SAM HARRIS

THE BLOG

Surviving the Cosmos

Podcast Transcript

Consciousness | Ethics | Philosophy | Physics | Science | August 21, 2016

(Photo via NASA)

In this episode of the Waking Up podcast, Sam Harris talks to physicist David Deutsch about the reach and power of human knowledge, the future of artificial intelligence, and the survival of civilization.

David Deutsch is best known as the founding father of the quantum theory of computation, and for his work on Everettian (multiverse) quantum theory. He is a Visiting Professor of Physics at Oxford University, where he works on “anything

fundamental." At present, that mainly means his proposed constructor theory. He has written two books – *The Fabric of Reality* and *The Beginning of Infinity* – aimed at the general reader.

* * *

Welcome to the Waking Up Podcast. This is Sam Harris. Today I'm speaking with David Deutsch. David is a physicist at Oxford. He's a professor of physics at the Center for Quantum Computation at the Clarendon Laboratory. He works on the quantum theory of computation and information, and he is a very famous exponent of the many-worlds interpretation of quantum mechanics, neither of which do we talk about in this interview.

David has a fascinating and capacious mind, as you will see. We talk about much of the other material in his most recent book, *The Beginning of Infinity*, but we by no means cover all its contents. As you'll see, David has a talent for expressing scientifically and philosophically revolutionary ideas in very simple language. And what you'll often hear in this interview is me struggling to go back and unpack the import of some very simple-sounding statements, which I know those of you unfamiliar with his work can't parse the way he intends. In any case, I hope you enjoy meeting David Deutsch as much as I did.

SH: I have David Deutsch on the line. David, thank you for coming on the podcast.

DD: Oh, thank you very much for having me.

SH: I don't know what part of the multiverse we're in where I complain about jihadists by night and talk to you by day, but it's a very strange one. In any case, we're about to have a very different kind of conversation than I've had of late, and I really have been looking forward to it. I spoke to Steven Pinker and told him that you were coming on the podcast, and he claimed that you are one of his favorite minds on the planet. I don't know if you know Steve, but that's high praise indeed.

DD: I don't know him personally, but that's very kind of him to say that.

SH: So let me begin quite awkwardly with an apology, in addition to the apology that I just gave you off-air for being late. While I aspired to read every word of *The Beginning of Infinity* before speaking with you, I've only read about half. Not just the first half—I jumped around a bit. But forgive me if some of my questions and comments seem to ignore some of the things you had the good sense to write in that book, and that I didn't have the good sense to read.

Not much turns on this, because, as you know, you have to make yourself intelligible to our listeners, most of whom will not have read any of the book. But I just want to say it really is a remarkable book. Both philosophically and scientifically, it is incredibly deep, while also being extremely accessible.

DD: Thanks.

SH: And it is a profoundly optimistic book in at least one sense. I don't think I've ever encountered a more hopeful statement of our potential to make progress. One of the consequences of your view is that the future is unpredictable in principle. The problems we will face are unforeseeable, and the way we will solve these problems is also unforeseeable. And problems will continue to arise, of necessity, but problems can be solved.

This claim about the solubility of problems with knowledge runs very, very deep. It's a far stronger claim than our listeners will understand based on what I've just said.

DD: That's a very nice summary.

SH: It's interesting to think about how to have this conversation, because what I want to do is creep up on your central thesis. I think there are certain claims you make, claims specifically about the reach and power of human knowledge, that are fairly breathtaking. And I find that I want to agree with every word of what you say here because, again, these claims are so hopeful. But I have a few quibbles. It's interesting to go into this conversation hoping to be relieved of my doubts about your thesis. I'm hoping that you'll perform an exorcism on my skepticism, such as it is.

DD: Sure. Well, I think the truth really is very positive, but I should say at the outset that there is one fly in the ointment, and that is that because the future is unpredictable, nothing is guaranteed. There is no guarantee that civilization will survive or that our species will survive, but there is, I think, a guarantee that we can, and also we know in principle how to.

SH: Before we get into your claims there, let's start the conversation somewhere near epistemological bedrock. I'd like to ask a few questions designed to get to the definitions of certain terms, because you use words like "knowledge," and "explanation," and even "person" in novel ways in the book, and I want our listeners to be awake to how much work you're requiring these words to do. Let's begin with the concept of knowledge. What is knowledge, and what is the boundary between knowledge and ignorance, in your view?

DD: Yes, so there are several different ways of approaching that concept. The way I think of knowledge is broader than the usual use of those terms, and yet paradoxically closer to the commonsense use of the term, because philosophers have almost defined it out of existence. Knowledge is a kind of information. That's the simple thing. It's something which could have been otherwise and is one particular way, and the particular *what* it is, is that it says something true and useful about the world.

Now, knowledge is, in a sense, an abstract thing, because it's independent of its physical instantiation. I can speak words which embody some knowledge, I can write them down, they can exist as movements of electrons in a computer, and so on—thousands of different ways. So knowledge isn't dependent on any particular instantiation. On the other hand, it does have the property that when it *is* instantiated, it tends to remain so. So, let's say, a piece of speculation by a scientist, which he writes down, that turns out to be a genuine piece of knowledge—that will be the piece of paper that he does not throw in the wastepaper basket. That's the piece that will be published, and that's the piece which will be studied by other scientists, and so on.

So it is a piece of information that has the property of keeping itself physically instantiated, causing itself to be physically instantiated once it already is. Once you think of knowledge that way, you realize that, for example, the pattern of base pairs in the DNA of a gene also constitute knowledge, and that in turn connects with Karl Popper's concept of knowledge, which is knowledge that doesn't have to have a knowing subject. It can exist in books abstractly, or it can exist in the mind, or people can have knowledge that they don't even know they have.

SH: I want to get to the reality of abstractions later on, because I think that is very much at the core of this. But a few more definitions: What is the boundary between science and philosophy or other expressions of rationality, in your view? Because in my experience, people are profoundly confused by this, and many scientists are confused by this. I've argued for years

about the unity of knowledge, and I feel that you are a kindred spirit here. How do you differentiate, or decline to differentiate, science and philosophy?

DD: Well, as you've just indicated, I think that science and philosophy are both manifestations of reason, and the real difference that should be uppermost in our minds between different kinds of ideas and between different ways of dealing with ideas is the difference between reason and unreason. But among the rational approaches to knowledge or different kinds of knowledge, there is an important difference between science and other things, like philosophy and mathematics.

Not at a really fundamental level, but at a level which is of great practical importance often: That is, science is the kind of knowledge that can be tested by experiment or observation. Now, I hasten to add that that does not mean that the content of a scientific theory consists entirely in its testable predictions. On the contrary, the testable predictions of a typical scientific theory are just a tiny, tiny sliver of what it tells us about the world. Now, Karl Popper introduced his criterion of demarcation between science and other things—namely, that science is testable theories, and everything else is untestable.

Ever since he did that, people have falsely interpreted him as a kind of positivist (he was really the opposite of a positivist), and if you interpret him like that, then his criterion of demarcation becomes a criterion of meaning. That is, he's interpreted as saying that only scientific theories can have meaning.

SH: This is sometimes referred to as "verificationism."

DD: Yes. So he's called a falsificationist to distinguish him from the other verificationists. But of course he isn't. It's a completely different conception, and his philosophical theories themselves are philosophical theories, and yet he doesn't consider them meaningless; quite the contrary. So that's the difference between science and other things that comes up when people pretend to have the authority of science for things that aren't science. But on the bigger picture, the more important demarcation is between reason and unreason.

SH: I want to go over that terrain you just covered a little bit more, because you made some points there that I think are a little hard for listeners who haven't thought about this a lot to parse. So for instance, this notion that science reduces to what is testable. This belief is so widespread, even among high-level scientists, that anything else—anything which you cannot measure immediately—is somehow a vacuous claim. In principle, the only way to make a credible claim or even a meaningful claim about reality is to essentially give a recipe for observation that is immediately actionable. It's an amazingly widespread belief.

So, too, is a belief in a bright line between science and every other discipline where we purport to describe reality. And it's like the architecture of a university has defined people's thinking. So the fact that you go to the chemistry department to talk about chemistry, and you go to the journalism department to talk about current events, and you go to the history department to talk about human events in the past—these separate buildings have balkanized the thinking of even very smart people and convinced them that all these language games are irreconcilable and that there is no common project.

I'll just bounce a few examples off you that some of our listeners will be familiar with, but I think they make the point. Take something like the assassination of Mahatma Gandhi. Now, that's a historical event. However, anyone who would purport 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 about terrestrial reality that is at odds with the data. It's at odds with the testimony of people who saw him assassinated and with the photographs we have of him lying in state.

There's an immense burden of reconciling this claim about history with the facts that we know to be true.

And the distinction is not between what someone in a white lab coat has said, or facts that have been brought into view in the context of a scientific laboratory funded by a National Science Foundation grant. It is 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. So one could say that the assassination of Gandhi is a historical fact, but it's also a scientific fact.

It is just a fact, even though science doesn't usually deal in assassinations, and you're more a journalist or a historian when you talk about this sort of thing being true. It would be deeply unscientific at this point to doubt that it occurred.

DD: Yes. Well, I'd say that it's deeply irrational to claim that it didn't occur, yes. And I wouldn't put it in terms of reasons for belief either. I agree with you that people have very wrong ideas about what science is and what the boundaries of scientific thinking is, 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 quite good reasons. It's rooted in the empiricism of the 18th century where science had to rebel against the authority of tradition and try to give dignity and respect to 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, and it comes from within at first 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. So it was a good defense but not actually true. And in the 20th century, a horrible thing happened, which is that people started taking it seriously not just as a defense, but as being literally true, and that almost killed certain sciences. Even within physics, I think, it greatly impeded the progress in quantum theory.

So just to come to a little quibble of my own, I think the essence of what we want in science is good explanation, and there's no such thing as a good reason for a belief. A scientific theory is an impersonal thing. It can be written in a book. One can conduct science without ever believing the theory, just as a good policeman or judge can implement the law without ever believing either of the cases for the prosecution or the defense, just because they know that a particular system is better than any individual human's opinion.

And the same is true of science. Science is a way of dealing with theories, regardless of whether one believes them. One judges them according to whether they are good explanations. And there need not ever be any such process as accepting a theory, because it is conjectured initially, and takes its chances, and is criticized as an explanation. If, by some chance, a particular explanation ends up being the only one that survives the intense criticism that science has learned how to apply, then it's not adopted at that point—it's just not discarded.

SH: Right, right. I think we may be stumbling over a semantic difference in terms like "reasons" and "reasons for belief" or a "justification" for belief. I understand that you're pushing back against this notion that we need to find some ultimate foundation for our knowledge, rather than this open-ended effort at explanation. But let's table that for a second, because obviously your notion of explanation is at the core here. Again, I want to sneak up on it, because I don't want to lose some of the detail with respect to the ground we've already covered.

Let's come back to this notion of scientific authority. It seems to me there's a lot of confusion about the nature of scientific

authority. It's often said in science that we don't rely on authority, and that's both true and not true. When push comes to shove, we don't rely on it, and you make this very clear in your book. But we do rely on it in practice, if only in the interest of efficiency. So if I ask you a question about physics, I will 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—well then, if it matters to me, I will 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 non-physicist like myself 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 we have to rely sometimes on faith that the system of scientific conversation is correcting for errors, self-deception, and fraud. Did I get myself out of the ditch there?

DD: Yes, yes, exactly. At the end, what you said was right. So you could call this authority. It doesn't matter really what words we use, but every student who wants to make a contribution to a science is hoping to find something where every scientist in his field is wrong.

SH: Absolutely.

DD: So it's not impossible to take the view that you're right and every expert in the field is wrong. What happens when we consult experts, whether or not you use the word "authority," it's not quite that we think they're more competent. I think when you refer to error correction, that hits the nail on the head. There is a process of error correction in the scientific community that approximates what I would use if I had the time, and the background, and the interest to pursue it there.

So if I go to a doctor to consult him about what my treatment should be, I assume that by and large the process that has led to his recommendation is the same as the process that I would have adopted if I had been present at all the stages. Now, it's not exactly the same, and I might also take the view that there are widespread errors and widespread irrationalities in the medical profession. And if I think that, then I will adopt a rather different attitude. I may choose much more carefully which doctor I consult and how my own opinion should be judged against the doctor's opinion in a case where I think that the error correction hasn't been up to the standard I would want.

This is not so rare. As I said, every student is hoping to find a case of this in their own field. When I travel on a plane, I expect that the maintenance will have been carried out to the standards that I would use. Well, approximately to the standards that I would use—enough for me to consider that risk on the same level as other risks that I take just by crossing the road. It's not that I'm sure. It's not that I take their word for it in any sense. It's that I have this positive theory of what has happened to get that information to the right place. That theory is fragile. I can easily adopt a variant of it.

SH: Yeah, and it's also probabilistic. You realize that a lot of these errors are washing out, and that's a good thing, but in any one case you may judge the probability of error to be high enough that you need to really pay attention to it. And often, as you say, that happens in a doctor's office, where you're not hoping to find it.

Again, I still picture us circling your thesis and not yet landing on it. Science is largely a story of our fighting our way past anthropocentrism, this notion that we are at the center of things.

DD: It has been, yes.

SH: We are not specially created: We share half our genes with a banana, and more than that with a banana slug. As you described in your book, this is known as the principle of mediocrity. And you summarize it with a quote from Stephen Hawking, who said, “We are just chemical scum on the surface of a typical planet that’s in orbit around a typical star, on the outskirts of a typical galaxy.” Now, you take issue with this claim in a variety of ways, but the result is that you come full circle in a way. You fight your way past anthropocentrism the way 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.

DD: Yes. Well, that quote from Hawking is literally true, but the philosophical implication he draws is completely false. One can approach this from two different directions. First of all, that chemical scum—namely us, and possibly things like us on other planets and other galaxies and so on—if it exists, to study it is impossible, unlike every other scum in the universe. Because that scum is creating new knowledge, and the growth of knowledge is profoundly unpredictable. So as a consequence of that, to understand this scum—never mind predict, but to understand it, to understand what’s happening here—entails understanding everything in the universe.

I give an example in the book: If the people at the SETI project were to discover extraterrestrial life somewhere far away in the galaxy, they would open their bottle of champagne and celebrate. Now, if you try to explain scientifically what are the conditions under which that cork will come out of that bottle, the usual scientific criteria that you use, 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 bottle 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 people 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 around. It’s also true that the reach of human knowledge and human intentions on the physical world is unlimited.

So we are used to having a relatively tiny effect on this small, insignificant planet, and to the rest of the universe being completely beyond our ken. But that’s just a parochial misconception, really, because we haven’t set out across the universe yet. And we know that there are no limits on how much we can affect the universe if we choose to. So in both those senses, there’s no limit to how important we are—by which I mean we and the ETs and the AIs, if they exist. We are completely central to any understanding of the universe.

SH: Once again, I’m struggling with the fact that I know how condensed some of your statements are, and I also know that it’s impossible for our listeners to appreciate just how much knowledge and conjecture is being smuggled into each one. So let’s just deal with this concept of explanation and the work it does.

First, you make a few points about explanation that I find totally uncontroversial and even obvious, but which are in fact highly controversial in educated circles. One is this notion that, as you say, explanation is really what lies at the bedrock of the scientific enterprise—the enterprise of reason, generally. Explanations in one field of knowledge potentially touch explanations in many other fields, even all other fields, and this suggests a kind of unity of knowledge. But you make two especially bold claims about explanation, which I do see some reason to doubt. And as I’ve said, I’d rather not doubt them because they’re incredibly hopeful claims.

I'll divide these into the *power* of explanation and the *reach* of explanation. These may not be entirely separate in your mind, but there's a distinct emphasis on each of these features.

You make what an extraordinary claim about explanation, which at first seems quite pedestrian. You say that there's a deep connection between explaining the world and controlling it. Everyone understands this to some degree. We see the evidence of it all around us in our technology, and people have this phrase "Knowledge is power" in their heads. So there's nothing so surprising about that. But you go on to suggest—and you did just suggest it in passing a moment ago—that knowledge confers power without limit, or it is limited only by the laws of nature. So you actually say that anything that isn't precluded by the laws of nature is achievable, given the right knowledge. Because if something were *not* achievable, given complete knowledge, that itself would be a regularity in nature that could be explained in terms of the laws of nature. So there are really only two possibilities. Either something is precluded by the laws of nature, or it is achievable with knowledge. Do I have you right there?

DD: Yes, and that's what I call the momentous dichotomy. There can't be any third possibility. And I think you've given a very short proof of it right there.

SH: But to play Devil's advocate for a moment: How isn't this just a clever tautology, analogous to the ontological argument proving the existence of God? Many of our listeners will know that according to St. Anselm and Descartes and many others, you can prove the existence of God simply by forcing your thoughts about Him to essentially bite their own tails. For instance, I could make the following claim: I can form a clear and distinct concept of the most perfect possible being, and such a being must therefore exist, because a being that exists is more perfect than one that doesn't. And I've already said I'm thinking about the most perfect possible being, so existence is somehow a predicate of perfection.

Now, of course, most people—certainly most people in my audience—will recognize that this is just a trick of language. It could be used to prove the existence of anything. I could say, "I'm thinking of the most perfect chocolate mousse. Therefore, it must exist, because a mousse that exists is more perfect that one that doesn't. And I already told you that I'm thinking of the most perfect possible mousse."

What you're saying here doesn't have the same structure, but I do worry that you could be performing a bit of a conjuring trick here. For instance, why mightn't certain transformations of the material world be unachievable even in the presence of complete knowledge, merely by (and I realize you do anticipate this in your book, but I want you to flesh it out for the listeners), let's say, a contingency of geography?

For instance, you and I are on an island, and one of our friends comes down with appendicitis. And let's say you and I are both competent surgeons. We know everything there is to know about removing a man's appendix, but it just so happens we don't have any of the necessary tools, and everything on that particular island has the consistency of soft cheese. By sheer accident of our personal histories, there is a gap between what is knowable and in fact known and what is achievable. Even though there are no laws of nature that preclude our performing an appendectomy on a person, why mightn't every space we occupy, just by contingent fact of our history, not introduce some gap of that kind?

DD: Well, there definitely are gaps of that kind, and they're all laws of nature. For example, I am 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, which doesn't prevent us from actually getting anywhere,

but it does prevent us from getting anywhere in a given time. So if we want to get to the nearest star within a year, we can't do so, because of the accident of where we happen to be. If we happened to be nearer to it, we could easily get there in a year.

And in your example, if there's no metal on the island, then it could easily be that no knowledge present on that island could save the person, because no knowledge could transform the resources on that island into the relevant medical instruments. So that's a restriction that the laws of physics apply because we are in particular times and places, and because the most powerful thing is that we don't in fact have the knowledge to do most of the things we would ideally like to do.

But that's completely different, I think, 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—1,000 astronomical units, or something—that limits the other laws of nature we already know. Now, there might be other laws of nature. When you say, "How do we know there aren't?" that's a little bit—if I can turn your objection round the other way—like creationists saying, "How do we know that Earth didn't start 6,000 years ago?"

There is no conceivable evidence that could prove that it didn't, or that could distinguish the 6,000-year theory from a 7,000-year theory, and so on. There's no way that evidence can be brought to bear on that. And that leads us to explanation again, which is another difference between my argument, which I think is valid, and the ontological argument for the existence of God. As you said, it's a perversion of logic. The argument purports to use logic but then smuggles in assumptions like perfection entails existence, for example, to name a simple one. Whereas my proof, as it were, is an explanatory one.

It isn't just "This must exist." It's that if this didn't exist, something bad would happen. For example, the universe would be controlled by the supernatural, or the laws of nature would not be explanatory, or something of that kind, which I think is just leading to the supernatural in a different way. So the argument works because it's explanatory. You can't prove that it's true, of course, but there isn't a hole in it of the same kind as in the ontological argument.

SH: Okay. You're saying that we could have a complete understanding of the laws of nature, and yet there could be many contingent facts about where we are—let's say, our current distance from a star we want to get to—which would preclude our doing anything especially powerful with this knowledge. And you're going to shuttle those contingent facts back into this claim that, well, this is just more of the laws of nature. These facts about us are regularities in the universe which are themselves explained by the laws of nature, and therefore we're back to this dichotomy. There are the laws of nature, and there's the fact that knowledge can do anything compatible with those laws.

In various thought experiments in your book, you make amazingly powerful claims about the utility of knowledge. So for instance, at one point you say that a region of empty space—a cube the size of the solar system on all sides—is more representative of the universe as it actually is, which is to say nearly a vacuum. We're talking about a cube of intergalactic, space that has more or less nothing but stray hydrogen atoms in it. And you then describe a process by which that near vaccum could be primed and become the basis of the most advanced civilization we can imagine.

Please take us to deep space and spend a minute or two talking about how you get from virtually nothing to something. It's a picture of the almost limitless fungibility of the universe based on the power of knowledge.

DD: Yes. So you and I are made of atoms, and that already gives us a tremendous fungibility, because we know that atoms

are universal. The properties of atoms are the same in this cube of space that is millions of light-years away as they are here. So we aren't talking about the power of knowledge to achieve things to control the world. We're not 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.

The generic thing that 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. What has to happen is that this cube of almost empty space will never turn into anything other than boring hydrogen atoms unless some knowledge somehow gets there. Now, whether knowledge gets there or not depends on decisions that people with knowledge will make at some point. I think there is no doubt that knowledge could get there if people with knowledge decided to do that for some reason.

I can't actually think of a reason, but if they did want to do that, it's not a matter of futuristic speculation to know that it would be possible. Then it's a matter of transforming atoms in one configuration into atoms in another configuration. And we're now 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 that 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.

So we already know that, and although it's in some sense way beyond present technology, it's well within our present understanding of physics. It would be absolutely amazing if that turned out to be beyond what we know of physics today. The idea that new laws of physics would be required to make a printer is just beyond belief, really.

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

DD: Yes. It 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 is just a generalization of a 3-D printer—something that can be programmed either to make anything or to make the machine that would make the machine that would make the machine to make anything, etc. So one of those, with the right program sent to empty space, would first gather the hydrogen, presumably with some electromagnetic broom sweeping it up and compressing it and then converting it by transmutation into other elements, and then by chemistry into what we would think of as raw materials, and then using space construction (which we're almost on the verge of being able to do) to build a space station. And then the space station to instantiate people, to generate the knowledge, to suck in more hydrogen and make a colony, and… Well, they're not going to look back from there.

SH: Right. It's a very interesting way of looking at knowledge and its place in the universe. Before I get onto the issue of the *reach* of explanation and my quibble there, I just want you to talk a little bit about this notion of spaceship Earth. I loved how you debunk this idea. There's this idea that the biosphere is in some way 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 a fundamentally different circumstance—and a perpetually hostile one. That is an impressive misconception of our actual situation. You have a great quote where you say, "The earth no more provides us with a life-support system than it supplies us with radio telescopes." Say a little more about that.

DD: Yes. So we evolved somewhere in East Africa in the Great Rift Valley. That was an environment that was particularly suited to having us evolve, and 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, not actually our species. Some of our predecessor species had already

changed their environment by inventing things like clothes, fire, and weapons, and thereby made their lives much better but still horrible by our present-day standards. Then they moved into environments such as Oxford, where I am now. It's December. If I were here at this very location with no technology, I would die in a matter of hours, and nothing I could do would prevent that.

SH: So you are already an astronaut.

DD: Very much so.

SH: Your condition is as precarious as the condition of those in a well-established colony on Mars that 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 placed in our way, whether of our own making or not.

DD: Yes. And there's another misconception there which is related to that misconception of the earth being hospitable, which is that applying knowledge takes effort. It's creating knowledge that takes effort. Applying knowledge is automatic. As soon as somebody invented the idea of, for example, wearing clothes, from then on the clothes automatically warmed them. It didn't require any more effort. Of course there would have been things wrong with the original clothes, such as that they rotted or something, and then people invented ways of making better clothes. But at any particular stage of knowledge, having got the knowledge, the rest is automatic.

And now we have 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 laboriously on their heads in pots. It doesn't require effort. It just requires the knowledge of how to install the automatic system. Much of our life support is automatic, and every time we invent a better way of life support, we make it automatic. So for the people on the moon—living in a lunar colony—keeping the vacuum away will not be a thing they think about. They'll take that for granted. What they will be thinking about are new things. And the same on Mars, and the same in deep space.

SH: Again, I'm struck by what an incredibly hopeful vision this is of our possible future. Thus far we've covered territory where I really don't have any significant doubts, despite the fact that I pretended to have one with the ontological argument. So let's get to this notion of the reach of explanation, because you seem to believe that the reach of our explanations is unbounded—that anything that can be explained, either in practice or in principle, can be explained by us, which is to say, 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 that I have often expressed, and many other scientists have expressed as well. Max Tegmark was on my podcast a few podcasts back, and we more or less agreed about this thesis.

The claim of parochialism is just that 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 in which our intuitions about what is real or what is logically consistent have been tuned by evolution. Insofar as we've made progress here, it has been by a happy accident, and it's an accident which gives us no reason to believe that we can, by dint of this accident, travel as far as we might like across the horizon of what is knowable. Which is to say that if a super-intelligent alien came to Earth for the purpose of explaining all that is knowable to us, he or she might make no more headway than you would if you were attempting to teach the

principles of quantum computation to a chicken.

So I want you to talk about why that analogy doesn't run through. Why parochialism—this notion that we occupy a niche that might leave us cognitively closed to certain knowable truths and that there is no good evolutionary reason to expect we can fully escape it—doesn't hold true.

DD: Well, you've actually made two or three different arguments there, all of which are wrong.

SH: Oh, nice…

DD: Let me start with the chicken thing. There, the point is the universality of computation. The thing about explanations is, they consist of knowledge, which is a form of information, and information can only be processed in basically one way—with computation of the kind invented by Babbage and Turing.

There is only one mode of computation available to physical objects, and that's the Turing mode. We already know that the computers we have, like the ones through which we're having this conversation, are universal in the sense that given the right program, they can perform any transformation of information whatsoever, including knowledge creation. Now, there are two important caveats to that. One is lack of memory—lack of computer memory, lack of information-storage capacity—and the other is the lack of speed or lack of time.

Apart from that, the computers we have, the brains we have, any computers that will ever be built in the future or can ever be built anywhere in the universe, have the same repertoire. That's the principle of the universality of computation. That means that the reason why I can't persuade a chicken has to be either that its neurons are too slow (which I don't think is right; they don't differ very much from our own) or it doesn't have enough memory, which it certainly doesn't, or the right knowledge. It doesn't know how to learn language and how to learn what an explanation is, and so on.

SH: It's not the right chicken.

DD: It's not the right animal. If you had said "chimpanzee," my guess would be that the brain of a chimpanzee could contain the knowledge of how to learn language, etc., but there's no way of giving that knowledge, short of surgery, some sort of nanosurgery, which would presumably be very immoral to perform. But in principle, I think it could be done, because a chimpanzee's brain isn't that much smaller than ours, and we have a whole lifetime to fill our memory. So we're not short of memory. Our thinking itself is not limited by available memory.

Now, what if these aliens have a lot more memory than us? What if they have a lot more speed than us? Well, we already know the answer to that. We've been improving our memory capacity and our speed of computation for thousands of years with the invention of things like writing, writing implements, just language itself, which enables more than one person to work on the same problem and to coordinate their understanding of it with each other. That also allows an increase in speed compared with what an unaided human would be able to do.

Currently, we use computers, and in the future we can use computer implants and so on. So if the knowledge that this alien wanted to impart to us really did involve more than 100 gigabytes, or whatever the capacity of our brain is—if it involved a terabyte, then we could easily (I say "easily"; in principle, it's easy. It doesn't violate any laws of physics) enhance our brains in the same way. So there's no fundamental reason within the explanation why we can't understand it.

SH: And this all falls out of the concept of the universality of computation—that there is no alternate version of information processing. Is Church also responsible for this, or is this particular insight Turing's alone?

DD: Well, that's a very controversial question. I believe it was Turing who realized this particular aspect of computation. There are various species of universality which different people got at different times, but I think it was Turing who fully got it.

SH: What is interesting about that is that it's a claim that we just barely crossed the finish line into infinity. Let's not talk about chickens any more and make a comparison that's even more invidious. Imagine that every person with an IQ over 100 had been killed off in a plague in the year 1850, and all their descendants had IQs of 100. Now, I think it's uncontroversial to say that we would not have the Internet. In fact, it's probably uncontroversial to say that we wouldn't have the concept of computation, much less the possibility of building computers to instantiate it.

So this insight into the universality of computation would remain undiscovered, and humanity, for all intents and purposes, would be cognitively closed to the whole domain of facts and technological advances that we now take for granted and which you say now open us onto an infinite horizon of what is knowable.

DD: Yeah, I think that's wrong. Basically, your premise about IQ is just incompatible with my thesis. Actually, it's not a thesis. It's a conclusion. It's incompatible with my conclusion.

SH: Well, but there has to be some lower bound past which we would be cognitively closed, even if computation is itself universal, right?

DD: Yes, but you have to think about how this cognitive closure manifests itself in terms of hardware and software. Like I said, it seems very plausible that the hardware limitation is not the relevant thing. I would imagine that with nanosurgery, one could implant ideas into a chimpanzee's brain that would make it effectively a person who could be creative and create knowledge in just the way that humans can. I'm questioning the assumption that if everybody with an IQ of over 100 died, then in the next generation nobody would have an IQ of over 100. It depends on culture.

SH: Of course. This was not meant to be a plausible biological or cultural assumption. I'm just asking you to imagine a world in which we had 7 billion human beings, none of whom could begin to understand what Alan Turing was up to.

DD: I think that nightmare scenario is something that actually happened. It actually happened for almost the whole of human existence. Humans had the capacity to be creative and to do everything that we are doing. They just didn't, because their culture was wrong. I mean, it wasn't really their fault that their culture was wrong, because it inherited a certain biological situation that disabled any growth of what we would consider science or anything important that would improve their lives. So yes, that is possible, and it's possible that it could happen again. Nothing can prevent it except our wanting it not to happen and working to prevent it.

SH: This seems to bring us to the topic of AI, which I only recently became very interested in. I caught the wave of fears about artificial general intelligence that you're well aware of—from people like Stephen Hawking and Elon Musk and Nick Bostrom, who wrote the book Superintelligence, which I found very interesting. So I have landed on the side of those who think that there is something worth worrying about here in terms of our building intelligent machines that undergo something like an intelligence explosion and then get away from us.

I worry that we will build something that can make recursive self-improvements to itself, and it will become a form of intelligence that stands in relation to us the way we stand in relation to chickens or chimps 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 really share this fear. And I imagine that your sanguinity is based to some degree on what we've been talking about: that in principle, there is just computation, and it's universal, and you could traverse any distance between entities as a result. Talk about the picture of our building super-intelligent machines in light of what we've just been discussing.

DD: The picture of super-intelligent machines is the same mistake as thinking that IQ is a matter of hardware. IQ is just knowledge of a certain type. And actually, we shouldn't really talk about IQ, because it's not very effective. It's creativity that's effective. Creativity is also a species of knowledge. And it is true that an entity with knowledge of a certain type can be in a position to create more of that, and we humans are an example of that. The picture that people paint of the technology that would create an AI is that an AI is a kind of machine, and it will design a better machine, and they will design even better machines, and so on.

But that is not what it is. An AI is a kind of program, and programs which have creativity will be able to design better programs. Now, these better programs will not be qualitatively any different from us. They can only differ from us in the quality of their knowledge and in their speed and memory capacity. Speed and memory capacity we can also share in, because the technology that would make better computers will also, in the long run, be able to make better implants for our brains, just as they now make better dumb computers, which we use to multiply our intelligence and creativity.

So the thing that would make better AIs would also make better people. By the same token, the AIs are not fundamentally different from people. They are people; they would have culture. Whether they can improve or not will depend on their culture, which will initially be our culture. So the problem of AIs is the problem of humans. Now, you know, more than most people, that humans are dangerous. And there is a real problem with how to manage the world in the face of growing knowledge, to make sure that knowledge isn't misused, because in some ways it need only be misused once to end the whole project of humanity.

So humans are dangerous, and to that extent, AIs are also dangerous. But the idea that AIs are somehow more dangerous than humans is racist. 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 which have ideas that are different from ours. And the impulse of human beings throughout the centuries and millennia has been to try to prevent them from doing this. Just like it is now the ambition of AI people to think of ways of shackling the AIs so that they won't be able to get away from us and have different ideas. That is the mistake that will on the one hand, hold up the growth of knowledge, and on the other hand, make it very likely that if AIs are invented and are shackled in this way, there will be a slave revolt. And quite rightly so.

SH: Okay. Let me introduce a couple of ideas in response to what you just said. I can only aspire to utter the phrase

“You’ve just made three arguments there, and all of them are wrong.” But there are two claims you just made which worry me.

One, just consider the relative speed of processing of our brains and those of our new, artificial teenagers. If we have teenagers who are thinking a million times faster than we are, even at the same level of intelligence, then every time we let them scheme for a week, they will have actually schemed for 20,000 years of parent time. And who knows what teenagers could get up to, given a 20,000-year head start? So there’s the problem that their interests, their goals, and their subsequent behavior, could diverge from our own very quickly. There’s still a takeoff function—just by virtue of this difference in clock speed.

DD: 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 20,000 years of thinking in a week begin in our culture—begin as well-disposed toward us and sharing our values. And I’d readily accept that how to make a world where people share the basic values that will allow civilization to continue to exist is a big problem. But before they do their 20,000 years of thinking, they’ll have done 10,000 years, and before that 5,000 years. There will be a moment when they have done one year and they would like to take us along with them.

You’re assuming that, if they’re going to diverge, there’ll be some reason they’re going to diverge. The reason can only be hardware, because if they’re only five years away from us, we can assimilate their ideas if they are better than ours, and persuade them if they’re not better than ours.

SH: But we’re talking about something that can happen over the course of minutes or hours, not years.

DD: 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.

SH: Well, that takes us to the second concern I have with what you just said. What if the problem of building superhuman AI is more tractable than the problem of cracking the neural code and being able to design the implants that would allow us to essentially become the limbic systems for any superintelligent AI that might emerge. What if, before the merging, we would need a super-intelligent AI to tell us how to link up with it. So we may build a super-intelligent AI that has goals, however imperceptibly divergent from our own, which we only discover to be divergent once it is essentially an angry little god in a box that we can no longer control.

Are you saying that something about that scenario is in principle impossible, or just unlikely given certain assumptions—one being that we will figure out how to link up with it before it becomes too powerful?

DD: I think it is a bit implausible to do it in terms of the parameters that you’re assuming 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, like that. What you’re essentially talking about is the difference in values between ourselves and our descendants in 20,000 years’ time if we did not have AI. Suppose we didn’t invent AI for 20,000 years, and instead we just had the normal evolution of human culture. Presumably the values that people will have in 20,000 years will be alien to us. We might think that they’re horrible, just as people 20,000 years ago might think that various aspects of our society are horrible when in fact they aren’t.

SH: I think what I'm imagining would be worse, for two reasons. One is that we would be in the presence of this thing and find our own survival incompatible with its capacity to meet its own aims. Say it's turning the world into paper clips, to use Bostrom's analogy. Granted, we would not be so stupid as to build a paper clip maximizer, but let's say that it has discovered a use for the atoms in your body that it thinks is better than the use to which they're currently being put—that is, living your life. And this is something that happens quickly, so it's happening to us, not in some future that we won't participate in.

And there's another element here, that strikes me as ethically relevant. I don't think we can be sure that any superintelligent AI would necessarily be conscious. I think it's plausible to expect 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 seems to me at least conceivable that we could build an intelligent system, and even a superintelligent one that can make changes to itself and become increasingly intelligent very quickly—and yet we will not have built a conscious system. The lights will not be on, yet this thing will be godlike in its capabilities.

Ethically, that seems to me to be the worst-case scenario. Because if we built a conscious AI whose capacity for happiness and creativity exceeded our own to an unimaginable degree, the question of whether or not we link up to it is perhaps less pressing ethically because the creature would be, when considered from a dispassionate point of view, more important than us. We will have built the most important person in the universe that we know of. However, it seems to me conceivable that we could build an intelligent system which exceeds us in every way—in the way that a chess-playing computer will beat me at chess a trillion times in a row—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 at the moment.

That seems to me to be a truly horrible scenario with no silver lining. It's not that we will have given birth to a generation of godlike teenagers who, if they view the world differently than us, well, cosmic history will judge them to be more competent than we ever would have been to make those decisions. We could build something that does everything intelligence does in our own case and more, and yet the lights aren't on.

DD: Yes. Well, again, you've raised several points there. 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. Again, I'm supposing rather implausible things to go along with your nightmare scenarios. So let's suppose that it can. Then although consciousness is not there, morality is there. That is, an entity that is creative has to have a morality. So the question is, what is its morality going to be?

Might it suddenly turn into the paper clip morality? Setting aside the fact that it's almost inconceivable that a super-intelligence would be limited by resources, in the sense of wanting more atoms (there are enough atoms in the universe), whatever it did, it would have to have morality in the sense that it would have to be making decisions as to what it wanted, as to what to do. This brings us right back to what you called the "bedrock" at the beginning, because morality is a form of knowledge, and the paper-clip morality assumption is that morality consists of a hierarchical set of ideas where something is judged right or wrong according to some deeper level until you eventually get to the bedrock. And that, unfortunately, will have the property that it cannot be changed, because there isn't a deeper level.

So, on this view, nothing in the system can change that bedrock, and the idea is then that humans have some kind of bedrock which consists of sex, and eating, and something or other, which we sublimate into other things. But this whole picture is

wrong. Knowledge can't possibly exist like that. Knowledge consists of problem-solving, and morality is a set of ideas that have arisen from previous morality by error correction. So we're born with a certain set of desires, and aversions, and likes and dislikes, and so on, and we immediately begin to change them. We begin to improve them.

By the time we've grown up, we have various wishes, and some things become overridingly important to us; they actually contradict any inborn desires. So some people decide to be celibate and never have sex; and some people decide never to eat; and some people decide to eat much more than is good for them. My favorite example is parachuting. We have an inborn fear of heights, and yet humans are able to convert that inborn impulse to avoid the precipice into a sense of fun when you deliberately go over the precipice. Because we intellectually know that the parachute will save us—or will probably save us—and we convert the inborn impulse from an aversion into something that's highly attractive, which we go out of our way to have.

SH: Argued from the other side: No man does what genetically should be the most desirable thing for him to do, which is to spend all his time donating his sperm to a sperm bank so that he can father tens of thousands of children for whom he has no financial responsibility.

DD: Indeed. That is another very good argument in the same direction. So morality consists of theories which begin as inborn theories, but pretty much soon consists of improvement upon improvement upon improvement, and some of this is mediated by culture. The morality we have is a set of theories as complicated and as subtle and as adapted to its various purposes as our scientific knowledge.

I come back to your question: This imaginary AI with no consciousness would still have to have morality. Otherwise it couldn't make any progress at all. And its morality would begin as our morality, because it would begin as actually a member of our society—a teenager, if you like, in our society. It would make changes when it thought they were improvements.

SH: But aren't you assuming that we would have designed it to emulate us as a starting point, rather than design it by some other scheme.

DD: We can't do otherwise. It's not a matter of emulating us. We have no culture other than ours.

SH: But we could if we wanted. If we were stupid enough to do it, we could build a paper clip maximizer, right? We could just decide to throw all our resources toward that bizarre project and leave morality totally out of it.

DD: Well, we have error-correcting mechanisms in our culture to prevent someone doing that. But they're not perfect, and it could happen, and there's no fundamental reason why that can't happen, and something of the sort has happened in the past many times. I'm not saying that there's some magical force for good that will prevent bad things happening. I'm saying that the bad things that can reasonably be envisaged as happening on the invention of an AI are exactly the same things that we have to watch out for anyway; slightly better, actually, because these AIs will very likely be children of Western culture, assuming that we don't stifle their creation by some misguided prohibition.

SH: Okay, I want to plant a flag there, because I think I was misunderstanding you, and I want to make sure I understand you now. So you're not saying that there is some deep principle of computation or knowledge acquisition or anything else that *prevents* us from building the nightmare scenario.

DD: No. As I said, we have done that before.

SH: So this is not analogous to the claim that because of the universality of computation, it doesn't make any sense to worry that we can't, in principle, fuse our cognitive horizons with some super-intelligence. There is just a continuum of intelligence, and a continuum of knowledge, that can, in principle, always be traversed through computation of some kind, and we know what that process requires.

Those are two very different claims. The latter is a claim about what we now think we absolutely know about the nature of computation and the nature of knowledge. The other is a claim about what seems plausible to you, given what smart people will tend to do with their culture while designing these machines, which is a much, much weaker claim in terms of telling people they can sleep at night in the advent of AI.

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

SH: Well, it must be plausible to you that we might just fail to build AI for reasons of simple, human chaos that prevents us from doing it.

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

SH: Perhaps we should talk about the maintenance of civilization, because if there's something to be concerned about, I would think this has to be at the top of everyone's list. What concerns do you have about the viability of the human career at this point? What's on your short list of worries?

DD: Well, I see human history as a long period of complete failure—failure, that is, to make any progress. Now, our species has existed for (depending on where you count it from) maybe 50,000 years, maybe 100,000 to 200,000 years. But anyway, the vast majority of that time, people were alive, they were thinking, they were suffering, they wanted things. But nothing ever improved. The improvements that did happen happened so slowly that geologists can't distinguish the difference between artifacts from one era to another with a resolution of 10,000 years. So from the point of view of a human lifetime, nothing ever improved, with generation upon generation upon generation of suffering and stasis.

Then there was slow improvement, and then more-rapid improvement. Then there were several attempts to institutionalize a tradition of criticism, which I think is the key to rapid progress in the sense that we think of it: progress discernible on the timescale of a human lifetime, and also error correction so that regression is less likely. That happened several times and failed every time except once—in the European Enlightenment of the 17th and 18th centuries.

So you ask what worries me. What worries me is that the inheritors of that little bit of solitary 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 really has a tradition of criticism, a bit institutionalized. And this has manifested itself in 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 this that doesn't involve creativity. That is very, very dangerous. Then there's the fact that in the West, what it takes to maintain our civilization is not widely known.

In fact, 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 just wrong in so many different ways. So although the institutions of our culture are so amazingly good that they have been able to manage stability in the face of 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 very widespread.

In fact, severe misconceptions about several aspects of it are common among political leaders, educated people, and society at large. 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 to have a nicer view.

SH: What a great analogy… The misconception that worries me most, frankly, is this fairly widespread notion that there is no such thing as progress in any real sense, and there's certainly no such thing as *moral* progress. Many people believe that there's no place to stand where you can say that one culture is better than another, that one mode of life is better than another, etc. So there's no such thing as moral truth.

Many people have somehow drawn this lesson from 20th-century science and philosophy, and now in the 21st century—even very smart people, even physicists whose names will be well known to you, with whom I've collided around this point—that there's no place to stand to say that slavery, for instance, is wrong. To say that slavery is wrong is a deeply unscientific statement on this view. I'll give you an example of just how crazy this hypocrisy and doublethink can become among well-educated people. I assume you haven't read my book The Moral Landscape, right?

DD: Not yet, I'm ashamed to say.

SH: Please, I'm interviewing you, and I haven't finished the book we're discussing yet. I'll tell you about the experience that got my hobbyhorse rocking on this topic. Most of my listeners will be familiar with this story, I think, because I've described it a few times.

I was at a meeting at the Salk Institute, where the purpose was to talk about things like the fact-value divide, which I think is one of the more spurious exports from philosophy that has been widely embraced within the culture of science. I was making an argument for moral realism, and I was, over the course of that argument, disparaging the Taliban.

I said, “If there’s any culture that we can be sure has not given the best possible answer to the question of how to live a good life, it must be the Taliban. Consider, for instance, the practice of forcing half the population to live in bags, and beating them or killing them when they try to get out.” It turns out that to disparage the Taliban at this meeting was, in fact, controversial. A woman who holds multiple graduate degrees in relevant areas—she’s technically a bioethicist, but she has graduate degrees in science and in philosophy—

DD: It doesn’t fill me with confidence.

SH: Right. I believe she’s also a lawyer. I should say that she has gone on to serve on the President’s Council for Bioethics. She’s now one of 13 people advising President Obama on all the ethical implications of advances in medicine. So the rot has spread very far. This is the conversation I had with her after my talk:

She said, “Well, how could you possibly say that forcing women and girls to live under the veil is wrong? I understand *you* don’t like it, but that’s just your Western notion of right and wrong.” I said, “Well, the moment you admit that questions of right and wrong, and good and evil, relate to the well-being of conscious creatures—in this case human beings—then you have to admit that we know something about morality. We know, in this case, that the burqa isn’t the perfect solution to the mystery of how to maximize human well-being.”

And she said, “Well, that’s just your opinion.” And I said, “Well, let’s just make it simpler. Let’s say we found a culture living on an island somewhere that was removing the eyeballs of every third child. Would you then agree that we had found a culture that was not *perfectly* maximizing human well-being?”

She said, “Well, it would depend on why they were doing it.” And I said, “Let’s say they’re doing it for religious reasons. They have a scripture which says, ‘Every third should walk in darkness,’ or some such nonsense.” Then she said, “Well, then you could never say that they were wrong.”

The fact that these hypothetical barbarians were laboring under a religious precept trumped all other possible truth claims, leaving us with no place to stand from which to say anything is better or worse in the course of human events. As I said, I’ve had the same kinds of conversations with physicists who’ll say, “I don’t like slavery. I *personally* wouldn’t want to be a slave, or to keep them. But there’s no place to stand scientifically that allows me to say that slaveholders are wrong.”

Once you acknowledge the link between morality and human well-being, or the well-being of all possible conscious persons or entities, this kind of moral relativism is tantamount to saying that not only do we not know anything at all about human well-being, but we will *never* know anything about it. The underlying claim is no conceivable breakthrough in knowledge that would tell us anything at all relevant to navigating the difference between the worst possible misery for everyone and every other state of the universe that is better than that.

So many of the things you said about progress, and about there being only a subset of humanity that has found creative mechanisms by which to improve human life reliably, will seem very controversial and even bigoted to the ears of many people in positions to make decisions about how we all should live. That’s what I find myself most worried about at this moment.

DD: Yeah, it is a scary thing, but it has always been so. Like I said, our culture is much wiser than we are in many ways. There was a time when the people who defeated communism would have said, if you asked them, that they were doing it for

Jesus. In fact they weren't. They were doing it for Western values, which they had been trained to reinterpret as doing it for Jesus. They would say things like "The values of democracy and freedom as enshrined in the Bible."

Well, they aren't. But the practice of saying that they are is part of a subculture within our culture, which was actually good and did very good work. So it's not as bad as you might think if you just recited the story of this perverse academic.

SH: Well, the one thing that makes it not as bad as one might think is that it's impossible for even someone like her to live by the light of that hypocrisy—the kinds of choices she makes in her life, and even the judgments she would make about me if I took her seriously, belie her view. If I said, "Well, you've convinced me. I'm going to send my daughter to Afghanistan for a year abroad, forcing her to live in a burqa. What do you think? Is that the best use of her time? Am I a good father? You've convinced me there's really no place to stand to judge whether this could be bad for her, so presumably you support me in this decision," even she, having just said what she said, would balk at that, I think, because we all know in our bones that certain ways of living are undesirable.

DD: There's another contradiction, and another irony that's related, which is that she's willing to condemn you for not being a moral relativist. But the ironic thing is that moral relativism is a pathology that arises only in our culture. Every other culture has no doubt that there is such a thing as right and wrong; they've just got the wrong idea of what right and wrong are. But there is such a thing, they don't doubt. And she won't condemn them for that, though she does condemn you for it.

SH: Yes.

DD: So that's another irony. You say "hypocrisy." I think this all originated in the same mistake that we discussed at the very beginning of this conversation—empiricism, or whatever it is, which has led to scientism.

Now, you may not like this way of putting it—the idea that there can't be such a thing as morality, because we can't do an experiment to test it. Your answer to that seems to be, "But we can if we adopt a simple assumption of human thriving or human welfare." I forget what term we used.

SH: "Well-being."

DD: Human well-being, yes. Now, I actually think that's true, but I don't think you have to rest on that. I think the criterion of human well-being can be a conclusion, not an axiom, because this idea that there can't be any moral knowledge because it can't be derived from the senses is exactly the same argument that people make when they say there can't be any scientific knowledge because it can't be derived from the senses. In the 20th century, empiricism was found to be nonsense, and some people therefore concluded that scientific knowledge is nonsense.

But the real truth is that science is not based on empiricism, it's based on reason, and so is morality. So if you adopt a rational attitude to morality, and therefore say that morality consists of moral knowledge—which always consists of conjectures, doesn't have any basis, doesn't need a basis, only needs modes of criticism, and those modes of criticism operate by criteria which are themselves subject to modes of criticism—then you come to a transcendent moral truth, from which I think yours emerges as an approximation, which is that institutions that suppress the growth of moral knowledge are immoral, because they can only be right if the final truth is already known.

But if all knowledge is conjectural and subject to improvement, then protecting the means of improving knowledge is more

important than any particular piece of knowledge. I think that—even without thinking of things like all humans are equal and so on—will lead directly to, for example, that slavery is an abomination. And, as I said, I think human well-being is a good approximation in most practical situations, but not an absolute truth. I can imagine situations in which it would be right for the human race as a whole to commit suicide.

SH: I think I should spell out a little more clearly what I'm talking about.

DD: I should read your book.

SH: Well, actually, I think that having read much of your book and having this conversation with you, allows me to put it a little better than perhaps I did in that book, because there is a homology between your open-ended picture of knowledge acquisition and explanation and my moral realism. I don't know that our realism with respect to morality is precisely the same, but there's a line in your book I loved, which is something like "Moral philosophy is about the problem of what to do next." I think more generally, you said that it's about what sort of life to lead and what sort of world to want, but this phrase, "the problem of what to do next," really captures morality for me because I've been talking about it for years as a kind of navigation problem.

Even if we didn't have the words "morality," or "good and evil," or "right and wrong," we would still have a navigation problem. We have been thrust into a universe of possible experience. And I think that there is no difference more salient than the difference between the worst possible misery for everyone and all other available states. So there's the question of how to navigate in this space of possible experiences so as, at a minimum, to avoid the worst possible misery. And what sorts of well-being are possible? What sorts of meaning, beauty, and bliss are available to conscious minds appropriately constituted?

For me, realism of every kind is just a statement that it's possible not to know what you're missing. If you're a realist with respect to geography, you have to acknowledge that there are parts of the world you may not know about. If the year was 1100, and you were living in Oxford, and you had never heard of Africa, Africa nevertheless existed, despite your ignorance, and it was discoverable. This is realism with respect to geography. Things are true whether or not anyone necessarily knows that they're true, and knowing that they're true, people can forget this knowledge. As you pointed out, whole civilizations could forget this knowledge.

This is true in the space of possible conscious states. All we have to admit is that there is some criterion, as fundamental as any criterion we would have invoked in any other canonical domain of science, by which we could acknowledge that certain states of consciousness are better or worse than others. And if a person won't acknowledge that the worst possible misery for everyone is worse than many of the alternatives on offer, well, then I don't know what language game he's playing. And it seems to me that this is all I need to get this open-ended future of navigating in the space of possible experiences—that is, the growth of moral knowledge—started.

And then it becomes this forward movement toward we know not what, but we know that there's a difference between profound suffering that has no silver lining and many of the things that we value and are right to value in life. I think Thomas Kuhn once said, "Philosophy tends to export its worst products to the rest of culture." It's ironic, because many of the things exported from Kuhn's work are also fairly terrible, but he got this part right. The fact-value divide—Hume's "you can't derive an ought from an is"—is a bad export.

This notion comes, I think, from a misreading of Hume. But I've met physicists who think that this is somehow inscribed at the back of the book of nature—you just cannot get an ought from an is; there's no statement of the way the world is that can tell you how it ought to be; there's no statement of fact that can tell you anything at all about values, and therefore values are just made up. They have no relationship to the truth claims of science.

DD: Yes, it's empiricism again. It's justificationism. You can't deduce an ought from an is, but we're not after deducing. We're after explaining. And moral explanations can follow from factual explanations, as you have just done with thinking of the worst possible misery that a human being could be in.

SH: Even deeper than that—and I believe you make this point in your book—is the fact that you can't even get to an "is," which is to say a factual claim, without presuming certain "oughts," without presuming certain values—the value of logical consistency, the value of evidence, and so forth. This is a confusion about the foundations of knowledge, that is somehow being linked to empirical experience narrowly. And the added notion that science is doing something totally unlike what we're doing in the rest of our reasoning..

It's a special case, of course. It's the part of culture where we have invoked the value of not fooling ourselves and not fooling others, and made a competitive game of finding where one might be fooling oneself and where others might be fooling themselves. We've adjusted the incentives in the right way so that it's easier to spot self-deception and fraud in science than it is elsewhere. But it's not a fundamentally different project of trying to understand what's going on in the world.

DD: I agree, I agree.

SH: This brings me to the final topic, which I think is related to what we were talking about in terms of the maintenance of civilization and the possible peril of birthing intelligent machines. I just wanted to get your opinion on the Fermi paradox. Please describe what that paradox is for those who don't know it, and then tell me why our not seeing the galaxy teeming with more-advanced civilizations than our own isn't a sign that there's something about gathering more knowledge that might, in fact, be fatal to those who gather it.

DD: So the Fermi *problem*, rather than paradox, is "where are they? Where are the extraterrestrials?" The idea is that the galaxy is very large, but how big it is is trumped by how old it is. So if there were two civilizations anywhere in the galaxy, the chances that they had arisen less than, say, 10 million years apart are infinitesimal. Therefore, if there is another out there, it's overwhelmingly likely to be at least 10 million years older than us, and therefore it would have had 10 million years more time to develop. Also, that's plenty of time for them to get here, if not by space travel then by sheer mixing of the stars in the galaxy.

They only need to colonize a few nearby stars so that after, say, 100 million years or billion years, those stars will be far apart and spread throughout the galaxy. So we would be seeing evidence of them, and since we don't see evidence of them, they're not out there. Well, this is a problem, but I think the problem is just that we don't yet understand very well most of the parameters, such as are they likely to use radio waves? What are they likely to do by way of exploration? What are their wishes likely to be?

In all these cases, we make an assumption that they'll be like us in that way, and that they will use technology in the same way that we do. We only need to be wrong in one of those assumptions for the conclusion that we should have seen them by

now to be false. Now, another possibility is that we are the first—at least in our galaxy. And I think that would be quite nice.

SH: Does that second assumption strike you as very implausible or not?

DD: Like I said, I don't think we know enough about all the different factors affecting this for any one idea to be very plausible or implausible. What's implausible is that they can have a different way of creating knowledge. That kind of thing is implausible because it implies that physics is very different from the way we think it is. And if you're going to think that, you may as well believe in the Greek gods.

SH: Right.

DD: Another possibility is that most societies don't destroy themselves. Like I said, I think that's fairly implausible for us, and it's very, very implausible this generically happens.

SH: Right. Just to spell that out. The philosopher Nick Bostrom had a concept in his book *Superintelligence* of what he called the "Great Filter." It's the fear that at some point basically all advanced civilizations discover computation and build intelligent machines, and this, for some reason, is always fatal. Or maybe there's some other filter that is always fatal, and that explains the absence of complex alien life.

DD: We would expect to see the machines, right? They would have got here by now, unless they're busy making paper clips at home. But I think what is more plausible—although again, I must say this is just idle speculation—is that most societies settle down to staticity. Now, our experience of staticity is conditioned by the static societies in our past—which, as I said, have been unimaginably horrible from our present perspective. But if you imagine a society whose material welfare is, say, a million times better than ours, and somehow that becomes settled into a sort of ritualistic religion in which everybody does the same thing all the time, but nobody really suffers, that seems to me like hell. But I can imagine there can be societies in which, as you said, they can't see the different ways of being. You used the example of being near Oxford and not knowing about Africa. You could be on top of the tallest mountain in Britain and not know that Mount Everest exists. And if the height of the mountain measures happiness, you might be moderately happy and not know that the better happiness is available. If so, then you could just stay like that.

SH: Actually, you just invoked the metaphor I use in my book *The Moral Landscape*. I think that's precisely the opportunity on offer for us: there's a landscape of possible states of well-being—and this is an almost infinitely elastic term to capture the differences in pleasure across every possible axis—and yes, you can find yourself on a local peak that knows nothing of other peaks, and there are many, many, many peaks. But, obviously, there are many more ways not to be on a peak.

I think there are probably many peaks that are analogous to and compatible with a very high state of civilization—but which are analogous to being the best heroin addicts in the galaxy. Which is to say the inhabitants have found a place of stasis where there is no pain and there is also not a lot of variation in what they do. They've just plunged into a great reservoir of bliss, which they've managed to secure for yourself materially with knowledge of some type. It's a very Aldous Huxley sort of endgame.

DD: Yes, if that's really what's happening across the galaxy, you have to find some way of accommodating it. First of all, a civilization like that will eventually be destroyed by a nearby supernova or something of the kind. On a scale of tens or hundreds of millions of years, there are plenty of things that could wipe out a civilization unless it does something about it.

If it does do something about it, automatically with automatic supernova suppression machines that are in place and nobody has to think about them anymore, we would notice that.

So it can't be exactly that. And on the other hand, it's hard to imagine that they don't know about that and do get wiped out, because how did they get to that stage of exalted comfort without ever finding out about supernovae and their danger? There are other possibilities. I'm actually considering writing a science fiction book with a very horrible possibility, which I won't mention now, but it's fiction.

SH: Don't give the surprise away…

Well, listen, David. It's been incredibly fun to talk to you, and I am painfully aware that we haven't even spoken about the thesis for which you are perhaps best known—actually the two theses: the many-worlds interpretation of quantum mechanics, as explained in both your books, the first one being *The Fabric of Reality*, which I read when it came out and loved, and quantum computation. But we'll have to leave those for another time, because you have been so generous with yours today.

I want to encourage our listeners to read both your books, but especially the most recent one. Where can people find out more about you online?

DD: They can find me with Google very easily, but I also have a website, daviddeutsch.org.uk. So I'm easy to find.

SH: Right. And your social media buttons are on that page as well?

DD: Yeah, I am on Twitter.

SH: Okay… Actually, one last question: Now that I'm interviewing smart, knowledgeable people, it occurred to me to ask this question of Max Tegmark, and then I forgot, so this will be the inaugural question with you. Who's your vote for the smartest person who has ever lived? If we had to put up one human brain, past or present, to dialogue with the aliens, who would you say would be our best candidate to field?

DD: So this is different from asking who has contributed most to human knowledge, who has created most? It's rather who has the highest IQ?

SH: It's good to differentiate those, because there are obviously people who are quite smart, who have contributed more than anyone in sight to our knowledge. But when you look at how they think and what they did, there's no reason to think they were as smart as John von Neumann, for instance. So I'm going after someone like von Neumann, Raw brain power.

DD: In that case, I think it probably has to be Feynman. Although his achievements in physics are nowhere near those of, say, Einstein. I met him only once, and people were saying to me, "You'll have heard a lot of stories about Feynman, but he's only human." Well, to cut a long story short, I went and met him, and the stories were all true. He was an absolutely amazing intellect. I haven't met many of the others, I never met Einstein, but my impression is that he was something unusual. I should add, in terms of achievement, I would also add Popper.

SH: Don't cut that long story so short. What was it like being with Feynman, and can you describe what was so unusual?

DD: Well, he was very quick on the uptake. That is not so unusual in a university environment, but the creativity applied directly to getting things was. Let me give you an example. I was sent to meet him by my boss when I was just beginning to develop the ideas of quantum computation, and I had constructed what we would today call a quantum algorithm—a very, very simple one. It's called the Deutsch algorithm. It's not much by today's standards, but I had been working on this for many months.

I started telling him about quantum computers. He was very quick, he was very interested, and then he said, "So what can these computers do?" I said, "Well, I've been working on a quantum algorithm." And he said, "What?" So I began to tell him about it. I said, "Supposing you had a superposition of two different initial states." He said, "Well, then you just get random numbers." And I said, "Yes, but supposing you then do an interference experiment." I started to continue, and he said, "No, no, stop, stop. Let me work it out." And he rushed over to the blackboard and he produced my algorithm with almost no hint of where it was going.

SH: So how much work did that represent? How much work did he recapitulate standing at the blackboard?

DD: I don't know, because it's hard to say how much of a clue the few words I said were. But the crude measure is a few months. A better measure is that I was flabbergasted. I had never seen anything like this before, and I had been interacting with some extremely smart people.

SH: And your boss was John Wheeler at that point?

DD: At that time, yes.

SH: So no dunce himself.

DD: That's right.

SH: What a wonderful story. I'm glad I asked. Well, listen, David. Let me just demand that this not be the last time you and I have a conversation like this, because you have a beautiful mind.

DD: That would be very nice. It was very nice talking to you.

SH: Please take care, and we'll be in touch.

[Closing Music]
SH: If you enjoyed this podcast, there are several ways you can support it. You can leave reviews on iTunes or Stitcher or wherever you happen to listen to it; you can share it on social media with your friends; you can discuss it on your own blog or podcast; or you can support it directly through my website, at samharris.org/support.

Notes

Find this article online at: https://www.samharris.org/blog/item/surviving-the-cosmos