Is Superintelligence Impossible?

43 min. read · View original

IS SUPERINTELLIGENCE IMPOSSIBLE?

John Brockman: It's a very exciting evening for me, and I don't like to show emotion. These two individuals have been at each other for twenty years debating very serious and consequential ideas. They appear to be friendly tonight. I hope they are.

These things matter. Ideas matter. They have very different conceptions of the world. They've been talking about the hard problem in consciousness for twenty years, and tonight we're not going to go there. We're going to be talking about the next shoe to drop, which is this growing world of AI. For me, although I met the original cyberneticists in 1965 and I've been here ever since with all of them, it became boring in the '80s, and I walked away from that world of expert systems. The Japanese had MITI, which was the their Fifth Generation agenda for computation. Everybody was nervous: "they're coming, they're coming." They came, they went. Nothing happened.

I happened to be at the meeting when the Japanese official directing MITI showed up for a lunch meeting at the New York Academy of Sciences when he sat down with Marvin Minsky, John McCarthy, Roger Schank, and Ed Feigenbaum, the leaders of the first two waves of AI. After all the build-up about The Fifth Generation, MITI just seemed to devolve into yet another AI Winter. Blink, and it's twenty years later, and we wake up to new developments in self-improving, unsupervised machine learning, as represented by AlphaGo, the software program developed by Demis Hassabis and his colleagues at DeepMind in the UK. It was extremely interesting. It was extremely interesting. I thought it would be valuable just to find out what's happening, so in April, 2016, I put together a dinner in London with Demis. The idea was to have him talk with David Deutsch, one of the sharpest people that I know, and get a sense of what's going on. In the group at dinner were people who have nothing to do with computing, but who have a lot to say about reality: the novelist Ian McEwan, the musician Brian Eno, the filmmaker Terry Gilliam, the arts curator Hans Ulrich Obrist. "The London Quality Chop House Society Dinner" was a fascinating start. We've had more since, and will continue to do so.

In September, 2016, I followed this up with a conference in Washington, Connecticut, with a number of people who had been thinking about

AI their entire lives, beginning with the Cybernetic world of Norbert Wiener. Included were people like Danny Hillis, who broke the von Neumann bottleneck with his parallel processing computer, Peter Galison, the science historian, Seth Lloyd, the quantum theorist, and Neil Gershenfeld, who, looking at Norbert Wiener's book, noted, "This is all so prescient, but it was written all these years ago. His worries about our culture, about the commercialization of science, it's all coming around again. We should do him a favor and rewrite his book." That's how this book started, and that's why we're here.

Tonight, we're going to talk about themes in the book, the title of which is *Possible Minds: 25 Ways of Looking at AI*. One theme is, is superintelligence possible? I thought we'd start off with five minutes by each of these gentlemen, and we'll get started with Dave.

DAVID CHALMERS: It's such a pleasure to be here. Thank you so much to John and to Janna for putting this event together. It's going to be fun. I guess one of our questions here is, is superintelligence possible or impossible? I'm on the side of possible. I like the possible, which is one reason I like John's theme, "Possible Minds." That's a wonderful theme for thinking about intelligence, both natural and artificial, and consciousness, both natural and artificial.

One of our questions here is, is superintelligence possible or impossible? I'm on the side of possible. I like the possible, which is one reason I like John's theme, "Possible Minds." That's a wonderful theme for thinking about intelligence, both natural and artificial, and consciousness, both natural and artificial. ... The space of possible minds is absolutely vast—all the minds there ever have been, will be, or could be. Starting with the actual minds, I guess there have been a hundred billion or so humans with minds of their own. Some pretty amazing minds have been in there. Confucius, Isaac Newton, Jane Austen, Pablo Picasso, Martin Luther King, on it goes. But still, those hundred billion minds put together are just the tiniest corner of this space of possible minds.

We can also add in all the non-human animal minds that there have been. I looked up on the web today how many organisms have lived, how many animals have lived in the history of the planet. The best estimate seems to be around 10^{29}. Most of them are worms it turns out. Their minds may not be so interesting. At least 10^{20} are very, very interesting minds. It's still the smallest corner of the space of possible minds. And what about the computer? One of the amazing things about the computer is the way that it enables us to explore and expand that space of possible minds. Arguably, for the first time since the history of the planet, the computer has enabled some wholly new kinds of

minds to come into existence, not by the standard methods of biological evolution, but by straightforward, intentional design and programming. So far, the minds have been limited but still interesting.

John mentioned AlphaGo and the successes in the AlphaZero family, which have managed to teach themselves to play Go from scratch in a way wholly out of whack, it seems, from the way in which a human would learn to play the game or would play the game at all. Nonetheless, it turned out to exceed human capacity, at least in that one very limited dimension of game playing. Likewise, deep learning has led us to surprising successes on things like image recognition, speech recognition, and, within limited domain, autonomous vehicle driving (they're not there yet on the autonomous vehicles). At least in speech recognition and image recognition, it's starting to exceed human capacity.

We've had a limited expansion of the space of actual minds to include some minds that we've designed, but so far it's only the smallest of expansions. One thing that we shouldn't do tonight is exaggerate where we've gotten to with AI to date. The advances are amazing but limited. They haven't gotten us yet anywhere near general human intelligence, and it's unlikely they're going to do so anytime soon. It's not happening in the next twenty years, but will it happen this century? Maybe.

People say that with any given technology people tend to overestimate its effects in the short term and underrate it in the long term. That's my attitude towards AI. There's a lot of hype right now. It may not change our lives completely in the next twenty years, but in the next 200 years it's probably going to transform everything. And one of the reasons why is because AI builds into it this self-enhancing and self-perpetuating mechanism of exploring and expanding the space of possible minds. With the early AI programming, you had to design them yourself. Alan Turing wrote a program in which he built some very simple rules of thumb for playing chess. And it played chess, though not very well. Now, the chess-playing systems, like AlphaGo, learn to play chess from scratch and do so amazingly well.

Learning serves as a method for moving ahead in this space of possible minds. Start from a pretty simple mind and the capacity to learn, and it gets somewhere. Evolution is another such method. I expect to see AI exploit the evolutionary methods, where we have some kind of system of artificial evolution among a bunch of different AI programs, and their capacities expand in surprising and unpredictable ways over time, thereby also getting us far beyond that starting point. Learning and evolution in computers is a way to expand that space of mind.

The most powerful method of all in exploring that space is one which is still to come, and that's once you have AI systems doing the designing, once AIs are designing AIs. Say we get to the first AI, which is that human level capacity for the various kinds of general intelligence, and in particular, a human level capacity for designing AI. Within a year or two later—these things always get better—this AI program will be even greater than the human level for designing AIs. It will therefore be able to design an AI that is, one way or another, better than itself. Why? It'll be better than humans at designing AIs. The humans designed it, and it'll design something better. This process of recursive self-improvement or recursive self-enhancement, first put forward by the philosopher and statistician I.J. Good, I see as an amazing bootstrapping method for exploring that space of mind.

We start from our little corner here in this space of possible minds, and then learning and evolution expands it still to a much bigger area. It's still a small corner, but the AIs we can design may help design AIs far greater than those we can design. They'll go to a greater space, and then those will go to a greater space until, eventually, you can see there's probably vast advances in the space of possible minds, possibly leading us to systems that stand to us as we stand to, say, a mouse. That would be genuine superintelligence. That's possible. I also think

it's not going to happen in twenty years or fifty years, but we do have to think about it and we do have to worry about it.

A working definition of intelligence here is the ability to fulfill your goals across a very wide range of goals, to solve problems, and find ways to achieve your goals in extremely powerful ways. These AIs by definition will be systems that are extremely powerful at achieving their goals, if they have goals. Unless there's some counterveiling force, they're probably going to achieve those goals. We have to then be very careful. What are the goals of these AI systems? As philosophers, we think about values. What are the values we put into these AI systems? We also need to think about consciousness. Are these AIs that we're creating conscious? How does their consciousness relate to human consciousness? Is this going to be a world of wonderfully enhanced subjective experience or a mindless world without consciousness at all? That's something maybe we can talk about as this goes along.

The final question we need to ask is where do we as humans stand with respect to these AIs? Are these AIs the systems that replace us or enhance us? Do we ourselves eventually become the AIs? Do we enhance ourselves? Do we upload ourselves to eventually become the AIs which are on the forefront of this expanding wave of superintelligence? That's an attractive

prospect in some ways, compared to the prospect where humans don't exist at all and get wiped out. Could you upload a human mind into a computer? This raises some of the oldest philosophical questions of identity and consciousness, so it's a great thing that John has managed to bring a number of philosophers, scientists, and engineers to think about these questions at this time.

JB: Thank you. I should add that Dave is a university professor of philosophy and neural science and co-director of the Center for Mind, Brain and Consciousness at New York University. Dan Dennett needs no introduction.

CHALMERS: I demand you take back that introduction.

JB: Dan Dennett is university professor and Austin B. Fletcher Professor of Philosophy and director of the Center for Cognitive Studies at Tufts.

DANIEL C. DENNETT: Thank you, John. Thank you, David. This is supposed to be a debate, but almost nothing that David said is anything I disagree with, although I wouldn't put the emphases where he does.

Let's talk about "possible" for the moment. There are lots of things that are possible, and philosophers love to talk about what's possible, but many things that are obviously possible are

never going to be actual. It's possible to build a bridge across the Atlantic. We're not going to do it, not now, not in a hundred years, not in a thousand years. It would cost too much money and would be a foolish endeavor. A lot of the imagined AI projects that are perfectly possible in principle are not worth doing. In fact, some of them are definitely things that we shouldn't do because they'll make more problems for us than they'll solve. Just bear that in mind.

Somebody said that the philosopher is the one who says, "We know it's possible in practice, we're trying to figure out if it's possible in principle." Unfortunately, philosophers sometimes spend too much time worrying about logical possibilities that are importantly negligible in every other regard. So, let me go on the record as saying, yes, I think that conscious AI is possible because, after all, what are we? We're conscious. We're robots made of robots made of robots. We're actual. In principle, you could make us out of other materials. Some of your best friends in the future could be robots. Possible in principle, absolutely no secret ingredients, but we're not going to see it. We're not going to see it for various reasons. One is, if you want a conscious agent, we've got plenty of them around and they're quite wonderful, whereas the ones that we would make would be not so wonderful.

For me, one of the important fears about the future is that long before we got to superintelligence, we would have human beings who are so dependent on non-superintelligences that we would become fragile and brittle in some very important ways. We might call that the GPS problem magnified. People have begun not being able to read maps anymore or know how to get anywhere without the help of GPS. Use it or lose it. Use it or lose it is going to play a big role in everybody's lives in the immediate future.

Is there anybody in this room that knows an algorithm for extracting the square root? I learned one in school when I was in about eighth grade. It's not easy, but there are algorithms for doing a square root, which nobody bothers with anymore. Nobody knows how to do that because you just hit that little button on your hand calculator. Many more important talents are going to atrophy and disappear except in the hands of cranky craftsmen. They'll still know how to make a horseshoe with a hammer, an anvil, and a simple forge. They'll be able to read a map and drive a car and other weird things like that while the rest of us are simply disabled in those regards. That's something that worries me.

Even more, what worries me is that we will for the very best of reasons turn over our responsibility for making major decisions to artificial in-

telligences that are not conscious and not super, they're just very intelligent tools that are great fabrics of pattern recognition and so forth. Who knew twenty years ago there could be such things? We know now that there are—deep learning, et cetera—but when you start delegating major life decisions to systems that are basically just smart tools, then this changes our human predicament in a very important way. My slogan about this is we want smart tools, intelligent tools, not artificial colleagues. The difference is that an artificial colleague is somebody who can take responsibility to be a co-author and be morally responsible for decisions made. We're nowhere near that with artificial intelligence.

Alan Turing, one of my all-time heroes, set in motion one thing which I regret, and that is that the Turing Test puts a premium on deception, on convincing human beings that they're talking to a human being. I know why he did it, it was a brilliant idea, but ever since then there has been this premium on what we might call the Disneyfication of artificial intelligence—making AIs that seem more human. They're basically false advertising. Whether we're talking about Siri, or Watson, or any of the others, they have this paper-thin human user interface which is deeply deceptive about what they understand. That's false advertising. It should not be honored, but rather criticized and condemned.

We should get out of the habit of treating AIs as agents when they're not. The reason this is going to be hard is that, as a number of people are foreseeing, the major market for AI is going to be elder care. And why not? Taking care of elderly folks who can't take care of themselves is not a good life for a regular human being. It's maybe worse than being an old-fashioned telephone operator. We don't regret the loss of those jobs. In elder care, there will be good market reasons for Disneyfying AI to a very great extent because old folks will want to have a companion, not just somebody that brushes their teeth and gets them fed and so forth. I do not like the future that is populated by millions and millions of old folks who are settling for an artificial companion that is a fake in most important regards.

JB: Thank you.

CHALMERS: We do agree on an awful lot here, but maybe we do have a disagreement about this core question of whether there will be genuine autonomous AI. Dan's piece in the book is wonderfully lucid and thoughtful on this. In his vision, while it's possible to create autonomous, intelligent, conscious AI, we shouldn't do it and maybe we won't do it. Instead, what we should do is create tools. He uses the wonderful analogy of Google Maps, where Google Maps tells you how to get someplace, but you still have to get there. If you want to get somewhere, it'll

show you a route, but you still have to follow the route. The human is still in the loop. You've got some advice, and then the human takes it. That's what I'm seeing as Dan's vision of AI. Maybe you've got a superintelligent AI and you want to know how to get to Mars or how to win a war or something, the AI will tell you what needs to be done, but the human will still be in the loop.

It's a beautiful vision, but I worry if it's realistic. There are going to be so many incentives to take the human out of the loop and to give these AIs the capacity to act on that advice directly and autonomously; in fact, this is already happening with Google Maps, with navigation software. With cars like a Tesla, for example, for a long time you'd tell them your destination and it would do the usual Google Maps thing. It would show you all the routes, but you had to still follow it and make the decision. Then, at a certain point they introduced a button called "navigate on autopilot." This means the car takes those instructions and follows the instructions itself. It turns the wheel and changes freeways and so on. There are limitations; for example, it can't drive on ordinary city streets.

In a very small domain, what you see happening is that the car has become, in a very limited way, autonomous. It has those goals and it acts on them. Yes, we can still stop it and change the goals and so on, but when I think about do-

mains like autonomous weapons in the military, well, sure, for a brief period when the stakes are low, maybe we'll just have AI systems that advise a soldier on target detection. Eventually, the AI is going to be so much faster and better at doing this kind of thing that when the stakes of a genuine international military conflict are present, it's hard to see that we're not going to move to genuine autonomous soldiers that have goals and execute them and fire the weapons.

More generally, biological systems are going to eventually be slow and creaky compared to these new super fast AIs. For financial purposes—the stock market—military purposes, and even scientific purposes, the incentives are going to be so strong to allow the AIs to achieve the goal directly. Autonomy is going to be very hard to avoid. If the tech companies are running it, certainly it's going to happen. If the government is running it, if the military's running it, it's going to happen. So, I don't know how you're going to avoid this from happening.

DENNETT: You raise a real and important issue, which is how much autonomy do you want? We don't want autonomous cars *that* autonomous. Dilbert a few weeks ago had a wonderful cartoon in which Dilbert's autonomous car says, "I want you to call me Carl. Self-driving car is my slave name." Dilbert says, "Shut up and drive me to the market." And the car says, "Says the self-walking man."

We don't want that much autonomy. Autonomy is synonymous with free will, and I don't think we want to give AI complete autonomy because the nature of the technology has a certain invulnerability we don't have. You can back them up and put them back together again and make another copy on Monday, and if human beings were capable of being completely backed up and then brought back on Monday, that would change the nature of human interactions and human relations dramatically. I for one don't want to go there, and I don't think many people do.

So, if you're right that it's inevitable that market pressures and cleverness will lead to genuinely autonomous AIs, then we're in for a very bad future indeed. When could that happen? Well, we can give them more autonomy than they can handle, and that's what I'm afraid of.

JB: Caroline Jones, one of the essayists in the book, sent a question today that pertains here in regard to this reliance on a computational way of looking at the world. She asks: "Can you address the complexity of our wet cognition, a much more distributed notion of intelligence that goes beyond the ideas of computation, our separate craniums. Craniums may not even be bounded by our own skin. In this regard, what is robot death? Without mortality, can there be a proper ethics in and of AI?"

This computational view is very West Coast. The cyberneticists—Wiener, Shannon, McCulloch, von Neumann—had a much more ecological view of how these connect and don't connect.

CHALMERS: That's a great question, and we can link it to the discussion we're having here. In terms of what is life and death in AI and what is autonomy in a computer, Dan said autonomy requires free will—well, then we're up against all the philosophical questions of what genuine free will is. Are you sure that it's essential that the AIs have free will or that they have consciousness? Those are very big questions. For questions of safety and human survival, what's really going to matter is what those systems can do.

For this debate, maybe we can just describe autonomy in very simplistic terms. A system is autonomous when it has a wide variety of goals and has the power to achieve them. To advance autonomous AI, it will be systems that not only have goals but can achieve them, compared to Dan's versions of AIs as tools that can advise you on how to achieve goals, and then you achieve them. This is a much more limited form of autonomy. I'm not sure that consciousness would be required for this. Once you have the AIs with goals and with the power to achieve the goals, that's already enough to get the party going.

DENNETT: The difference comes out if you compare good old-fashioned AI with contemporary AI. At the moment with deep learning and all the rest, what we have are these wonderful pattern-finding fabrics. They're great at finding needles in haystacks and doing other amazing things, but they haven't been formed into an architecture that's anything like an agent with its own goals and so forth. In principle, there are two ways you could go. You could go back to good old-fashioned AI and say, we've got these great fabrics, now we're going to do intelligent tailoring. We're going to do it from the top down. We're going to figure out what goals we want to install, and we're going to put in Asimov's rules. That's one way we could imagine going. That's very brittle, very unlikely, and much harder than people think.

The other way is to let her rip, bottom up, and let these things evolve and learn, and it'll be all done by bottom up quasi-Darwinian methods. If we go that route, then what we know right from the outset is that we will not be in control. We will not be in control, so we will be setting in motion something where the amount of autonomy the systems have will not be up to us.

I am not deathly afraid right now because the people who imagine this scenario and think this is coming soon are just wrong. Orders of magnitude of difficulty stand in the way. Take Watson, brilliant in its own way, I don't know

how many person-centuries of brilliant work went into the creation of Watson, which uses up the power of a small city. What percentage of an intelligent conscious AI is it? I would say a fraction of one percent. Turning Watson into an actual autonomous agent would be the work of many person-centuries of work, and nobody even knows how to do it yet.

CHALMERS: Watson is basically an exercise of what they call knowledge engineering. Give it a big enough database and a good way of dealing with all that data and knowledge and it can retrieve and apply that information in all kinds of ways.

DENNETT: It does some wonderful things. It really can. It's a great tool.

CHALMERS: Did you see those ads of IBM? "Watson does this. Watson does that." There are forty or fifty different Watsons out there. Watson is more a brand name at this point.

DENNETT: And some of it is hype, to put it politely.

CHALMERS: What excites people right now is machine learning, where you take systems from a whole lot of data and train them to do certain things. Supervised learning right now leads to amazing results in, say, image classification. Reinforcement learning, which is what was used to drive AlphaGo and AlphaZero, where you ba-

sically get the reinforcement of winning and losing—it turns out that's enough to drive learning and eventually unsupervised learning. You're right that it's where learning and evolution are involved that all this becomes extremely messy and hard to control.

When you have machine learning, you're always optimizing something. In machine learning, there's something called an objective function. The ideal of perfect behavior for your system—completely matching the training set on the images, or classifying a language right, or winning every game of GO—that's an objective function, and a good machine learning system will eventually approximate that objective function better and better. How it does it is not up to us. The objective function may be up to us, what you want your system to maximize and the behavior you want it to model. Once these systems have autonomy, that is, the ability to act and achieve their goals, that puts an enormous responsibility on us as the creators of the AI to get the objective function right, to make sure our systems are maximizing the right objective function. Roughly, they might have the right goals. The goal with your self-driving car is to get you to the destination, but also not to run into anybody on the way, to obey traffic laws, and so on. Once you've got systems with human level autonomy, then you want to get that objective function just right.

In some sense that's going to be the challenge of autonomous AI, finding a way to make sure our systems have the right goals and values, and this is where all this stuff about the messiness of wet cognition also enters. As human beings, we don't have *one* objective function, but rather we have many. We were thrown out not by a straightforward designer, but by a whole process of evolution with the ultimate value of reproducing our genes, so there are any number of little, messy objective functions along the way. It may well be that some form of artificial evolution will eventually produce AIs as wet and as messy in a way as biological systems. Humans are so unpredictable in every way, and in an international and sociopolitical context, that's not really a good thing. If AIs are as unpredictable as us in those ways, at a certain point we may wish we had simple AI with a simple objective function we knew about. Then at least we have an AI we can understand.

JB: The subtitle of this evening is "Philosophy and AI," and I have a question for both of you. How do you distinguish your work as cognitive scientists from that of philosophers?

I'll tell you a story. Twenty-five years ago, I did a book called *The Third Culture*, which involved a chapter with Dan. I went to everybody else in the book and had them talk about each of the other contributors. Marvin Minsky got on the

phone and I said, “Tell me about Dan Dennett.” He said, “Oh, the greatest philosopher since Russell.” Six months later I had to do fact-checking, so I went back to Minsky and said, “Let me read you what you said about Dan Dennett: 'He's the greatest philosopher since Russell.'” He said, “I said what? What I meant was, sure, he’s great, but he’s the only philosopher that understands what we do.”

You are one of the people that does the real stuff, so where does cognitive science end and philosophy begin? Let’s talk about the role of philosophy in AI because frankly I don’t get it. I don’t understand it.

DENNETT: There is a subfield in philosophy that has blended with cognitive science to such a degree that it roughly occupies the position that theoretical physics has relative to experimental physics. If you’re with people who have done their homework, they know the technologies. They’ve got hands that know how to code, but they’re interested in the theoretical questions and they’re interested in helping the engineers and the AI people sort out and understand what they’re up to.

It has been my fortune to be tutored over the last thirty or forty years by some of the leaders in AI. I’m not a good coder, but I have done some programming. The current generation of philosophers of cognitive science are superbly

well-trained and know a whole lot more than I did when I got into this or when Dave did when he got into this, and I was there when he was a graduate student. That's a very good sign. I was reading a dissertation today on predictive processing and the Bayesian Brain hypothesis, and it's a very technical dissertation by a philosopher.

JB: That's not going to get you the Berggruen Prize. In terms of the philosophy community that you call mainstream, I don't see how anybody focusing on AI would win one of their prizes at this point.

CHALMERS: Oh, I think you're wrong. Many of the philosophers who have thought about the mind, have thought about AI, and someone like Dan is a prime example. He's one of the leading philosophers of mind on the planet, and some very large part of that is from his thinking about AI. It's been very central in philosophy over the last few decades. The trend has been towards integrating the two pretty closely.

I did my PhD in AI, and my PhD advisor was not a regular academic philosopher. He was the AI researcher, cognitive scientist, and writer Douglas Hofstadter, well known to many of you for writing *Gödel, Escher, Bach*. He was also co-editor of a book with Dan called *The Mind's I*. Though I haven't done a lot of coding in the last couple of decades, for a philosopher to have

that experience of getting their hands dirty and building these systems, it just stayed with me. My technical knowledge is now thirty years out of date, but it gives you something to build on. That's partly how philosophers can educate themselves in the science and engineering and can contribute to it.

There's a big part of AI and cognitive science which is software engineering, but there's another big part of it which looks at what it tells us about, say, the human mind. That's no longer engineering; that's science and philosophy. We better start thinking about the relationship of these artificial systems to, say, human systems. Someone like Dan has done a lot there, whereas someone like John Searle on the other side has argued it tells us nothing about the human mind or about human consciousness. Anyway, we need philosophers to think about what this is telling us and what it's explaining.

There's also the social, political, and moral questions that ask not only what AI systems can we build, but what AI systems should we build? Dan just offered a proposal about that. Other people would offer a different proposal, but at some point someone's going to sit back and reflect on the ethical questions, which are going to involve reflecting on human values. What do we want as a society? Philosophers know how to think about human values, and

that's increasingly becoming pretty central to thinking about AI.

JB: That's very useful.

DENNETT: It's interesting, too, that in AI over the years there's been a similar gradient of philosophical interest. There are some people who are basically just engineers and that's all they want to be. They don't want to think about philosophical issues, and it doesn't mean they aren't doing great work. Some of them do important work. There's technical work by people who yawn when the issues involve how it relates to cognitive science or to the mind.

What I think is ironic is that if you go back to the early days of Herbert Simon and Allen Newell and others, there was an attempt to divide the field into AI and cognitive simulation. The idea was that cognitive simulation was using the computer to simulate human cognition, whereas AI was by hook or by crook anything that worked was fine. Oddly enough, the people who tried to do cognitive simulation ended up with these creaky GoFi models that didn't do a very good job, while the people who treated it as by hook or by crook ended up inventing deep learning and other systems like that, which we now realize maybe that's how the brain is doing it. It's come full circle, which is very interesting.

JB: Dan, Stuart Russell had a question specifically addressed to you. Stuart Russell is one of

the eminent computer scientists we all respect greatly: “Dan, you seem to divide AI systems into conscious entities and tools. Is there no middle ground (agent programs that presume explicitly represented goals in the real world)? Such agents could be arbitrarily competent, as AlphaGo in this little world, and yet non-conscious. Do you believe that consciousness will necessarily creep in as we make agent programs more and more competent in general? Can you tell us how *not* to make conscious AI system?”

DENNETT: Good question, and I’m glad he asked it. Indeed, we can have very intelligent systems that are not conscious in any interesting way, but they will seem conscious in some ways, and they won’t have important features that we have. It’s very much a matter of whether they are capable of taking their own interstates as objects of scrutiny and doing that recursively and indefinitely; that’s a very special feature. No non-human animal has that capacity, and that's the big difference between human consciousness and animal sentience.

I’m not going to argue about where consciousness stops or starts, but it’s important to realize that a lot of the techniques and structures that have been developed in recent years, which are just wonderful at analyzing causation, for instance—the directed acyclic graphs and Judea Pearl’s do-calculus and so forth—that can all be

accomplished unconsciously. We can tell a story where it looks like conscious hypothesis testing, but it doesn't have to be conscious. We can get all those benefits without any bit of acquaintance by the system itself with its own interstates.

JB: You mentioned Judea Pearl, so here's a question for you. Judea Pearl is the father of the Bayesian network, without which we wouldn't have AI as we know it today. He's a real giant in the field. He asks: "Is it too bold to assume that philosophy will soon melt into AI in the sense that all philosophical questions, especially those concerned with consciousness, will be reduced to problems in AI?"

CHALMERS: I would put it the other way around. Philosophy is pretty good at spinning off its problems into the sciences as we solve them. Isaac Newton considered himself a philosopher and figured out some good methods for solving the problems of space and time and so on. Okay, so we spun it off and called it physics. Along the way, philosophy spun off psychology and linguistics and so on. It's never the case that the spin-off solves the entire original philosophical problem, but we find some part of it that is tractable where we can find methods on which people can agree where they didn't agree before.

Did physics solve every philosophical problem of space and time? Absolutely not. Some of the biggest ones are unsolved. Did psychology solve the mind-body problem? No, absolutely not. There are as many views on the mind-body problem now in the age of psychology as there were before. Is AI going to solve the problem of consciousness? No, almost certainly not on its own. On the other hand, what will certainly happen is it will give us a whole lot of new insights. We'll get AI engineered systems that behave in remarkable ways where we're tempted to suspect that they're conscious and that someone may even think there's good reason to think they're conscious, but we're still going to need philosophical reasoning to think about it.

This now gets back to the question of the elephant in the room: Are these AI systems genuinely going to be conscious? This is not something you can just dismiss as a philosophical question. Why? It's deeply baked into our moral system as human beings that an entity has moral status. It's a system that we should care about if and only if it is conscious. If a computer system doesn't have any consciousness, then it's basically a tool. It might as well be like a car or a loudspeaker, in which case it doesn't deserve moral consideration. If the systems are conscious, then at least they enter into the moral sphere. They're systems that we have to start caring about. So, if most AI systems eventually are conscious, then we can't simply use

them as our tools, and we have to start thinking about these questions of whether they deserve equal respect, equal rights, and so on. It is, at least in my view, a crucial question.

My suspicion is, as AI systems develop and become more and more autonomous, more and more capable of reflecting on their own processes, more and more capable of reasoning, they're going to have a sense that they are in fact conscious systems. We're going to talk with them eventually. I imagine a conversation going like this: "You said there are some people over there. How do you know?" "Well, I saw them." "What was that like?" "Oh, I had an experience," and maybe they'll start reflecting on philosophy: "Well, I've read the owner's manual, and I know I'm just a whole bunch of circuits, but I feel like so much more." So, now they're going to say they've got the illusion of consciousness.

JB: That's qualia for you.

DENNETT: Well, that's all any of us have. There's one thing I think you underestimated. When I was working with Rod Brooks on Cog, one of the take-home messages from that whole experience for me is how little it takes in the way of animation and speed, particularly speed and grace, to convince most people that a robot is conscious. Cog was never within a country mile of being conscious, and yet there were MIT students who were banding together

to think about our moral obligations. Though it wasn't planned this way, Cog did have some strikingly persuasive behaviors, unconscious though Cog was. If you walked in the room when Cog was on, Cog's eyes would follow you across the room. That would freak people out. Shaking hands with Cog was a good one. I took one of my TAs over to the Cog lab to see Cog, and Cog's arm wasn't even attached to Cog's shoulder, it was c-clamped to the bench. Then Matt Williamson said, "Go ahead and shake his hand." She reached, and she shook its hand, and she screamed "It's alive!" because it wasn't clunky. It had elastic actuators, and that was enough.

What I am quite sure of is that we're not going to have a problem convincing people that robots have moral rights and are conscious. It's going to go the other way around. We're going to have a problem convincing them that, no, these aren't conscious. Not yet. You're being fooled by the tempo.

CHALMERS: There's some great psychological data on this, on when people are inclined to say a system conscious and has subjective experience. You show them many cases and you vary, say, the body—whether it's a metal body or a biological body—the one factor that tracks this better than anything else is the presence of eyes. If a system has eyes, it's conscious. If the system doesn't have eyes, well, all bets are off.

The moment we build our AIs and put them in bodies with eyes, it's going to be nearly irresistible to say they're conscious, but not to say that AI systems which are not in body do not have consciousness.

There's a website you can go to, People for the Ethical Treatment People of Animals. Well this is the AI analogue: People for the Ethical Treatment of Reinforcement Learners. And the idea is that very time you send a negative signal to a reinforcement learner not do something again, it's going to get a little bit of suffering. And every time you give it a reward, a little bit of pleasure. We have to make sure we give it a lot more reward than suffering. Okay, well, maybe that's not a good way to consider consciousness, but we have to think about these questions eventually. Once we get to the level of genuine autonomous agents, as Dan said, it's going to be very hard not to treat them as conscious, and that's going to raise many social philosophical questions.

JB: We're talking about ethics, and the elephant in the room is the ethics of the big five and what they do with your data, and how your reality is being programmed without your vote and your permission. Let me give you just a few words from George Dyson's chapter in the book:

> Norbert Wiener became increasingly disenchanted with the "gadget worshippers"

> whose corporate selfishness brought "motives to automation that go beyond legitimate curiosity and are sinful in themselves." He knew the danger was not machines becoming more like humans but humans being treated like machines. "The world of the future will be an ever more demanding struggle against the limitations of our intelligence," he warned, "not a comfortable hammock in which we can lie down to be waited upon by robot slaves."

Comment? You have to address these like this if you're going to talk about what you're doing and what AI people do.

DENNETT: Reading Wiener's book to write my essay for this was astonishing in a way because I had read it when I was an undergraduate and thought, okay... Then I read it today, and it's remarkably prescient in some regards. Some of the essays in the book are genuinely scary. People ought to read those essays and decide for themselves if some of the proponents there shouldn't be sat down and argued out of some of their blithe confidence about what the future holds. We have some serious problems looming, and we should take them very seriously.

CHALMERS: He talks about humans being treated like machines. I don't think I'm being treated like a machine, I think I'm gradually becoming a machine. Half of my memories are

now either stored on my smart phone or sitting in the cloud. I was trying to figure out the other day who has a bigger part of my brain. Is it Google, Apple, or Facebook? I think for now it might be Google. They've got an awful lot of my memories, my plans, my calendar system, my navigation system. We've all long since become these giant exo-organisms with this giant exo-cortex, as Charles Stross called the computer systems we're coupled with and the cloud. I don't go anywhere without consulting the Internet at least five or ten times in the process. What is it I'm going to be doing again? How do I get there? Who's going to be there and so on? It is true that these corporations own some rather large portion of my mind. If they wanted to do bad things with it, I'm in trouble. We're certainly in some sense in the situation of having to give them rather a large amount of our trust.

JB: If they want to do bad things with it?

CHALMERS: They're doing small bad things with it, relatively. They're not yet taking your mind and reprogramming it. Of course, they're brainwashing us bit by bit via the Facebook algorithm and the ads. I don't think they're malicious, the big corporations, they just have structural incentives. If someone genuinely malicious got control of those systems, then we'd have a dystopia, so we do have to think about that.

Q&A

Q: I was wondering what you thought the impact of this exo-cognition would be on evolutionary biology long term, where technology moves faster than genetic iteration. What does that mean for the long term of humanity as technology becomes more integrated into humans, too?

CHALMERS: What does the exo-cortex mean for our long term? What is happening is our minds are gradually migrating onto computational systems, so far only relatively small parts of the mind—memories, planning systems, navigation systems. We still have this conscious core that is mostly in the biology, and we're still exerting free will— if we ever had it—in the usual ways.

Long term, even that conscious biological core might itself migrate onto computational technology. After all, if we get to the age where we've got artificial intelligence systems at a level greater than human intelligence, biology is going to be slower and it is going to degrade itself, whereas over time we're eventually going to have the option of uploading our entire core onto computational systems. We're going to have to make a decision about whether that's something we want to do.

Doing this is going to offer the promise of immortality, probably of super fast processing of

enhanced reasoning. We'll be able to enhance ourselves so much more easily. Will there be down sides? Some people would see it again as dystopic. Some people would see uploading as a form of death. Maybe the uploaded system will no longer be conscious. Maybe it will just no longer be me. Maybe it'll be someone else. It'll be a copy of me and my twin. This is something Dan has written about beautifully in his past work "Where Am I?" We're going to need philosophers to think about those questions. Once this kind of uploading becomes a realistic possibility, that the exo-cortex turns into our whole cortex and we're faced with that choice about uploading, we have to ask ourselves if we want to step into that system. Will this still be me? My sense is at this point people are going to have to start reading copies of the great ancient texts like Derek Parfit on personal identity and Dan Dennett's "Where Am I?" in order to make an extremely practical decision.

JB: Do you want to be on a DVD or streamed?

DENNETT: I have nothing to add to what Dave said, but in reply to that question, it's important just to take a deep breath and remind ourselves how unbelievably complex brains are. The latest count says 86 billion neurons, but there is some reason to think that the glial cells and the astrocytes, which outnumber the neurons, are playing an important role in cognition. If they are, we've been studying the phone system all the time

while leaving the communicators out of the picture. When I started out thinking about brains, I had this simple, elegant model of McCulloch Pitts' logical neuron. I thought I could understand how this works, and I didn't want to see it get much more complicated. But it's become much more complicated by orders of magnitude. When you start realizing that it might even turn out that viruses play a role in modulating our brains, you realize that we may have a laughably impoverished view of the actual dynamics of the human brain.

CHALMERS: It's all in the quantum superpositions, Dan.

DENNETT: No, it isn't. I draw the line there.

JB: Venki Ramakrishan, president of the Royal Society and Nobelist in biology for the ribosome, has a question about evolution: "Is the evolution of carbon-based intelligence simply a catalyst in the evolution of silicon-based intelligence? One that can survive far greater extremes of environment? And does evolution even care about intelligence?" Speaking of viruses, in one of his interviews with me, he noted that at the MRC, which is an eminent biological lab where he's been deputy director, he said, "We make viruses. Deep learning is opaque. We can't go there. We have to know every step of the way and check it. Right now, it's a problem."

DENNETT: The problem of black box science hasn't been mentioned yet, but you basically raise it here. That's important. We're at the point now, thanks to the deep learning technologies, where we can delegate to black boxes finding the patterns in all sorts of very large datasets, and we don't know how the systems work. We're making oracles and trusting them. We can even have proofs that they're trustworthy, that they give very good answers, but this means a diminution in the role of the individual conscious scientist and also the distribution.

We're now moving away from the great scientist, the individual mind, and we're beginning to deal with distributed understanding, where no one person understands the results, rather, it's a team. That's a good thing. It's changing the whole structure of science, and it may do the same thing with philosophy, where the idea of an intelligent designer, whether it's a designer of a theory, or the discoverer of a scientific model, will be a role. They might have to just discontinue the Nobel Prizes, for instance.

CHALMERS: The point about evolution is an interesting one. It's commonplace, at least in humans, that the force of biological evolution is now largely being supplanted by the force of cultural evolution, which moves only so fast—the development of language, and writing, and computers and so on. It wouldn't at all surprise me if cultural evolution will continue to be a force.

But if this vision of AI is right, then at some point, cultural evolution itself may be supplanted by a different kind of artificial design evolution, where systems move ahead by leaps and bounds by humans designing artificial intelligences, which design ever greater artificial intelligences and so on. That could be a kind of evolution which greatly outstrips even ordinary cultural evolution. Then the question is, is that the future evolution of humanity? One view is the future evolution of a wholly different species.

DENNETT: That is in fact the view that Sue Blackmore argues for, that human hosts will no longer be necessary for memes to evolve, and we'll have what she calls temes, which are a technologically hosted meme. If you just want to have an example of how that might work, right now there are algorithms that are being used to predict the popularity of popular songs. They're getting better. The day may come where a song goes platinum without ever having been heard by a human being.

JB: That's the best take-away of the evening.

CHALMERS: I'm all for the future of where we're conscious of the song and someone's actually experiencing it, because that's where this teme is to have some value to someone.

Q: Is there any regulation happening among the U.S. government or any governments? Dave

mentioned that AI is going to take over elder care. I know there are already sex robots that are taking over so much, sadly, of people's lives. What keeps this from spiraling into this evolution where human beings don't matter anymore? Is there any regulation? What can we regular people do to make sure it doesn't go in that direction?

CHALMERS: It's a good question. Right now, there's a lot of discussion around the issue of regulating AI. The risks, the down sides, and the ethics of AI have become extremely prominent in the popular discussion over the last two, three, four years. I went to a conference in Asilomar about two years ago, which was devoted to coming up with ethical principles for guiding AI. We came up with twenty-three principles that were supposed to play some role. There's something called the Partnership on AI, which involves some of the leading companies involved in AI—Google, Facebook, DeepMind, and so on—supposedly coming up with some principles. There's also a fair bit of skepticism about how much difference that's making. It's easy for people to pay lip service to this kind of regulation, these kinds of ethical principles, but what happens when incentives are involved, financial, or military, or otherwise? "Avoid an AI arms race": That's one of the twenty-three principles. We don't want that because then it will go in unpredictable ways. It's a great thing to say, but what happens once the Americans and the

Chinese are in competition, possibly in a military situation? Do you think people are then going to say, “Well, there’s this regulation from a seminar”?

I did a talk at West Point, the military institute, a couple years ago and I asked people, “What do think is the military’s attitude towards superintelligence and the singularity? Is this something we should prevent?” They said no. Their attitude is, better American superintelligence than Chinese superintelligence. It’s an incredibly important question. I don’t know exactly what it is an ordinary citizen can do right now, but thinking about this, talking about it, and keeping the issue active in the public eye is a very good first step.

Q: I’m not sure exactly how to phrase it, and unfortunately it might go into a harder question of consciousness. You guys have been talking about the idea of uploading your mind or translating it into another medium. It doesn’t click with me because I’m thinking you could clone me, create another me with all of my experiences, but I’m not going to be able to share my awareness with that person. I feel as if it’s the same issue if, say, I’m just putting my intelligence into a computer. You’re erasing me now and then just copying it. It'll be someone who’s going to have my memories, but the me who exists now no longer will.

DENNETT: Just imagine this happens slowly: Your brain is dying and, thanks to technology, you get to upload a little bit every day, and so you get used to the fact that more and more of your brain is residing in the cloud and interfaced with you. Eventually, your biological brain is dead and you move right on. That's one of the possibilities. If you think of it that way, it's like the Ship of Theseus. You go right on living.

CHALMERS: If I'm ever going to upload myself, I'm going to do it that way, one neuron at a time. Stay conscious throughout. Here I am. Here I am. Now I'm here.

Q: We're easily convinced of consciousness. What would it take for you to be convinced of consciousness? Doesn't the Chinese Room state that it's practically impossible?

CHALMERS: Red flag. Take it away.

DENNETT: No, it doesn't. The Chinese Room is a defective thought experiment. I've said so for years. I don't want to talk about it. You can read the endless pages of what's technically wrong with John Searle's arguments. I know a lot of people don't care about the technicalities, they just like where he comes out. If you like where he comes out and you don't want to know the technical arguments, then go with it, but don't think that you're being convinced by a good argument. That's enough.

CHALMERS: What would convince me that a system was conscious? A lot of things, but probably the single thing that would carry the most weight would be talking to them. How am I convinced that another person is conscious? How do I learn what they're conscious of? Through talking to them, and they tell me about their consciousness.

If we encounter Martians and we talk to them, I hope they're going to be able to tell me about their conscious experiences. If an AI system says, "When I look at the world, I've got these experiences that seem to have a certain qualitative character," and I go, "Yes, I wonder whether another AI system would experience red things the way that I experience green things. Could an AI that just knew about my circuit diagram know what it's like for me to be experiencing certain smells, the smell of ammonia at a certain time?" That's the kind of thing that would convince me that it was onto something. Now, maybe it's just going to turn out that it's read one of those Turing Test guidebooks: "Talk like a human," and it internalized how to reproduce the illusion of consciousness. Assuming it's not something like that, that would be pretty strong evidence to me.

Q: It seemed from your earlier discourse that you hold human cognition paramount, and that the human needs to be in control of AI, but we know that humans are very pathological in their

decision making. You spoke of possible lines but not of possible problems. I'd like to hear your thoughts on uniting those two things. So, could you see a subset of the problems that we face today be outdated to AI since they might actually be better solutions than humans do?

DENNETT: I can foresee it, but I'm not sure I like what I see when I foresee it. Ask yourself whether you would be content. Maybe the answer is yes. I'll deliberately choose an example that is on the knife's edge. You've been charged with a heinous crime, and you have your choice between a trial by jury, a trial by judge, or a trial by AI. Now, would you want to have a trial by AI? If so, under what conditions?

CHALMERS: I'm not going to have one of those judges who's really hungry just before lunch.

DENNETT: That's why it's an interesting question. There is evidence of judges who are all too human.

Q: You spoke earlier, Dr. Dennett, that there's a difference between consciousness and intelligence. You have a superintelligence, not necessarily conscious, although it can be possible in principle, but who cares. That's not the point. You brought up Hofstadter earlier. Hofstadter has a book, *I Am a Strange Loop*—thinking is consciousness.

So, if superintelligence is not necessarily being conscious, could a superintelligence not be thinking? If so, would it be a dumb superintelligence? Where do you have this line of demarcation between very intelligent but not thinking at all, versus probably stupid but thinking, so, intelligence but also consciousness? Do you have any thoughts on that?

DENNETT: Yes, very simply, we tend to intellectualize our mission and think of it as thinking done by the brain, but in fact a lot of what we've learned in the last fifty years is that we now know lots of ways that unconscious processes can mimic conscious thought. In one of my books I introduce—actually in several of them—the idea of a Popperian future, which is one that tests hypotheses before trying them out in the real world. When people think about that, they think about somebody doing that, knowing that that's what they're doing and thinking about it in those terms. But in fact, you can get all the benefits of a Popperian creature completely unconsciously. Do you want to call that thinking? Not necessarily.

If, with Doug, you think that thinking ought to be reserved only for conscious cognition, that raises the best issue of how do we get a personal level out of a system that has all of this wonderful stuff going on at the sub-person level? Part of the answer comes from a wonderful phrase that comes from Jean Piaget that has

been improved by Guy Claxton, a British psychologist. Piaget said—Claxton says—"Intelligence is knowing what to do when you don't know what to do." If you think about that, you realize that if you're already equipped with instincts or training, then you know what to do under many circumstances. That's what your intelligence consists of. If you know what to do when you're given a novel problem that you have no training for and you haven't evolved for, what do you do now? Well, if you're really intelligent, you know what to do, which is think about it. It requires the recursions that both David and I have been talking about.

JB: Catherine Bateson asked recently: "Could you have a superintelligence that knows what it doesn't know? Can we get one that receives a question and says, 'I'll have to think about it. Let me sleep on it. I'll be back in the morning.'"

Q: It seems weird to me that so much of the cultural exception with where AI is going has to do with malice and an intentional AI doing bad things to people, because it doesn't take much more computing power than we have now combined with a malicious person to do bad things to everybody. If you could steer the conversation in some other way, what do you think is like a more productive bent than thinking about just AI getting conscious and doing bad stuff?

CHALMERS: I don't think malicious agents are the biggest problem. Here's something that's a soft problem, which involves people trying to exploit AI for evil ends. Even at the point of view of thinking about agents, many of the problems will arise just in thinking about structural factors—ordinary incentives to use AIs to make money, or to win wars, or to solve problems. We're going to have very strong incentives to build very powerful AI systems, and then that's going to have some spin-offs. It's going to have large effects, of which we have to be very careful. Those effects are just as concerning as Terminator-style scenarios, and probably much more likely because they can be almost expected to happen through ordinary forces. A lot of it also comes down to this question of having to get the objective function, the goals and values of your AI systems, just right.

The Terminator scenario is an AI whose value function is to destroy humanity. Well, that's not so good. To Nick Bostrum's point, maybe there's AI whose value function is just to create paperclips, and the AI decides that humans are taking up space where we could have paperclips, so bye-bye humans. Maybe it turns out the value function is we want the humans to be happy. What's my test for happiness? Make sure they have a smile on their face. If they've got a smile on their face, they're happy. Then, AI goes around plastering smiles on everyone. There are just a lot of ways for this to go wrong even with

well-meaning agents, and this is why we have to be careful in designing control.

JB: Final question to wrap it up. I'd like to ask each of you, what, if anything, have you learned from each other tonight?

CHALMERS: Tonight?

DENNETT: I've learned a lot from David over the years. Aside from getting clearer about what some of his current views are on these issues, there's much more agreement than I was expecting.

CHALMERS: I've learned a lot from Dan over the years. He was one of the first philosophers I read at my mother's knee as it were. There are many things on which we agree, and there are things on which we have very strong disagreements. Consciousness, for example, and to what extent that can be explained by the standard methods of science and to what extent it's a significant problem. It wasn't our focus here tonight, but from Dan one of the things I've learned is to think very hard about the relationship between consciousness and the way systems think about their own consciousness.

Now, Dan sees all this as a matter of delusions, that our thinking about consciousness is basically one giant delusion. There's the god delusion and the consciousness delusion. He thinks the greater truth underneath it, but at least

philosophical problems got going by this delusion. I love the view. It's an interesting view. I don't buy it myself, but I have learned to think about consciousness in human systems or in artificial systems, think about the mechanisms by which these systems are modeling their own minds, thinking about their own minds, and thinking of themselves as objects.

As Dan said tonight, that's the right perspective to start from in thinking about consciousness in artificial systems, even though Dan and I end up diverging on the path we take from there.