AI and You

Transcript

Guest: Anil Seth, part 1

Episode 137

First Aired: Monday, January 30, 2023

Hello, and welcome to episode 137!

My guest today is a renowned TED speaker with over 13 million views. **Anil Seth** is Professor of Cognitive and Computational Neuroscience at the University of Sussex in the UK, and author of the bestselling book, *Being You: A New Science of Consciousness*, which is exactly what we're going to talk about.

Now, if you're thinking it's off the topic of AI to be talking about consciousness – surely this is psychology, philosophy, or neuroscience at best – it's actually of primary relevance, and if you're a long-term listener to this show, you're aware of that connection. Actually, there's more than one connection. The most obvious one is a bit out there, and it's the question of when AI will become conscious. This clearly isn't very useful to people who are wanting to use AI right now to process their CRM data, of course, but bear with me, we'll come off this ledge in a bit.

In order to know when AI might become conscious, we need to know what being conscious means in the first place. This sounds like a *duh* sort of test – what could be more obvious than being conscious – but it isn't. It's so hard, in fact, that – and I'm not a philosopher, but it appears there is a whole field of philosophy devoted just to studying the ways in which we can't explain what consciousness is. Daniel Dennett said, and I'm paraphrasing here, because I didn't understand what he was saying nearly as well as I wanted, that consciousness doesn't even exist, and I find that very tempting for not only giving a convenient excuse for sidestepping the whole problem, but it has a certain Zen appeal to it. Of course, if you still think you have something going on between your ears that doesn't happen inside a rock or a rhododendron or a racoon then you may find that answer deeply unsatisfying.

There's also a viewpoint that we can't understand our own consciousness because it would be like trying to fit a bucket of water inside itself, but that's also a bit suss. We have ways of breaking down really complex things to understand them, so why not this one too?

Now, all this sounds fascinating, but really only a topic for philosophers and late-night college dorm arguments. Which until fairly recently was true. But now we're building AI that's conversational enough that it starts to make us wonder. Blake Lemoine was notorious for saying that Google's AI called LaMDA was sentient, which some people say is the same as being conscious or is at least closely related. Even if we don't think an AI is conscious, what if we can't tell one way or another? If AI imitates consciousness well enough, is that good enough, in the same way that an imitation of a chess game *is* a real chess game anyway?

But even if you don't think or don't care if AI could become conscious, there's also the question of when enough people will start *thinking* that it is. Because the very debate could reverberate throughout society. If we don't have good ways of answering the questions that will come up when AI starts looking

even more human, how disruptive might that debate alone become? So there you've got some of the AI context for looking at this.

Anil has been researching consciousness for more than two decades, is Editor-in-Chief of the academic journal *Neuroscience of Consciousness,* and has a PhD in artificial intelligence. Without further ado, let's get into the interview with Anil Seth.

Anil, welcome to the show.

Thanks. It's a pleasure to be here.

So, maybe you can give us an idea how someone gets started, in your case, how you got started in a field as fascinating, complex, and universal as consciousness?

I think you gave it away, it's because it's fascinating, universal, and complex that rather attracted me. I think it's not that there was a particular decision that I made, it was more that it was just an interest that I always had. I think a lot of people are fascinated by these questions about consciousness at a very young age. I remember as a kid, having these questions about, "Who am I, why am I me and not somebody else? What happens when I die?" Which became elaborated later on, when I was studying into questions about how does the brain generate consciousness, or how do they relate? Maybe it doesn't generate consciousness, maybe there's some other relation and the time when I was an undergraduate student in the 1990s, this was not really catered for in the courses that I was doing in natural sciences and physics, and then in psychology, but it still seems to be this fascinating question and my PhD was something rather different. I was doing a lot of Computer Science and Artificial Intelligence, evolutionary robotics, very different kinds of AI than people are doing these days with deep networks and so on. But nonetheless, AI and neural networks were involved. And it just occurred to me - I mean, all of this, in the end contributed to how I later came to think about consciousness as something deeply embedded, deeply embodied in an organism, something not associated intimately or not constitutively with intelligence. Some of the tools from early ideas of cybernetics that I learned about during my PhD turned out to be very relevant for how I now come to think about the brain as a prediction machine and how *that* is fundamental to conscious perception. But the honest answer is I sort of staggered around between different disciplines for most of my early academic career, and only when I was doing my postdoc, when I moved to San Diego in California, was I in an environment for the first time - this was now 21 years ago, 22 years ago, it's been a new year, 22 years ago - that I was in a place where focusing on consciousness, studying consciousness was something out there, in the air, explicit that people did and this was an era where Francis Crick was working there with Christof Koch at Caltech and my former boss, Gerald Edelman, and other Nobel laureates who also cashed in his Nobel capital to focus on consciousness. It was a very exciting place to be and so that was, that was the transition point for me and from then on, I've guided my research more explicitly to be about consciousness, its relation to other topics in psychology and computer science, and all the other things that go along with this, as you say, very universal, complex and fascinating area.

It sounds like you were involved in sort of Woodstock of consciousness, at the early days there with these people who were just giants in moving the field along with you.

It didn't quite feel like that at the time. I think that looking back, I do feel I was very lucky to be there at that time. But even then, most of what I was doing was not directly consciousness-related, I was helping build the so-called brain-based devices, which were robots controlled by neural networks that were heavily biologically inspired, which was very interesting and we learned a lot by doing that. Yet it was a more general I think, turning of the tide that was happening in neuroscience, in psychology, in general, then. That consciousness was not a purely philosophical problem. It was not something that only had fringe theories that sort of associated consciousness with some new physics or some weird quantum effect or something like that. There's a general appreciation that consciousness, yeah you can study it. It's a fundamental property that we know brains and bodies are intimately related to and there are a number of reasonable experimental and theoretical ways in to getting a handle on it. Even if you don't get all the way you can make progress and I think that's really the key, for a scientific enterprise to take off. You don't have to necessarily be confident you can answer the big question you set out with, but you need to be able to make progress and develop your understanding as you go.

Then let's get into that and make progress because I'm itching to ask questions I've been storing up for, oh, 50 years or so and for anyone that's been listening to this show for a while, this topic won't come as a surprise. But if this is your first time listening to the show, you might be wondering, "What are we doing talking about consciousness on a show about artificial intelligence?" And one of the answers to that is that we will shortly as a species, be confronting the question of how to decide whether an artificial intelligence is conscious or not, and we ought to have a good way of deciding that. So, that's one of the places we're going to end up, hopefully, in this interview. But consciousness, you know it's this word everyone uses 'conscious' like, "oh yeah of course" and then you say, "well, what does that mean?" and it's like "well duh, isn't it obvious?", but it's not like other "duh" quantities like mass or temperature. It eludes definition, at least for me in any sort of satisfactory sense and I think I should just say that, since I have had so much time to think about this, if any of the arguments I have stored up sound combative, they're not directed at you, they are just as a spirit of academic inquiry, right? So what should we use as a definition and one of the definitions is, is something awake or not, which I think is useful to surgeons and high school teachers, but not interesting to the rest of us. If you see the term consciousness in science fiction, it never means that, it's talking about something more fundamental. But is the word itself - do we even know what we're talking about when we - I don't mean you - but the rest of us throw this word around? Like we know what we mean. Are we even using the word to mean the same thing?

Well, that's the key, I think there are two reasons a definition can be useful. One is to actually specify the properties of the thing we're talking about, and I would agree that we don't have a fully consensus satisfactory definition of consciousness for that criterion. The other one is to make sure we don't talk past each other by assuming we mean the same thing when we actually mean different things. And so I, at the moment, prefer quite a generic definition of consciousness. It's an old definition that comes from a philosopher, Thomas Nagle and he put it like this, he said, something very simple, that for a conscious organism, there is something it is like to be that organism. It feels like something to be me, it feels like something to be you. But

intuitively, it doesn't feel like anything to be a table or a chair, or a cellphone. People might argue about the cell phone, but at least, the intuition is clear that for some systems, there is some experiencing going on, it feels like something to be that system and then for other things, they're merely objects. And you may say, well, that's just a very trivial, very circular definition; but I think it just, it's important, because it highlights the essential property that we talk about when we're talking about consciousness, which is the presence of experience and it also is interesting for what it leaves out. And here is where I think definitions are very useful, because there can be assumptions we make or intuitions we have about what necessarily goes along with consciousness. So, you mentioned wakefulness. Wakefulness is a good example. Most of the time it does, but not always, right? We dream when we're asleep, we're conscious, but we're not awake. And on the other side of that coin, we have people in the so-called vegetative state, where they might be awake, but the idea is that in a way that they're not having conscious experiences. So, wakefulness can be doubly dissociated from consciousness. There are other things that are different as well: intelligence, and I'm sure we'll come to this in the context of AI. Intelligence is not the same thing as consciousness, they're very different, in the definition of intelligence, very, equally hard to define, but maybe a very broad definition of intelligence is doing the right thing at the right time, or behaving according to goals and desires. That doesn't necessarily entail any awareness at all, and being aware doesn't necessarily entail and perhaps more than a minimal amount of intelligence. And there's also not the implication that we have to have a very explicit sense of self, as some people identify consciousness with self-consciousness, because we humans we typically experience the world and we also experienced ourselves as the subject of these experiences. But again, that's just the particular instantiation that adult humans have. Even in states of extreme meditation, people can experience things without having an experience of self. So, I think starting with a very simple definition like that is a good idea because in science as well, you mentioned mass and temperature. Now, these definitions weren't just given from the beginning. They were arrived at after a long process of trying to understand phenomena such as heat and temperature. Eventually, we now have a lovely definition of heat; but that was the end of the road rather than start of the road.

Well, that's a great overview and it was an overview that raised a lot of directions for us to go down. And you mentioned Nagle's definition, which is one of the ones that's always irritated me, because I was like, that's just begging the question. What do you mean by *like*, and *feels* and *be*? And it reminded me of a description of Einstein's relativity, and the similarly kindergarten definition of that is, what does it feel like to be on a beam of light? But that's reduced from two pages of intense mathematics, which you can *do* something with. And so, with Nagle's definition, what can you do with that? What does it open up? What does it enable?

Right, I think you can do a lot with it. So, it is a starting point, not the end point, again. So, we don't have the equivalent of general relativity, you know this mathematical glory yet for consciousness. Although there are some theories out there that that attempt to do that, like Integrated Information Theory is a very ambitious, very mathematically precise theory. But what can we do with the with Nagle's definition? Well, I think what we do is something like what we did, as scientists with a definition of life. And this is a parallel that is imperfect, but I think its

historically illuminating. Life was at one point thought to be a big scary mystery: how can you get living things from nonliving elements, there must be something supernatural almost going on here, some spark of life, some *élan vital.* This was the philosophy of vitalism and, of course, that turned out to be not correct. It's not that we understand everything about life, but instead of treating it as one big, scary mystery, like, "Where's the special sauce? What is it?" Maybe we don't need one at all. Biologists of the day, like nascent biologists understood that life is not one thing. It has many different attributes, there's metabolism, homeostasis, reproduction, all sorts of things. And began to explain how systems could deliver, produce, generate these different aspects, that collectively we associate with life. And so this problem of life, the hard problem of life, was sort of chipped away at. And now we don't feel that life is something beyond the reach of science, even though some parts of it are still unknown. And I think we can play a similar game with consciousness. Instead of treating it as one big scary mystery, like what is the magic special sauce that that gives systems experience, we say, "Ok what are the properties that conscious systems exhibit?" These can be both functional properties, things that conscious systems can do, by virtue of being conscious. And I think more critically what to use a bit of philosophical jargon we call phenomenological properties. So, properties of that experience, like why is visual experience the way it is? What's the character of visual experience compared to emotional experience, compared to… I don't know, an experience of intending to do something. And trying to explain these different properties of consciousness broadly - and this is the strategy I take in the book and in my work over the years - is to identify three rough aspects to consciousness: One is conscious level. So, this is the difference between being conscious at all and losing it, like you might do in general anesthesia; or global alterations, like in sleep, or perhaps in altered states, like psychedelics, something like that. Then there's conscious content which is when you're conscious, you're conscious of something. So, what explains a particular character of the experiences that I might have at any one time, and why are they that way and not some other way. And then the third bucket of things is Self. What is the experience of self all about? What are the mechanisms that underlie that and make that the way it is? What comprises the experience of self? And my hope is that by dividing up the big scary challenge of consciousness this way, and making progress on some of its elements, it's not sweeping the problem of consciousness away under the carpet, because you're still trying to explain aspects of the experience. You will make progress. Will you get all the way, and will there be an Aha! moment where we can say, "Oh, yeah, you know we understand consciousness, because this and it all makes sense and if you're sufficiently *au fait* with the theories and the math's, you can understand it in the same way that a smart physicist can understand general relativity and then you use those analogies like, well, it's about what it feels like to ride a beam of light, you might say something like it's what it feels like when the brain is integrating information or making predictions"? But that will be the endpoint. So that's the route for me. I think it's a perfectly plausible route for a science of consciousness to follow.

And so, is it fair to say that, then we're approaching this sort of deconstructive, descriptive approach here, like with your analogy to life, it seems to me that what we did, what the scientists that came up with definitions of life, said, "Look, we all have this idea that something's alive or not, we can look at something and say, 'oh, that thing's alive. That thing

isn't'," and there just edge cases where we might disagree like viruses. But by and large, we will have an idea that something is alive or something isn't; can we break that down into something more measurable than just asking someone, 'is this thing over here alive or not?'" Are we doing the same thing with consciousness? Is it that we all have this basic idea of what something what that is, we just don't know how to describe it, and now all this work is going into, how can we come up with a common understanding that would break down that way?

I think in part, yes. So, one of the things with consciousness compared to life is that, as you said, for life, there's some general agreed intuition about what's alive and what's not, both in terms of the difference between an organism when it dies, and the things that we have around us and there are some edge cases, as you say. Now for we need these kinds of intuitions to make progress and for consciousness we have them, but they're not as readily available, right? We have intuitions about ourselves when we lose consciousness, whether it's in dreamless sleep or anesthesia. We have some intuitions about some other animals, but not all other animals and there's lots of disagreement there about comparisons between different species with respect to consciousness. But we also have the humans. We have good intuitions, I think, about when we're conscious of some sensory information or rather, when some sensory information makes a difference to our conscious perception compared to when it doesn't. So, there are things to hold on to and that that's a good place to start. The second place to start is thinking more about what are the properties of consciousness that are were trying to explain? So, I did a very rough and ready job of saying there are three levels of content itself. But of course, there's whole traditions of work in philosophy more than phenomenological philosophy that try to describe what experiences are like, rather than just saying you're conscious versus not or did you see the dog or not? They try to develop a more sophisticated language that is a language of what experiences are like. And I think that is a very useful resource to draw on. If we can map that to the brain in ways that are not just arbitrary, but to have some sort of explanatory and predictive power, then we make progress. This is not a new idea. It's not a new approach at all. It's Francisco Varela, one of the pioneering Chilean neuroscientists who helped rehabilitate the study of consciousness. I call this neuro phenomenology. It's not purely looking for what happens in the brain when you lose consciousness. It's really trying to draw explanatory links by which particular patterns of brain activity and bodily processes make sense that they're related to particular kinds of conscious experience. And that's the way to go

And to save people a trip to the dictionary, when you're talking about phenomenology, what's the capsule definition of that?

The study of what experiences are like. I hope that's roughly accurate. But it's focusing on the experience itself. How do we how do we understand the nature of experience? How do we describe it? Focusing on the properties like, why is a visual experience the way it is what distinguishes a visual experience from an emotional experience? That's a phenomenological question. They have different characters, but what are they? You know, one's spatial, one has valence for instance, that's a difference. But there may be other differences there, too. They have different aspects of time, you know the experience of time is another classic question in phenomenology, like what does is it? It's not a succession of moments, it seems to have some

extension into the future and a little bit into the past. It has a character that again, marks it as distinctive, from other kinds of experience. That's phenomenology.

So, I like your breakdown into the level, the content, and the self, because I can relate to those and in particular, this one about the self seems to make a distinction that's worth drawing in this question that we will one day have about whether to decide if an AI is conscious. And there is this concept called The Philosophical Zombie that from the outside looks conscious, reacts like it's conscious, but allegedly on the inside, doesn't have any conscious experience. And so, we didn't in that case, is the philosophical zombie - and maybe you can give us some of the background of that as well - is the philosophical zombie, something that has the level and the content, but not the self?

No: the philosophical zombie - and this is a classic thought experiment in philosophy of mind that many philosophers have written about, David Chalmers has written about it a lot, but others have too - it's a conceivability experiment. The idea of a philosophical zombie is that it lacks consciousness entirely. So, a philosophical zombie me would be behaviorally indistinguishable from real me. But there would be - back to Nagle there would be nothing it would be like to be philosophical zombie Anil; no consciousness going on at all, not of level, not of content, not of self. So, the point of the zombie thought experiment, it makes the following argument. It sort of says that, if it's conceivable to at least imagine that there might be a version of a conscious creature like you or me, that is indistinguishable from the outside. So, it would be saying things about consciousness, it will be having exactly this conversation, which is fairly ridiculous, given there's no consciousness going on, but that's the point. It would be behaviorally indistinguishable, but without any consciousness happening. So, the idea is that if that's at least conceivable, then consciousness is not purely a property of the functions and materials that the system is made of. And then it gets a bit more sophisticated about this, it sort of says, "Well, it depends whether it's conceivable given the laws of physics we have in our universe" or whether that's something you might change and there are other versions of the zombie that that say, well, the inside can be different, but it's just indistinguishable on the outside. But then there are neurological philosophical zombies, which are also the same on the inside too. Anyhow, I've not found that much utility in these kinds of conceivability arguments, because to my mind - I mean, they're just generally rather weak arguments in philosophy, because they tend to turn on people's intuitions rather than logical force. And our intuitions are very changeable and, in some sense, they're inversely proportional to knowledge. So, if you know nothing about, let's say, aerodynamics, and somebody asks you to imagine a Boeing 777 flying backwards, well, you know, of course, you can kind of do that, right, you can just imagine a plane up there going backwards. But then when you learn about engineering and aerodynamics, and how planes work, you realize that actually, actually you can't build a 777 that flies backwards. Wings don't work that way; planes don't work that way. It has to go forwards if it's going to go. And so it becomes inconceivable. And so, I think that as we learn more about the brain and its relationship to consciousness, the idea that we could have a functionally equivalent system that had no consciousness at all, that might become much less conceivable. As we learn more.

Are you saying there that there couldn't be a philosophical zombie? Like one could not be created? Or could not exist?

I'm saying that one cannot assume that there could be. Now this does have some relation to this question of consciousness in AI, though, because we're already getting to the point where our natural human tendency to attribute consciousness to things that have some reasonable degree of functional capability is on the horizon. We have these quite powerful chat bots now, which you can have conversations with. They're still very easy to trip up if you want to trip them up. But they play on our anthropomorphic tendencies to attribute self-consciousness awareness to things that behave in a human-like way, to some extent. So, we have things that, if you like they're sort of playing at being philosophical zombies, because they might lull us into an idea that they're conscious based on the fact that they approximate functional capabilities of conscious humans in some in some sort of, I don't know, emotionally powerful way. But they're very far from the philosophical idea, which is truly behaviorally indistinguishable, you would never be able to know the difference between a conscious me and a zombie me in in that philosophical thought experiment.

Well, that's an interesting distinction there, because people right now find it easy to believe that the day will come when there will be an AI that produces a completely convincing impression of being conscious, alive, self-aware. And the Turing Test is, you know, worth mentioning, but it is rapidly becoming too dated. But if that is possible - and it seems like it may not be that far off - then one can imagine the sort of confrontation going on where people are saying, "Turn it off, it's just a machine" and it's going "No, no don't do that. That'll hurt. I don't want that" and they say, "No, you've got nothing on the inside" and it says, "Yes, I do. I know I do." And I think there are two questions. One is, have you declared that that's not a state that we could ever arrive at? Or, if it is, are we doomed to be unable to answer the question of whether there's anything on the inside?

At the risk of annoying you and your listeners I think we have to make a number of distinctions here. And there are four different questions going on, and there are in fact four different kinds of tests that you can think of as relevant here. So, you mentioned the Turing Test. The Turing Test is classic test of machine intelligence, right? If a machine passes the Turing Test, which is operationalized by humans not being able to reliably distinguish between the human and machine when both are pretending to be another human, then the Turing Test is passed and the conclusion is that the machine is intelligent, right? So, it's a test of the machine. But another way of thinking about the Turing Test is that it's actually a test of the human it's a test of what it would take for a human being, to be convinced that the system is intelligent. And in fact, many people have criticized the Turing test for exactly this interpretation that it's a test of human gullibility, and humans very often fail. We've seen examples of this over the years. But it's both the Turing Test and this inverse Turing Test are tests of intelligence. They're not tests of consciousness and consciousness is, as we said earlier, it's different from intelligence. So, then you've got another pair of tests, which is which, interestingly, came up in science fiction film, which is, of course, where a lot of the best ideas come up about these things. This was *ex Machina* by Alex Garland. And there's a piece of dialogue in that film, where they mention the

Turing test, but he actually twists the idea of it in a fascinating way, and makes it about consciousness, rather than the back intelligence and there's a piece of dialogue where he says, the inventor of the robot says, "Do you know what the Turing Test is?" And Caleb, who's this hotshot programmer, says, "I know what the Turing Test is" and explains it and then the inventor, Nathan says, "No, the real test is to show you that she's a machine, and see if you still feel that she has consciousness" and so that's now a test of that. It's a little unclear if it's a test of the machine or a test of the human, but again there's this pairing, you could say that the Garland Test is passed to give it a name? My colleague and friend and Murray Shanahan coined that term I know, there's the Garland test, which is, Can a person tell the difference between a machine and a conscious system by criteria of consciousness rather than intelligence? And equivalently, the inverse Garland Test: what would it take for a human being to attribute consciousness to a system? And so we now have these tests four distinct things that that are floating about. I think the conflation of these different questions is where we get into a lot of trouble. Because people can, for instance, say, "Okay, on the basis of a sort of Turing Test-type thing, because I conflate consciousness and intelligence. So, GPT 3, or 4 or 10, or whatever it is, or LaMDA, there was that big fuss about LaMDA last year, because it's convinced me that there's some intelligence going on, I am now convinced that it is conscious" and this is what the Google engineer Blake Lemoine basically was doing. I'm convinced that it's intelligent, therefore, it's conscious because it was saying things about philosophy of mind, apparently, that were unexpected to him, that it would say.

I think the word he used was "sentient."

Yeah, that's another slippery term. So, I tended to avoid that. There's three terms that come up in the discourse about this: There's consciousness; there's awareness, which I treat as synonymous with consciousness; and there's sentience, which has this, this semantic affiliation with feeling. And the problem with sentience is that it's also used by some people without any connotation of consciousness at all. So, some people just say, a system that's responsive to its environment is sentient. And because of the sort of dual use of that term, I tend to avoid it, and say, "Okay, let's just be clear, and restrict the use of consciousness and awareness to systems where we're making a substantive claim that experience is happening for that system, independently of whether it's sensing the external world or not." But yeah, so that so you can end up thinking, you're answering one question, but actually answering another. And if you take a test, ultimately, of what it takes for a human to attribute intelligence, if you mistake that for a test of whether a system itself is conscious, well no wonder you're going to get into problems. And I think that's a lot of the problems we've seen in some of the hype and fury about conscious AI and it's whether it's around the corner or not.

That's the end of the first half of the interview. We're splitting this one up because this is obviously already pretty dense. The rest will be next week.

So we discussed the Turing Test there, and although that's gone pretty mainstream now, thanks to the movie *ex Machina* among many other reasons, if you're not familiar with it, let me give the brief description. Alan Turing was a genius on a par with Einstein and Newton, and that's no exaggeration. Listen to episodes 128 and 129 when I interview his historian Johnathan Bowen to hear a lot more about

Turing. He thought about this question of how to judge whether a machine was thinking like a human back in a time when computers barely existed, and he'd just helped to invent one of the early ones that was used to crack Nazi codes during World War II. And he created what he called the Imitation Game that was also the title of a movie about him starring Benedict Cumberbatch. That evolved into what we call the Turing Test, which puts a human judge in front of a computer terminal in text conversation with either a human or a computer. After some period of time the judge is asked whether they think they've been talking with a human or an AI. The results of multiple tests and multiple judges are averaged together and if an AI is mistaken for a human enough times, Turing said we should think of it as thinking like a human. This rationale has been pretty good but it some deficiencies, and you heard Anil explore that a bit just now.

In today's news ripped from the headlines about AI, researchers at DeepMind and the University of Oxford issued a paper in AI Magazine that concludes that "a sufficiently advanced artificial agent would likely intervene in the provision of goal-information, with catastrophic consequences." We've heard guests on this show argue that the value alignment problem means that superintelligences could end up posing an existential threat to humanity through attempting to pursue the goals that we set them to the exclusion of what we need for survival. The archetypal form of this argument is the Paperclip Maximizer put forward by Nick Bostrom in his 2014 book Superintelligence, where humanity gets consumed by an AI that was given the goal of maximizing the output of a paperclip factory. Well, this paper, by Michael Cohen, Marcus Hutter, and Michael Osborne, titled, "Advanced artificial agents intervene in the provision of reward," approaches the problem rigorously. It discusses some potential approaches for mitigating the problem, with esoteric names like myopia, imitation learning, and quantilization, but while it certainly doesn't say that extinction is certain, Cohen tweeted that "our conclusion is much stronger than that of any previous publication — an existential catastrophe is not just possible, but likely." That may sound like the ultimate downer to end this episode with, but the encouraging thing about this is that it shows how much serious work is now being performed on this issue, and funded, in this case by the Future of Humanity Institute, the Leverhulme Trust, the Oxford-Man Institute, and the Australian Research Council.

Next week, we'll conclude the interview with Anil Seth, when we'll put Anil on an imaginary witness stand to hear his testimony, and we'll also discuss hallucinations, his opinion of ChatGPT, and how you can participate with Seth in what he calls the Perception Census. That's next week

, on *AI and You.*

Until then, remember: no matter how much computers learn how to do, it's how we come together as *humans* that matters.

http://aiandyou.net