McKinsey
Global Institute

How to ensure artificial intelligence benefits society: A conversation with Stuart Russell and James Manyika

Leading artificial-intelligence researcher Stuart Russell shares in a conversation with James Manyika why a new approach for AI is necessary.

January 2020

Stuart Russell, a leading artificial-intelligence (AI) researcher at the University of California, Berkeley, and author of the book *Human Compatible* (Penguin Random House, October 2019), sits down with McKinsey Global Institute chairman James Manyika to discuss our future as AI transforms our world. In this broad conversation, they explore the immense benefits ahead and what our role will be as AI becomes more pervasive. They also delve into potential challenges we may face with our current approach to AI, and how we can redefine AI to ensure it helps humanity achieve its full potential.

How AI could improve everyone's quality of life

AI has the potential to change the world. UC Berkeley's Stuart Russell shares with James Manyika what excites him the most about AI.

James Manyika: When you look at the AI field today and you see all these announcements and breakthroughs, what excites you the most?

Stuart Russell: With today's technology, delivering high-quality education to everybody on Earth is just the beginning. Even fairly simple AI tutoring tools have been shown to be very effective. So that can only get better if we figure out how to roll it out to the people who really need it. There are hundreds of millions of people for whom education simply doesn't exist. It's out of reach. It's too expensive. It's too difficult for governments to organize. And AI could really help.

And from the beginning of time, it's been a struggle for us to have enough for everyone to have a good standard of living. And now perhaps we have a chance to get there.

Imagine 200 years ago, and you decide, "OK, I want to go to Australia." That would be a five- or ten-year project costing, in modern terms, probably billions of dollars. And you'd have about an 80 percent chance of dying, right? Incredibly complicated. You would need hundreds, if not thousands, of people to put together the ships, the expeditions, to fit them out, to man them, and so on.

Now I take out my phone, tap, tap, tap, and I'm in Australia tomorrow. It costs a bit, but compared to what it used to cost, it's nothing. Travel is now travel as a service; it's a utility just like electricity and water.

There's still lots of stuff that is expensive and complicated—construction projects, education, scientific research, and so on. But with human-level AI, it's "everything as a service." Everything takes on the characteristics that travel has today.

James Manyika: In other words, what you're describing is a bountiful economy. We'll have access to services. We'll have access to things that are otherwise prohibitively costly today.

I want to ask something about that. One of the things that you sometimes hear from AI researchers is the idea that AI assistance could actually allow us to do better science, to discover and work on things like climate science, new-materials discovery, or other things. Is that a real possibility, that we could actually do new kinds of science and achieve new kinds of discoveries?

Stuart Russell: I think we already are. Computational chemistry is an example where people use AI systems along with the ability to simulate the properties, the materials, or simulate reactions and the behavior of molecules, or search for molecules or materials that have a given property, electron energy density, or whatever it might be. AI's starting to accelerate these processes dramatically. And similarly, it's accelerating just doing the basic science research of putting together a complete picture of all of the molecular processes that occur in cells.

Again, it gets beyond the capability of the human mind. It wasn't built for us to be able to understand it. It was built by a process of evolution that's produced something that's incredibly complicated. And every few years we discover another realm. I was just reading about small proteins the other day.

There's an entire realm of things going on in the cell that we didn't even know.

I think climate science is another problem around the totality of the picture where AI can help. You've got atmospheric specialists, you have ocean people, you have cloud people, you have economists who look at migration and mitigation and so on, you have got the biosphere people who look at bacteria and processes of putrefaction of peat bogs and Siberian permafrost, and all the rest of it. But does anyone have the whole picture in their mind? AI systems could have the whole picture.

What's our role in an AI-driven world?

UC Berkeley's Stuart Russell discusses with James Manyika what skills people will need as AI systems automate more tasks.

James Manyika: Everybody worries about work and the role of humans. When we start to have these highly automated systems producing everything and doing a lot of things, what's the role of humans?

Stuart Russell: What's left is our humanity—the characteristics of human beings that machines don't share. Sometimes people use the word *empathy* for this.

James Manyika: But machines could mimic that.

Stuart Russell: They can mimic it, but they don't know what it's like.

James Manyika: Does the distinction matter?

Stuart Russell: It matters a lot, because what it's like then directly affects how you respond to it. For example, if I'm writing a poem, by reading my own poem, I can get a sense of how it would feel for you to read the poem. Or if I write a joke, I can see if it's funny.

This is one of the things where machines are funny right now, but only by accident. In principle, AI could superficially learn by learning five million jokes and five million non-jokes, and AI would try to learn a distinguishing thing. But that's not the same as actually finding anything funny. The machine just doesn't find it funny. It doesn't really know. In the more complex settings of interpersonal relationships, I think we have this comparative advantage. And we may even want to reserve the realm of interpersonal relationships for humans.

James Manyika: There are some who would actually argue that you could perhaps show empathy much better with automated systems.

Stuart Russell: You can have them simulate empathy. But part of what I value is that you actually care, not that you appear to care. And I think that's really important.

Interpersonal relationships are great, and there are some professions that do this already: the executive coach, the childcare provider, psychotherapists, and so on. It's a mixed bag as to whether these are well-paid, high-status jobs. But mostly, in terms of numbers, the vast majority are childcare and elder care, which are low-status, low-pay jobs. The question is, why? Our children and our parents are the most precious things. And yet we're paying $6 an hour and everything you can eat from the fridge for someone to look after our children.

Whereas, if I've got a broken leg, I'm paying $6,000 an hour to the orthopedic surgeon to fix my broken leg. Why? Well, because the orthopedic surgeon is resting on hundreds of years of medical research and ten years of training. Whereas for childcare there's almost no training. There's very little science on how to do a good job.

How does one person improve the life of another? We know there are people who can do it. But, generally speaking, there's no how-to manual. There's no science. There's no engineering of this. We put enormous resources, in the trillions of dollars, into the science and engineering of the cell phone, but not into the science and engineering of how one person can improve the life of another. And I think that's what we're going to need in spades. Because having material abundance is one thing. Having a rich and fulfilling life is another.

What could go wrong with AI—and how we can make it right

James Manyika and UC Berkeley's Stuart Russell discuss why it's dangerous to give AI systems "fixed objectives," and how we can redefine AI to ensure it's beneficial to humanity.

James Manyika: As you think about all these technologies and this enormous bounty and economic and societal benefit and potential, but, at the same time, if we do achieve these breakthroughs, what could go wrong with AI? I can imagine all the things that would go well.

Stuart Russell: Making machines that are much more intelligent than you: What could possibly go wrong, right? Well, this thought hasn't escaped people over time. In 1951, Alan Turing raised the same question. He said, "We would have to expect the machines to take control." And he actually referred to an earlier book. He says, "In the manner described in Samuel Butler's *Erewhon*," which was written in 1872—fairly soon after Charles Babbage developed a universal computing device, although he never built it. But Babbage and Ada Lovelace speculated about the use of this machine for intellectual tasks. The idea was clearly there.

In *Erewhon*, what Samuel Butler describes is a society that has made a decision, that they have gone through this debate between the pro-machinists and the anti-machinists. The anti-machinists are saying, "Look, these machines are getting more and more sophisticated and capable, and our bondage will creep up on us unawares. We will become subservient to the machines and eventually be discarded."

But if that was the only form of argument, which says, "Look, smarter thing, disaster," you might say, "OK, then we better stop." But we would need a lot more evidence. And also, you would lose the golden-age benefits—all the upside would disappear.

I think to have any impact on this story, you have to understand why we lose control. The reason actually lies in the way we've defined AI in the first place. Our definition of AI that we have worked with since the beginning is that machines are intelligent to the extent that they act in furtherance of their own objectives, that their actions can be expected to achieve their objectives.

James Manyika: Objectives that we give them, presumably.

Stuart Russell: Yes, so we borrowed that notion from human beings. We're intelligent to the extent that our actions achieve our objectives. And this, in turn, was borrowed from philosophy and economics, the notion of rational choice, rational behavior, and so on. We borrowed that notion of intelligence from humans. And we just said, "OK, let's just apply it to machines." We have objectives, but machines don't intrinsically have objectives. We plug in the objective, and then we've got an intelligent machine pursing its objective.

That's the way we've done AI since the beginning. It's a bad model because it's only of benefit to us if we state the objective completely and correctly. And it turns out that, generally, that's not possible. We've actually known this for thousands of years, that you can't get it right. King Midas said, "I want everything I touch to turn to gold." Well, he got exactly what he wanted, including his food and his drink and his family, and dies in misery and starvation. Then there's all the stories, where you rub a lamp and the genie comes up, what's the third wish? "Please, please, undo the first two wishes because I ruined everything."

The machine understanding the full extent of our preferences is sort of impossible, because *we* don't understand them. We don't know how we're going to feel about some future experience.

In the standard model, once you've plugged in the objective, certainly it may find solutions that you didn't think of that end up tweaking part of the world that had never occurred to you. The upshot of all this is that the best way to lose control is to continue developing the capabilities of AI within the standard model, where we give machines fixed objectives.

The solution is to have a different definition of AI. In fact, we don't really want intelligent machines, in the sense of machines that pursue objectives that they contain. What we want are machines that are beneficial to us. It's sort of this binary relationship. It's not a unary property of the machine. It's a property of the system composed of the machine and us, that we are better off in that system than without the machine.

A paradigm shift: Building AI to be beneficial rather than simply intelligent

UC Berkeley's Stuart Russell shares with James Manyika principles for creating beneficial AI systems.

James Manyika: The notion you described, these kind of provably beneficial systems—what makes it beneficial, by definition?

Stuart Russell: We don't really want intelligent machines, in the sense of machines that pursue objectives that they contain. What we want are machines that are beneficial to us.

What we're actually doing is instead of writing algorithms that find optimal solutions for a fixed objective, we write algorithms that solve *this* problem, the problem of functioning as sort of one half of a combined system with humans. This actually makes it a game-theoretic problem because now there are two entities, and you can solve that game. And the solution to that game produces behavior that is beneficial to the human.

Let me just illustrate the kinds of things that fall out as solutions to this problem—what behavior do you get when you build machines that way? For example, asking permission. Let's say the AI has information, for example, that we would like a cup of coffee right now, but it doesn't know much about our price sensitivity. The only plan it can come up with, because we're in the Georges V in Paris, is to go ask for a cup of coffee that costs €13. The AI should come back and say, "Would you still like the coffee at €13, or would you prefer to wait another ten minutes and find another cafe to get a coffee that's cheaper?" That's, in a microcosm, one of the things it does—it asks permission.

It allows itself to be switched off because it wants to avoid doing anything that's harmful to us. But it knows that it doesn't know what constitutes harm. If there was any reason why the human would want to switch it off, then it's happy to be switched off because it wants to avoid doing whatever it is that the human is trying to prevent it from doing. That's the exact opposite of the machine with a fixed objective, which actually will take steps to prevent itself from being switched off because that would prevent it from achieving the objective.

When you solve this game, where the machine's half of the game is basically trying to be beneficial to the human, it will do things to learn more, and asking permission allows it to learn more about your preferences and it will allow itself to be switched off. It's basically deferential to the human. And it doesn't matter—unlike the case where you've got the fixed objective, which is wrong, the more intelligent you make the machine, the worse things get, the harder it is to switch it off, the more far-reaching the impact on the world is going to be. Whereas with this approach, the more intelligent the machine, the better off you are.

Because it will be better at learning your preferences. It will be better at satisfying them. And that's what we want. I believe that this is the core. I think there's lots of work still to do, but this is the core of a different approach to what AI should've been all along.

James Manyika is co-chairman of the McKinsey Global Institute and a senior partner in McKinsey's San Francisco office. **Stuart Russell** is a professor of computer science at the University of California, Berkeley, and author of the book *Human Compatible.*

Designed by Global Editorial Services