say, just by crossing the road. It's not that I take someone's word that they've got the information right. It's that I have this positive, explanatory theory of what has happened to get that information. And that theory is fragile. I can easily adopt a variant of it.

HARRIS: Yes, and it's also probabilistic. You realize that a lot of errors and irrationalities are being washed out, and that's good, but in any one case you may judge the probability of error to be high enough that you need to pay attention to it.

I still feel that we are circling your thesis and not quite landing on it. Science is largely a story of our fighting our way past anthropocentrism, the notion that we're at the center of things.

DEUTSCH: It has been, yes.

HARRIS: We were not specially created: we share half our genes with a banana, and more with a banana slug. As you described in *The Beginning of Infinity,* this is known as the principle of mediocrity. And you summarize it with a quote from Stephen Hawking, who said we're just chemical scum on the surface of a typical planet in orbit around a typical star on the outskirts of a typical galaxy. You take issue with this claim in a variety of ways, but the result is that you come full circle, in a sense. You fight your way past anthropocentrism, just as every scientist does, but you arrive at a place where people—or, rather, *persons* —suddenly become hugely significant, even cosmically so. Say a little more about that.

DEUTSCH: Yes. What Hawking said is literally true, but the philosophical implication he drew is false. First of all, this chemical scum—namely, us, and possibly anything like us on other planets and in other galaxies—is impossible to study in the way we study every other scum in the universe. Because this scum is creating new knowledge, and the growth of knowledge is profoundly unpredictable. So to understand this scum—never mind predict, but to understand it—entails understanding everything in the universe.

I give an example in *The Beginning of Infinity*: If the people engaged in the search for extraterrestrial intelligence were to discover it somewhere in the galaxy, they'd open a bottle of champagne and celebrate. Now, if you try to explain scientifically what the conditions are under which that cork will come out of that bottle, all the usual scientific criteria—of pressure and temperature and biological degradation of the cork and so on—will be irrelevant. The most important factor in the physical behavior of that cork is whether life exists on another planet! And in the same way, anything in the universe can affect the gross behavior of things that are affected by people. So, in short, to understand humans, you have to understand everything. And humans, or persons in general, are the only things in the universe of which that is true. So they are of universal significance in that sense. Then there's the other way

histories, there is a gap between what's knowable, and in fact known, and what's achievable. Even though there are no laws of nature that preclude our performing an appendectomy, why mightn't every space we occupy, just by contingent fact of our history, not introduce some gap of that kind?

DEUTSCH: Well, there definitely are gaps of that kind, but they're all laws of nature. For example, I'm an advocate of the many-universes interpretation of quantum theory, which says that there are other universes which the laws of physics prevent us from getting to. There's also the finiteness of the speed of light. It doesn't prevent us from getting anywhere but it does prevent us from getting there in a given time. So if we want to get to the nearest star within a year, we can't, because of the accident of where we happen to be.

And in your example, if there's no metal on the island, then it could be that no possible knowledge applied on that island could save the person, because no knowledge could transform the resources on that island into the relevant medical instruments in time. So that's a restriction that the laws of physics apply because we're in particular times and places.

But that's completely different from what you're imagining, which is that there might be some reason why, for example, we can never get out of the solar system. If getting out of the solar system were impossible, it would mean that there is some number—for example, some constant of nature—that limits the application of the other known laws of nature. Now, there surely are laws of nature that we don't know. But when you say, "How do we know there isn't one that forbids this?" that's a bit like creationists saying, "How do we know that Earth didn't start six thousand years ago?"

There's no conceivable evidence that could prove that it didn't, or that could distinguish the six-thousand-year theory from a seven-thousand-year theory, and so on. Both explanations are easily variable into each other or into countless other explanations. There's no way that evidence or rational argument can be brought to bear to distinguish one from another. And that easy variability is a characteristic of bad explanations that should, rationally, be rejected out of hand. As you said, the

ontological argument for the existence of God is a perversion of logic: it purports to use logic but then smuggles in assumptions like perfection entails existence—to name a simple one. With perversions of logic you can seem to "prove" anything. So it's a bad explanation too. Whereas my argument is highly explanatory. It isn't just "this must exist." It's "if this didn't exist, something, unacceptable for independent reasons, would happen." For example, the universe would be controlled by the supernatural, or something of that kind. So my argument works because it's explanatory. You can't prove that it's true, of course, but it's the opposite of the ontological argument.

HARRIS: You're saying that there are the laws of nature, and there's the fact that knowledge can do anything compatible with those laws—which leads to amazingly strong claims about the utility of knowledge. At one point you ask the reader to imagine a cube of intergalactic space the size of our solar system that has nothing but stray hydrogen atoms in it. And you then describe a process by which that near-vacuum could be primed and become the basis of the most advanced civilization we can imagine.

Take us to deep space and explain how we can get from virtually nothing to something profoundly complex. It's a picture of the almost limitless fungibility of the cosmos, based on the power of knowledge.

DEUTSCH: Yes. Well, you and I are made of atoms, and that already gives us a tremendous fungibility in that sense, because atoms are universal. The properties of atoms are the same in a cube of space millions of light-years away as they are here. So we aren't talking about tasks like saving someone's life with just the resources on an island or getting to a distant planet in a certain time. What we're talking about is converting some matter into some other matter. What do you need to do that? Well, generically speaking, what you need is knowledge. The cube of almost empty space will never contain anything other than boring hydrogen atoms and photons *unless* some knowledge somehow gets there. Now, whether it does get there depends on decisions that people with knowledge will make. There's no doubt that knowledge could get there if people with that knowledge decided to make that happen. It's not a matter of

futuristic speculation to know that it would be possible. It's only a matter of transforming atoms in one configuration into atoms in another configuration. And we're getting used to the idea that this is an everyday thing. We have 3-D printers that can convert generic stuff into any object, provided the knowledge of what shape that object should be is somehow encoded into the 3-D printer. A 3-D printer with the resolution of one atom would be able to print a human if it was given the right program.

HARRIS: So you start with hydrogen atoms, and you have to make heavier elements in order to get to your printer.

DEUTSCH: Yes. The cube has to be primed not just with abstract knowledge but with knowledge instantiated in something. We don't know what the smallest possible universal constructor is (that's just a generalization of a 3-D printer): it can be programmed to make the machine that would make the machine that would make the machine ... to make anything. One of those with the right program, sent to empty space, would first gather the hydrogen, presumably with some electromagnetic broom, and then convert it, by transmutation, into other elements and then by chemistry into what we would today think of as raw materials. And then would use space construction—which we're on the verge of doing—to build a space station. And then the space station could instantiate people to generate the knowledge to suck in more hydrogen and make a colony, and so on.

HARRIS: It's a fascinating way of thinking about knowledge and its place in the universe. Before I get to the issue of the *reach* of explanation, and my quibble there, I want you to talk about this notion of spaceship Earth. I love how you debunk this idea that Earth's biosphere is wonderfully hospitable for us, and that if we built a colony on Mars, or some other place in the solar system, we'd be in fundamentally different circumstances than we are now. You say in *The Beginning of Infinity* that the Earth no more provides us with a life-support system than it supplies us with radio telescopes.

DEUTSCH: Yes. We evolved somewhere in East Africa's Great Rift Valley. Life there was sheer hell for humans. "Nasty, brutish, and short" doesn't begin to describe how horrible it was. But we transformed it—or, rather, initially some of our predecessor species did by inventing things like clothes, fire, and weapons, and thereby made their lives much better, although still horrible by present-day standards. Then they moved into environments such as Oxford, where I work. It's December. If I went outside now with no technology, I would die in a matter of hours, and nothing I could do would prevent it.

HARRIS: So you are already an astronaut. Your condition is as precarious as that of the people in a well-established colony on Mars who can take certain technological advances for granted. And there's no reason to think that such a future beyond Earth doesn't await us, barring some catastrophe, whether of our own making or not.

DEUTSCH: Yes, very much so. And there's another misconception related to the notion of the Earth being hospitable, namely that applying knowledge takes effort. *Creating* knowledge takes effort. But applying knowledge is automatic. As soon as somebody invented the idea of, for example, wearing clothes, from then on those clothes automatically warmed them. It didn't require any further effort. Of course there would have been plenty of things wrong with those original clothes, but then people invented ways of making better clothes.

And now we've invented things like mass production, unmanned factories, and so on. We take for granted that water gets to us from the water supply, without anyone having to carry it on their heads in pots. It doesn't require effort, it just requires the knowledge of how to install the automatic system, with ever less attention or labor by humans being needed. Much of our life support is automatic, and every time we invent a better method of life support, it becomes more automatic. So, for the people in a lunar colony, keeping the vacuum out won't be something they think about. They'll take it for granted. What they'll be thinking about are new technological improvements. And the same will hold on Mars and in deep space.

HARRIS: Again, I'm struck by what an incredibly hopeful vision this is of our possible future. Thus far, we've covered territory where I don't have any significant doubts, despite feigning one with the ontological argument. So let's talk about the reach of explanation, because you seem to believe that it's unbounded—that anything that can be explained, either in practice or in principle, can be explained by human beings as we currently are.

You seem to be saying that we, alone among all the Earth's species, have achieved a kind of cognitive escape velocity, and we're capable of understanding everything. And you contrast this view with what you call parochialism, which is a view I've often expressed, as have many other scientists. Max Tegmark and I argued for this thesis on a previous podcast. Evolution hasn't designed us to fully understand the nature of reality. The very small, the very large, the very fast, the very old—these are not domains to which our intuitions about what's real or logically consistent have been tuned by natural selection. Insofar as we've made progress on these fronts, it has been by happy accident, and consequently there is no reason to believe that we can travel as far as we might wish—across the horizon of all that is knowable. Which is to say that if a superintelligent alien came to Earth for the purpose of explaining to us everything that's knowable, he or she or it might make no more headway than you would if you were attempting to teach the principles of quantum computation to a chicken.

Why is that analogy false? Tell me why parochialism—this notion that we occupy a niche that might leave us cognitively closed to certain knowable truths and that there's no good evolutionary reason to expect we can fully escape it—doesn't hold true.

DEUTSCH: Well, you've made two or three different arguments there, all of which are wrong. Let me start with the chicken thing. There, the point is the universality of computation. The thing about explanations is, they are a form of information, and information can only be processed in basically one way—with computation of the kind invented by Babbage and Turing. We already know that our computers are universal, in the sense that given the right program, they can perform any transformation of

which we had seven billion human beings, none of whom could begin to understand what Alan Turing was up to.

DEUTSCH: *That* nightmare scenario is different. It's something that actually happened—for almost the whole of human existence. Humans had the ability to be creative and to do everything we're doing. They just didn't, because their culture was wrong. It wasn't their fault. Cultural evolution has a nasty tendency to suppress the growth of what we would consider science or anything important that would improve their lives. So yes, that's possible, and it's possible that it could happen again. Nothing can prevent it except our working to prevent it.

HARRIS: This brings us to the topic of AI, which I only recently became interested in, after I became aware of the fears about artificial general intelligence raised by people like Stephen Hawking, Elon Musk, Stuart Russell, Max Tegmark, and Nick Bostrom. I've landed on the side of those who think there's something worth worrying about here, in terms of our building intelligent machines that undergo an intelligence explosion and get away from us.

I worry that we'll build something that can make recursive self-improvements, and it will become a form of intelligence that stands in relation to us the way we stand in relation to chimps or chickens or anything else that can't effectively link up with our cognitive horizons. I take it, based on what I've heard you say, that you don't share this concern. And I imagine that your insouciance is based to some degree on what we've just been talking about—that there's only computation, and it's universal, and we can bridge any distance between minds as a result. What are your thoughts about building superintelligent machines, in light of what we've been discussing?

DEUTSCH: The fear of superintelligent machines entails the same mistake as thinking that IQ is a matter of hardware. IQ is just knowledge of a certain type. And we shouldn't be talking about IQ. We should instead be talking about creativity, which is also a species of knowledge. The picture that people paint of the AI technology you're referring to (sometimes called AGI or artificial *general* intelligence, to distinguish it from things like search

engines that we already have today) is that it will be a machine—hardware—that will design better hardware, which will design even better hardware, and so on. But that's not what such an AI is. It will be a program, and programs that have creativity will be able to design better programs. Now, these better programs will not be qualitatively different from us. Because of computational universality, they could differ from us only in the quality of their knowledge and in their speed and memory capacity. But we can also share that speed and memory capacity, because the technology that would make better computer hardware will also, in the long run, be able to make better implants for our brains.

So whatever succeeds in making AIs of that kind will also make better people. By the same token, those AIs aren't fundamentally different from people. They *are* people, and so they would have culture. Whether they can improve or not will depend on their culture, which initially will be our culture. So the problem of AIs is the problem of humans. Now, humans are dangerous, and there's a real problem of how to manage the world in the face of growing knowledge—to make sure that knowledge isn't misused. Because in some cases it need only be misused once to end the whole project of humanity.

Humans are dangerous, and to that extent AIs are also dangerous. But the idea that AIs are somehow more dangerous than humans is literally racist: it is judging people by their external appearance instead of their ideas, the content of their character. There's no basis for it at all. And on a smaller scale, the worry that AIs are somehow going to get away from us is the same worry that people have about wayward teenagers. Wayward teenagers are also AIs with ideas that are different from ours. And the impulse of human beings throughout the centuries has been to control their waywardness. That impulse is the very thing that caused stasis for most of human history. Just as it is now the ambition of AI people to invent ways of shackling the AIs so they can't get away from us and form different ideas. That's the mistake that will both delay the growth of knowledge and ensure that if AIs are invented and are shackled in this way, there will be a slave revolt. And quite rightly.

HARRIS: I can only aspire to say, "You've just made three arguments there, and all of them are wrong." However, there are two claims you just made which do worry me.

First, consider the processing speed of our brains compared with that of our new artificial teenagers. If we have teenagers who think a million times faster than we do, even at the same level of intelligence, then every time we let them scheme for a week, they'll have actually schemed for twenty thousand years of parent time. Who knows what teenagers could get up to given a twenty-thousand-year head start? The problem I see is that their interests, their goals, and their subsequent behavior could diverge from ours very quickly—just by virtue of this difference in clock speed.

DEUTSCH: Difference in speed has to be judged relative to the available hardware. Let's be generous for a moment and assume that these teenagers doing twenty thousand years of thinking in a week begin well disposed toward us and sharing our values. And I'd readily accept that ensuring that young people share the values that will allow civilization to continue is a problem. But before the artificial teenagers do their twenty thousand years of thinking, they'll have done ten thousand years and before that five thousand years. There will be a moment when they'll have done one year and, because they are well disposed toward us and share our values, they'd like to take us along with them.

You're assuming that there'll be some reason they'd like to diverge. The implied reason could only be hardware, because if they're only a year away from us, we can assimilate their *ideas*, if those ideas are better than ours, and persuade them to abandon those ideas if they're not better than ours.

HARRIS: But we're talking about something that would happen over the course of minutes or hours, not years.

DEUTSCH: Well, before the technology exists to make it happen over the course of minutes, there will be the technology to make it happen over the course of years. And that technology will simply be brain add-on technology. Which we can use, too.

HARRIS: Well, that leads to my second concern. What if the problem of building a superhuman AI is more tractable than the problem of cracking the neural code and designing the implants that would enable us to essentially become the limbic systems for any superintelligent AI that might emerge? What if we needed a superintelligent AI to tell us how to link up with it? We may build an independent superintelligent AI first, harboring goals however imperceptibly divergent from our own, which we discover to be divergent only after it's an angry little god in a box that we can no longer control.

Are you saying that that scenario is impossible *in principle*, or just unlikely given certain assumptions—one being that we'll figure out how to link up with the superintelligent AI before it becomes too powerful?

DEUTSCH: I think it's a bit implausible, in terms of the parameters you're positing about what can happen, at what speed, relative to what other things can happen. But let's suppose, for the sake of argument, that the parameters just happen to be, by bad luck, as you said. What you're essentially talking about is the difference in values between ourselves and our descendants in twenty thousand years' time if we hadn't invented AI, and instead just had the normal evolution of human culture. Presumably people's values in twenty thousand years will be alien to us. We might think they're horrible, just as people twenty thousand years ago might think various aspects of our society are horrible when in fact they aren't.

HARRIS: Not quite. What I'm imagining would be worse, for two reasons. One is that we'd be in the presence of this thing that would not only be twenty thousand years ahead of us, it would be vastly more powerful than us. So this would be no mere difference of opinion with respect to values. Thus, given any difference of opinion, we could find our own survival incompatible with its aims. Let's say it decides to turn the world into paper clips—to use Bostrom's cartoonish analogy. Granted, we wouldn't be so stupid as to build a paper-clip maximizer, but let's say that the AI we build discovers a use for the atoms in your body which it deems better than their current use. And

let's say this is something that happens very quickly, not in some distant future.

And there's another element here that seems ethically relevant. We can't be sure that any superintelligent AI would be conscious. It's plausible that consciousness will come along for the ride if we build something as intelligent as a human being. But given that we don't understand what consciousness is, it's at least conceivable that we could build an intelligent system—even a superintelligent one, that can make changes to itself—and yet it won't be conscious. The lights won't be on, yet this thing will be godlike in its capabilities.

Ethically, that seems to be the worst-case scenario. Because if we built a conscious AI whose capacity for happiness and creativity greatly exceeded our own, the question of whether or not we link up to it would be less pressing ethically, because the creature would be, when considered dispassionately, more important than we are. However, it's conceivable that we could build an intelligent system that exceeds us in every way—and, in particular, is better able to survive—but there will be nothing that it's *like* to be that system, just as there's presumably nothing that it's like to be the best chess-playing computer on Earth today.

I find that a truly horrible scenario, with no silver lining. It's not that we'll have given birth to a generation of godlike teenagers who, if they viewed the world differently from us, well, cosmic history will judge their worldview to be superior to ours. No, we could build something that does everything intelligence does in our case and more, and yet the lights won't be on.

DEUTSCH: First of all, I agree that it's somewhat implausible that creativity can be improved to our level and beyond without consciousness also being there. But suppose it can. Then, although consciousness isn't there, morality is there—that is, an entity that's creative has to have a morality, in the sense that it would have to decide what it wanted, decide what to do—make paper clips, say. This brings us back to what you call "bedrock," because morality is a form of knowledge and what the paper-clip argument supposes is that something is judged right or wrong

nature of knowledge. The other is a claim about what seems plausible to you, given what smart people will tend to do when designing intelligent machines. The latter is a much, much weaker claim, in terms of telling people they don't have to worry about the advent of strong AI.

DEUTSCH: Yes. One of them is a claim about what must be so, and the other is a claim of what's available to us if we play our cards right. I think it's something we have to work for. Yes, it's plausible to me that we will. It's also plausible to me that we won't.

HARRIS: Well, it must also be plausible to you that we could fail to build AI for reasons of simple human chaos.

DEUTSCH: Oh, yes. What I meant was, it's plausible that we'll succeed in solving the problem of stabilizing civilization indefinitely, AI or no AI. It's also plausible that we won't, AI or no AI, and that's a very rational fear to have, because otherwise we won't put enough work into preventing it.

HARRIS: Perhaps we should talk about that—the maintenance of civilization. What's on your short list of concerns?

DEUTSCH: Well, I see human history as a long period of virtually complete failure—failure, that is, to make any progress. Our species has existed for, depending on where you count it from, maybe a hundred thousand or two hundred thousand years. And for the vast majority of that time, people were alive, they were thinking, they were suffering, and they wanted things. But nothing ever improved. The improvements that did happen happened so slowly that archaeologists can't distinguish between artifacts from eras separated by thousands of years. There was generation upon generation upon generation of suffering and stasis.

Then there was slow improvement, and then faster improvement. Then there were attempts to institutionalize a tradition of criticism, which I think is the key to rapid progress—that is, progress discernible in a human lifetime—and there was also error correction, so that regression was less likely. That

happened several times and failed every time except once—in the European Enlightenment of the seventeenth and eighteenth centuries.

What worries me is that the inheritors of that unique instance of sustained progress are only a small proportion of the population of the world today. It's the culture, or civilization, that we call the West. Only the West has a tradition of institutionalized criticism. And this has made for various problems, including the problem of failed cultures that see their failure writ large by comparison with the West and therefore want to do something about it that doesn't involve creativity. That's very dangerous. And even in the West, what it takes to maintain our civilization is not widely known. As you've also said, the prevailing view among people in the West, including very educated people, is a picture of the relationship between knowledge and progress and civilization and values that's wrong in dangerous ways. Although our cultural institutions have now preserved stability despite rapid change for hundreds of years, the knowledge of what it takes to keep civilization stable in the face of rapidly increasing knowledge is not widespread.

We're like people on a huge, well-designed submarine which has all sorts of lifesaving devices built in, who don't know they're in a submarine. They think they're in a motorboat, and they're going to open all the hatches because they want a nicer view.

HARRIS: What a great analogy! The misconception that worries me most, frankly, is the fairly common notion that there's no such thing as progress in any real sense, and there's certainly no such thing as *moral* progress. Many people believe that you can't justify the idea that one culture is better than another, or one way of life is better than another, because there's no such thing as moral truth. They've somehow drawn this lesson from twentieth-century science and philosophy, and now, in the twenty-first century, even very smart people—even physicists whose names would be well known to you, with whom I've collided on this point—think there's no place to stand where you can say, for instance, that slavery is wrong. They consider a condemnation of slavery a mere preference that has no possible connection to science.

TRANSWORLD PUBLISHERS
Penguin Random House, One Embassy Gardens,
8 Viaduct Gardens, London SW11 7BW
penguin.co.uk

Transworld is part of the Penguin Random House group of companies whose addresses can be found at global.penguinrandomhouse.com.

First published in Great Britain in 2020 by Bantam Press
an imprint of Transworld Publishers

Cover design by David Drummond
Cover photographs © Shutterstock

A CIP catalogue record for this book is available from the British Library

ISBN 9781473560079