# AI and You

Transcript

## Guest: Stuart Russell

## Episode 87

First Aired: Monday, February 14, 2022

Hello, and welcome to episode 87! We are going to conclude the interview with Stuart Russell, one of the key figures in the world of AI today. He is professor of Computer Science at the University of California at Berkeley, and director of their Center for Human-Compatible AI. He is the author (with Peter Norvig) of "Artificial Intelligence: A Modern Approach," *the* standard text on AI at the university level. Stuart is an Honorary Fellow of Wadham College, Oxford, and a Fellow of the Association for Computing Machinery, and of the American Association for the Advancement of Science. In 2021 he received the Order of the British Empire, and was chosen by the BBC to present artificial intelligence as the subject of their annual Reith lectures, which if you're outside Britain you can catch on the BBC's podcast.

His 2019 book, *Human Compatible: Artificial Intelligence and the Problem of Control*, is a must read for anyone interested in the impact of artificial intelligence on humanity's future.

Last week we talked about lethal autonomous weapons – Stuart is the spokesman for the Future of Life Institute's gripping viral video *Slaughterbots* – and how Stuart explains the need for attention right now on aligning AI's values with ours.

One of the reasons that I do this work is the big difference between the way technologists and lay people view the way AI is developing. The people who are developing it are gung-ho, excited, can't wait, want to make this stuff happen as fast as possible. The average lay person feels like a civilian during the Cold War who feels caught between the superpowers throwing their weight around and might at any time go nuts. It's like being in the back seat of a car being driven at breakneck speed along the freeway by someone who's chugging a Red Bull while texting a friend and turning around to say, "Hey, isn't this great? I've no idea where we're going but it sure is fun!" So I try to help those folk in the back seat—which is all of us in a way—understand what AI is and where it's going, and when I talk to technologists I remind them of what it's like for everyone else and ask them to make their work relatable. And I can tell that Stuart has a similar mission, on a much larger scale of course.

A couple of things we mention in the interview that are worth telling you about in advance. Firstly, the Chinese Room. We've mentioned this before, but it's been a while, so here goes. The *Chinese Room* is the name of a thought experiment by the philosopher John Searle, from the University of California at Berkeley – which, you may recall I just said is where Stuart is. Lot going on at UCB. Searle believes that only humans are capable of understanding, and he constructed a thought experiment to make his case. Imagine a giant box, like a shipping container, completely sealed, except for a hole on one side and a hole on the other. Into the first hole we put a piece of paper with a question written in Chinese on it. Chinese because Searle didn't understand Chinese—you'll see why in a moment. Some time afterwards, a piece of paper comes out of the other hole with the answer, also written in Chinese. Searle's question is, does the box understand Chinese? And then he goes on to say that the box hollow, and inside is a

man and a book. The man does not understand Chinese at all. But the book contains instructions that tell him what to do with the pieces of paper that come through the input hole. Like, "If you see a mark with this pattern, then go to this page of the book and execute instruction number 17. If it's followed by a mark like this, go to this page. Write a mark that looks like this on the output paper." And on and on, for many, many instructions. We know that this is a feasible scenario for working with Chinese because computers can answer questions put to them in Chinese by following instructions, albeit a great many of them. You just need a big enough book and a lot of time, but since it's a philosophical thought experiment, the size of the book and the amount of time are nobody's business. Searle says that since the man does not understand Chinese, the room does not understand Chinese, and that therefore if a computer gives the appearance of answering a question put to it in Chinese it doesn't understand Chinese either, and the same goes even if the question is in English.

Many people, myself included, disagree with Searle; there's a Wikipedia page full of rebuttals. Now, we're not adjudicating the Chinese Room in this interview, but I bring it up in the context of a question about Alan Turing's empirical approach to thinking, which is that by constructing the Turing Test (which he called the Imitation Game) to answer the question of whether a computer was thinking like a human, he declared it was only necessary to look at external signs, not to look inside the room, as it were. And so I was thinking, at what point do we actually take that on board with respect to computers, and consider only external signs and stop acknowledging that we can look inside the room, with the programs.

As we rejoin the interview, Stuart was just talking about how in science fiction AI has the element of being a big calculator until it spontaneously becomes conscious, which he has said elsewhere is an attribute of AI that is an unknown distance away, but isn't necessary in order for AI to be an existential risk. Here we go.

And you use the word "conscious" and you've said that that's far enough away in machine intelligence that it's not something that we need to look at now. And so much of the risks and benefits can arise from artificial general intelligence that doesn't need to be conscious. I actually feel like getting into this field has decreased my IQ, because beforehand, I knew what words like "consciousness," "intelligence," and "understanding" meant. And now I don't. It's not a promising trajectory. We talk about artificial general intelligence as though it's one thing and that is that we know what that means, and human level intelligence as well. But I think that we will likely find there are great many gradations to that. Do we lack a vocabulary or science of describing what artificial general intelligence would manifest as?

Yes, I think we do. And I think when it matters, it's not accurate to talk about human level intelligence; but mostly it doesn't matter at this point. What really matters actually is what my colleague Andrew Critch calls *transformative intelligence*, which means systems that are capable of bringing about transformative change in the world which could be for the worse. And certainly, when people write about machine IQ increasing or machine IQ exceeding human IQ, this is all nonsense. The construct of IQ is basically an approximation that says roughly - very roughly - our various intellectual faculties are correlated. Our ability to understand complex language, our ability to make complex plans about the future, our ability to discover regularities in our experience and extrapolate them. Those tend to be highly correlated. And so experts who

work in in that field of psychology often talk about, they call it the *g*-factor, which is sort of this general propensity. But they certainly acknowledge that people have different intellectual faculties in different degrees. And the extreme other end of that is the Multiple Intelligences theory that basically says you could be a genius at understanding complex language and an idiot at detecting regularities and extrapolating them into the future. So, and I think the empirical evidence, I'm not gonna say where it falls on that spectrum, but for AI systems, they are far beyond the multiple intelligence end of the spectrum, right? So Google, for example, has this massively superhuman memory, right? It's got 80 million books and hundreds of billions of web pages and trillions of hours of video and it remembers it all and you can retrieve it, and so on, but it can't plan its way out of a paper bag. It doesn't have a reasoning capability at all, and you can't make a human being like that. Similarly, AlphaGo is amazing at planning move sequences on Go boards and AlphaZero can also do chess and Shogi, and DQN can play all these Atari video games, but none of those systems can understand a word of any language and they can't remember anything, and they can't reason. So all they can do is plan. So the concept of intelligence for machines is completely different. The notion of general-purpose intelligence is valid in the sense that if you look at AlphaZero and its capabilities, it only generalizes across a certain range of problems. For example, it doesn't work for games like bridge and poker, where you can't see the cards that the other person has, you have--

--Perfect knowledge--

-- you have to be able to see the pieces on the board, otherwise the algorithm just doesn't work. So if you can relax that requirement, then all of a sudden you go from a narrow range of games that you can play to a much wider range of games, but you still need to know the rules, right? What if you don't know the rules? Well, that requires yet more different kinds of algorithms and system designs, and if you can relax that assumption, then now you're getting somewhat close, and that's kind of like general purpose intelligence, but there's more of those broadenings have to happen. But each of those broadenings is a research project that people are working on. We already have solutions to many of them, we need to integrate those solutions and there are a few more that we need to get. That's why I said, I don't think it's a hundred Nobel prizes, it's maybe half a dozen. And then we really would have systems that, just as AlphaGo can wipe the floor with humans on the Go board and the chess board and so on, these general-purpose systems would wipe the floor with humans anywhere.

I think the big surprise has been for me how many tasks AI was able to exceed human performance at, that we thought would take general intelligence. Like when Deep Blue beat Kasparov, Douglas Hofstadter said, "My God, I used to think that playing chess required thinking now I realize it doesn't," which had some collateral fallout for chess players, but that wasn't his target.

I think that it does require a particular kind of thinking,

But narrow - it can be achieved with narrow thinking. It may not be how humans do it, but it can be how computers do it.

Chess programs use only one form of reasoning about action, right? They only think about sequences of moves beginning from the current position going forward in time. They don't reason backwards from a goal. They don't say, "Oh maybe I could capture his queen, let me see if I can come up with a plan to do that." So it's a very restricted form of reasoning about action, but there are other programs that do work back from goals. There are other programs that do subdivide problems into different sub goals and then work on them separately and combine the solutions and so on. And that covers quite a lot of human reasoning about action. We also do a lot of kind of coarse-defined reasoning where we have a coarse plan, like, I want to go on vacation, I want to go to Hawaii, I want to stay on the Big Island, I guess I should fly there, all the way down to, when you actually do it, it's millions of motor control actions with your feet and your fingers and your hands and your vocal apparatus and all the rest to actually go on vacation. And so we are incredibly good at that sort of coarse-defined thinking, what we call *hierarchical abstraction.* And again, we've got algorithms that do some of that. And some of those came from the kinds of research that people were doing on chess back in the '60s and '70s. So I think there there's no good reason to believe that we can't achieve general purpose intelligence in machines.

Do you think we're currently hardware-bound or software-bound on that?

I think absolutely we are software-bound; we need new conceptual breakthroughs.

So if we had the software, we could run it on today's hardware.

I think so. Interestingly, Arthur Samuel wrote a machine learning program that taught itself to play checkers at a really quite good level, even went on to beat some professional players at some point, he was using a computer that was, I think, 10 trillion times less powerful than the one that AlphaZero used to learn to play Go and chess. Right? We don't think of checkers as 10 trillion times easier than Go and chess, right? So I think the evidence suggests that we have way more computing power than we know how to use efficiently.

We were talking about - and we've been talking a lot here about machines not really *understanding* a problem, and as I said, "understanding" is one of the words that I no longer know what it means, and we are quick as computer scientists to say about an AI, "It doesn't really understand anything, it doesn't understand these things that you think it does in order to solve this problem." And I wonder whether we are unconsciously or accidentally rejecting Turing's empiricism as a result of not giving any quarter to just the results that are achieved. I sense that somewhere in the background John Searle is laughing and saying, "See I told you so, this is the Chinese Room argument. Now you're saying that it doesn't count unless you can look inside the room." And I have this fantasy that one day there will be something that is an AI that is so externally empathetic, compassionate, communicative, creative, humorous, all of those things, that laypeople will be saying "This thing is alive," and computer scientists will be saying "No, no, no, no, it's a transformer. We know every line of code." And I wonder whether we could end up where computer scientists are not the first people to declare an AI conscious, but the last. Is that a future you think is possible?

So you said a lot of things, let me try to unpack. So the first thing is the critique that the AI system doesn't really understand can be interpreted on two levels. One, often philosophers often use the word "understand" this way, it means "understand" in the way that humans do, which is, in some sense, bound up with our subjective experience. And then there's "understand" in the sense that the AI system is going through reasoning steps and drawing on knowledge that is sort of analogous to the way a human might work on the problem. So for example, when the chess algorithm rejects the move, because it's looked ahead and said, "Oh yeah, you know, if I do that, he's gonna take my queen," that understanding is roughly analogous to the understanding that the human has of why that move is a bad move. So there are plenty of cases where at least in that sense, AI systems are reaching the conclusion for the right reasons in ways that roughly map on to how humans are reaching the same kinds of conclusions. But a lot of the critique now is about, are deep learning systems doing anything that can be mapped onto something that would be called understanding in humans. Are they understanding the images in any sense that - so when, when a human looks at an image of the house, we, in some sense parse the image, right? We can say, yeah, there's windows and a roof, it's put together in the right way, a particular style of architecture etc, etc. And so we can recognize it as a house, by that process. But what we're finding with some of these deep learning systems is that they recognize the breed of dog by looking at the color of the carpet. And clearly in that case you would say it doesn't really understand what "Norfolk Terrier" means because it's not even looking at the damn dog. And the state of the art in in actually diagnosing what deep learning systems are doing is pretty weak. There's an interesting theory going around developed by Adi Shamir, whose who was the "S" In RSA Security. And his proposal is essentially that deep learning systems are really acting as look-up tables; that effectively they're just storing all the training examples and around each training example, there's sort of a little ball of other images that are sufficiently similar that it would recognize it as being the same thing. And in between, it's No Man's Land and it's anyone's guess what it's going to say. And I don't know if this is fully verified, but it does a good job of predicting some of the weird phenomena with what we call adversarial examples where you can take a photograph of a school bus and you can change a few dozen pixels of a million-pixel image and it still looks exactly like the same school bus and now the algorithm is confident that it's an ostrich. And those kinds of examples, clearly humans are not susceptible to them, and so this is pretty clear evidence that the deep learning systems are not understanding the images in the way that humans are understanding the image in a robust way in terms of understanding the components and their spatial organization and coherence and so on. So if that's true, if deep learning systems really are sort of glorified look-up tables on steroids, that has a lot of implications. One is we're never going to get to general purpose intelligence by scaling up deep learning systems right? And in some sense we're making the same mistake that we did in the '50s and '60s where we said, oh look, you know, we can find a two-move plan to solve this problem. So it's just a question of scaling up right? And then we'll have human intelligence, right? At that time we didn't understand the theory of computational complexity which says that a lot of problems simply cannot be solved by scaling up the amount of computational power.

NP-complete.

Yeah the NP-Completeness. But, fairly soon after NP-Completeness was invented, late sixties, early seventies, and fairly soon after that AI sort of got to grips with it and understood the implications. But I think we're making the same mistake with data. We're saying, "Oh look, we can learn this concept from only a million examples. So if we just scale up the data, we'll be able to learn everything." And I just think that the universe will never contain enough data for that to be the case. Coming back to your point about Turing's empiricism. He actually in one of his lesser-known papers, he actually says it actually does matter what's going on inside, because if a system passed the Turing test using a look-up table then you'd have to imagine that one's conclusions about the consciousness of the system would be different.

It feels like a cop-out; like you could just make a longer test and that would defeat any look-up table?

Yeah. So he's saying as a thought experiment, would we say passing the Turing Test implies consciousness? And he says, well, clearly it matters. And if we found out that it was actually just using a look-up table, that would change our conclusions about whether the machine was conscious, or at least whatever consciousness was occurring would be of a very different nature than if it was doing something much more human-like inside.

That sounds exactly like Searle's book inside the room.

Well, no, I think it's the other way around, right? What Turing is saying is, if the operation of the machine producing intelligent behavior does generate consciousness, if there is subjective experience happening, surely you'd have to accept that the nature of that experience is going to be different in the two cases, whether it's a look-up table or some very sophisticated process of reasoning and similarity judgments and so on and so forth. Because otherwise you're saying that the subjective experience is completely nonphysical. In other words, it doesn't matter at all what physical processes are occurring. It's the same subjective experience. Well, I think nobody believes that.

It's fascinating. I'm aware that we're running out of time. I'm sure you can tell that I would love to talk about this for days on end. It's one of the limitations of humans that we usually can't. Do you see any technology on the horizon, any approaches that are outside of the box of current deep learning methods that could lead to breakthroughs in this area?

Yes, I think all the technologies that we developed before deep learning, they're already out there: for example, Google's knowledge graph, which is billions of facts, is answering about a third of all the questions asked on the Internet. So arguably that technology has had more economic impact, has more economic value than all the deep learning systems put together.

And the difference between the symbolic and pattern AI makes me think of the joke about, you ask a computer 2+2 and it says 3.97 and the deep learning expert goes "Great, close enough," because that would be a good result in there. But there are some problems where we need 100% and have every right to expect a 100% result. Is there anything in AI at the moment that enables it to switch between a deep learning approach and symbolic approach according to the type of problem?

I think there are there are the beginnings of some kind of integration of deep learning with a technology called probabilistic programming, which is a symbolic technology based on probability theory and either Universal Turing Machines or first order logic. And a key advantage of probabilistic programs—well, there's several actually—one is that it's based on a rigorous semantic theory. So we can analyze the semantics of each piece of the program separately and we know when we compose them what the result is going to be. And they are symbolic, so we can actually look at the reasoning processes and understand what they are doing. But there are clearly some processes that it doesn't make sense to try to do them symbolically. So if I ask you to differentiate between "cat" and "cut" as sounds, the symbolic process wouldn't get you very far, right? There's just a flavor in the vowel, and that seems to be a very sort of continuous thing. In fact, there is a pretty much a continuum in vowel space and different things—

And would give wrong answer with different accents.

So the natural thing seems to be that up to certain levels of the perceptual hierarchy, deep learning would be the kind of thing that would be happening. And then as you go further up, you start to get into places where symbolic reasoning makes more sense. And that would be my bet for how we'll get to general purpose intelligence.

That sounds like a good place to look for a conclusion here. What would you like to draw people's attention to in what you have done recently and what you will be doing shortly?

If people are interested in the general question of how we handle superhuman intelligence than the book *Human Compatible* is one place to start. Max Tegmark's book *Life 3.0*, is another one. and Nick Bostrom's *SuperIntelligence*. So they all offer different perspectives with different degrees of optimism. And I think the Reith Lectures, if you want something a bit less dense, the Reith lectures are available on the BBC website and those cover both the superintelligence question and autonomous weapons, and something we haven't talked about, which is the future of work and the human role in the world when we have general purpose AI systems.

Indeed, we have only so much time, and you got into the economy and the impacts on that as well. So I recommend those lectures to people. To have people not come away from that feeling, it was a "absolutely terrifying dystopian vision of the obliteration of the human race" what context do you want people to listen to them from? Assuming they're not *Guardian* reporters.

I'm wondering if the *Guardian* reporter just listened to the first lecture. And I did leave it hanging at the end of the first lecture. But the fourth lecture is: okay, can we do something different? Can we do AI in a different way that doesn't leave us vulnerable to this existential risk from superintelligent AI systems and humans losing control? Which is what Alan Turing predicted. And I think if you listen to the first lecture, you should definitely listen to the fourth lecture right after that.

Right. And you had a very good, and to my ears, novel, take on solving the value alignment problem there, which I encourage people to listen to. Thank you very much for coming on the show.

All right, thank you, Peter.

That's the end of the interview. I like how Stuart leverages his considerable academic prowess and platform at AI into exhorting calls for action that can and should be taken today.

By the way, you may notice that my fantasy of the empathetic AI that everyone except computer scientists think is conscious is what David Brin brought up a couple of weeks ago. I continue to think that this is a likely scenario not too far away.

When Stuart talked about AI recognizing the breed of dog by the color of the carpet, by the way, he was referring to a thorny problem of image recognition that it often trains on things that have nothing to do with the subject in the image, or what we would think the subject is. Like if you're looking for sheep, many pictures of landscapes will come up as being of sheep even when there are no animals there at all, because the model was trained on so many pictures of sheep in landscapes – since that's where sheep are usually found – that it's decided that the landscape is actually a part of what it's recognizing. It's very hard to get AI to tell us what parts of an image it has incorporated into its recognition model. There are other examples, like cancerous cells being recognized by rulers because all the cancer cells in the training set had rulers next to them. You've got to be careful with this stuff.

And when Stuart mentioned the g-factor of intelligence, my guest Kristóf Kovács, the supervisory psychologist of International Mensa, discussed that way back in episodes 10 and 11.

Speaking of episodes, we are now well over 50,000 downloads, by the way! In fact it'll probably be 60k by the time you hear this. Thank you, thank you, thank you. The league table of listenership by the way runs USA, UK, Canada, Australia, India for the top 5, but there is an impressive long tail of about 80 other countries listening. Please, please, please share about this show, write 5-star reviews, because that's where people are going to heard about this.

I have to say I'm not in agreement with the idea that if AI is a look-up table it can't be conscious. It seems like a slippery slope argument, because next you're saying that a Markov Chain can't be conscious, and pretty soon neither can an artificial neural network, because you know every line of code. It seems to me, on my more judgmental days, that some people are operating pretty much like large look-up tables themselves, if you think about how predictable some people are. Does it make a difference how big the look-up table is? Isn't the argument for consciousness as an emergent phenomenon basically saying that it happens if you connect enough neurons together, that it's a property of large scale? Anyway, that's where I'm at about that.

And by the way, Alexa's way of doing calculations pretty much parallels my quip about 2+2 = 3.97. If I ask her any calculation with numbers much longer than that, odds are she'll get it wrong. She seems to be treating them like every other question, which is to find the closest example from the Internet training data. On the other hand, I think Siri hands those kinds of question off to Wolfram Alpha. Don't quote me on that.

In today's news ripped from the headlines about AI, MIT researchers have developed a model that understands the underlying relationships between objects in a scene. (We're not going to quibble about the word "understand" after so much time on it in these episodes.) Co-lead author Yilun Du said that we don't think of, say, a table as having an object on it at certain (x,y,z) coordinates. We think of it more like, "A wood table to the left of a blue stool. A red couch to the right of a blue stool." Their system can take a description like that and use it to build an image, using a machine learning technique called energy-based models. It can also work the other way, and take an image and extract relationships from it. It outperformed other deep learning methods and was robust when dealing with descriptions it hadn't encountered before. This has obvious and exciting application in robotics.

Next week, we will have a panel of experts and researchers all working in the field of AI music on the show. I get to ask them all about how AI can understand and compose music. Did you know there's now an AI Song Contest? If you think about the Eurovision Song Contest – it's nothing like that.

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**