McKinsey Analytics

Why we need to rethink the purpose of AI: A conversation with Stuart Russell

Stuart Russell, one of the pioneering thinkers in the field of artificial intelligence, explains how to ensure that the technology benefits humanity.

May 2020

In this episode of the *McKinsey on AI* podcast mini-series, we share a conversation between James Manyika, co-chairman of the McKinsey Global Institute, and University of California, Berkeley, professor Stuart Russell. They explore how we can ensure AI truly benefits humanity rather than causing us harm. According to Russell, doing so begins with abandoning the idea of creating "intelligent" machines altogether.

Podcast transcript

David DeLallo: Artificial intelligence is poised to expand the realm of what's possible in every facet of our lives. Already it has transformed everyday activities, from banking to shopping to the way we interact with our phones. But as AI becomes more powerful, some worry about what could happen if these intelligent systems become, well, *too* intelligent. Will they begin to tell us humans what to do, rather than the other way around? I'm David DeLallo with McKinsey Publishing. Welcome to this edition of our McKinsey podcast series on AI.

In this episode, we'll be talking about this question of building AI systems that are as smart as or even smarter than we are, and how we can ensure that AI truly benefits humanity rather than causing us harm. To explore this topic, McKinsey Global Institute chairman James Manyika sat down with Stuart Russell, one of the world's foremost thought leaders on AI. Stuart is well known for coauthoring the seminal textbook on developing AI systems nearly three decades ago.

Today, Stuart is poised to guide the next generation of AI with his latest book, *Human Compatible*. Released last year, it's been called the most important book so far on AI and tackles the problem of control, as Stuart calls it, as machines become more intelligent and could potentially ignore our requests. To start, it's probably good for us to learn exactly how Stuart defines AI. As James points out early in their conversation, there's a lot of hype in the press and misunderstanding about exactly what AI is.

James Manyika: This is actually an important point, because I think normally these days, when you read the typical press, you would think AI equals deep learning. So how do you define artificial intelligence?

Stuart Russell: There are lots of parts of AI that actually don't rely on deep learning at all. It's still logic based. You can think of the whole database industry as a branch of logic-based AI, along with all the business rules or business-intelligence systems that effectively use logical rules on logical data.

Those systems run a big chunk of the economy and the web. So it's not as if that stuff died or disappeared or was wrong. It just found a niche where it's entirely applicable. There are other technologies—for example, probabilistic programming—which took the major network technology of the late '80s, early '90s, and essentially lifted it up to the next level.

Take, for example, the monitoring system for the Nuclear Test Ban Treaty. This is a system running in Vienna that is listening to the entire planet through very, very sensitive seismic detectors and trying to understand all the seismic activity on Earth and figure out what is "manmade," so to speak, such as clandestine nuclear explosions. It generates models with hundreds of thousands or millions of variables doing probabilistic reasoning in real time.

I'll give you the classical definition, the one that's in the textbook: AI is about building machines that do the right thing, that act in ways that can be expected to achieve their objectives. This covers learning systems, robotic systems, the game-playing systems, the natural-language systems—they can all be understood in this framework.

David DeLallo: That sounds pretty straightforward and logical: building machines that act in predictable ways to meet their objectives. But in actuality, Stuart says, it's this classical definition that may be leading us down the wrong path, one in which AI systems begin acting in *unpredictable* ways. More on that shortly. But first, let's get into what Stuart thinks about the potential for super-intelligent machines and where they could fit into the future of our world.

Stuart Russell: We want to be able to endow machines with intelligence at least comparable with—

or, in relevant respects, superior to—our own. It would be a tool that we could use to increase the power and capabilities of our civilization.

David DeLallo: But what exactly are super-intelligent machines? And how are they different from the AI systems organizations are building today, such as those that can predict which products we'll buy or that alert factory workers when a piece of equipment is about to fail before there are any visible indications?

James Manyika: One of the questions that I think is also at the center of this discussion is the idea of what a super intelligence or a general intelligence would look like. As you know, in the AI community, there's often this distinction made between AI and AGI—artificial general intelligence. Is that distinction useful?

David DeLallo: Interestingly, the difference, Stuart says, is more of an artificial distinction, if you will, than a real one.

Stuart Russell: AGI is a little bit of a marketing term, although it comes from the academic community. It's intended to actually mark off the group of people who think of themselves as working on the long-term goal, creating a human level of superhuman intelligence. Their story is that most people in AI are just working on narrow applications and spinoffs and have lost sight of the long-term goal. I actually don't think this is true at all.

David DeLallo: To make his case, he shares the story of an AT&T lab group that was trying to solve a fairly mundane business problem back in the 1990s.

Stuart Russell: They were working on recognizing handwritten digits for the US Postal Service and for banks, so that they could recognize handwritten checks. It couldn't get much more narrow and boring and tedious than that.

David DeLallo: It was this very ordinary goal that led to an important advance in AI: the development of convolutional neural networks. These networks, or CNNs, as they're often called, are a type of deep learning model that enables us to infer information from unstructured data sets, such as images. Convolutional neural networks make it possible, for example, for AI systems to diagnose diseases from health scans or to detect a product defect on a production line through images.

Stuart Russell: So really, there isn't a whole lot of evidence that narrow AI actually exists. Yes, we build AI systems for particular applications, but in order to make them work, we tend to develop technology that has lots of other applications. The process of moving AI forward is, first of all, understanding the limits of what that technology can and can't do. And then, can we remove those limits one by one until they're all gone?

David DeLallo: So what limitations do we need to remove to get there? An important one Stuart shares is the inability of AI systems to create what he calls abstractions, which bring together existing ideas and create entirely new things.

Stuart Russell: So I didn't invent the idea of taking the Metro, right, but it's there. Civilization created it as an operation, which I can then combine into more complex operations. I didn't invent the PhD, but I could choose to get one because it existed as a step. Our civilization over centuries has produced layer upon layer upon layer of these abstractions, which we then have, like, a library. We're taught what they are, and then we can put them together in new ways to make new things. In recent years, we've added Ubering and Googling and emailing, which didn't exist before. Taking a flight to Australia used to be something almost impossible, and now it's just a thing. You just do it.

James Manyika: So this idea of assembling things and achieving higher and higher levels of abstraction is a problem.

Stuart Russell: Right, so creating those new abstractions is one of the big open problems. We don't know how to [enable AI to] do that.

David DeLallo: So it seems we're well on the way to creating these super-intelligent machines, but how long will it be until they're truly a reality?

Stuart Russell: Most people think that super-intelligent AI is going to arrive sometime this century.

David DeLallo: And in fact, it's likely to happen faster than we think.

Stuart Russell: If you look at AlphaGo, for example—it's a system built by DeepMind that beat the best Go players in the world. Go is a game that's considered to be much more complex than chess. It certainly has a bigger state space, with many more possible legal moves at any given point. So even after the human world chess champion was beaten back in '97 [by a machine], people predicted it would take another 100 years before machines could beat the world Go champion. But that happened less than 20 years later, in 2016, using machines that were, I think, almost a billion times faster. So a lot more computational power—and a lot more training.

David DeLallo: So we know super intelligence is possible, and we're well on our way to solving the technology barriers to creating machines that are as smart as or even smarter than we are. How then do we ensure we stay in control and avoid a scenario where robots can and do take over? Some have suggested the only or best solution is abandoning the development of super-intelligent systems altogether. After all, they say, this is a foolproof way to ensure these systems don't take over the world. But that's not the way to go, Stuart says.

Stuart Russell: You would lose the golden-age benefits. All the upside would disappear. You have to understand *why* we lose control. That was sort of the genesis of the new book—thinking about why we lose control.

David DeLallo: Remember earlier when Stuart shared the classical definition of AI as "building machines that do the right thing to meet their objectives"? This idea of meeting certain objectives gets a bit thorny, and it's basically the crux of the problem, Stuart says.

Stuart Russell: It's a bad model, because it's only of benefit to us if we state the objective completely and correctly. It turns out that that's not possible in general. In the lab, with not very bright computers, what typically happens is you state the objective, you see the behavior, you don't like it, and you say, OK, I guess I got it wrong. We've actually known this for thousands of years: 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 he dies in his room of starvation. You know, 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." Even if the machine understands the full extent of our preferences, which I think is impossible, because *we* don't understand them, we don't know how we're going to feel about some future experience.

David DeLallo: So halting AI development isn't a good idea, nor is it likely possible. But continuing on the current path could lead to some significant problems. What's the solution? Stuart believes we can avoid disaster by shifting our focus from building intelligent systems to building beneficial ones, which operate under three principles.

Stuart Russell: The key characteristics—which I express in the book as three principles of just entirely coincidental resemblance to Asimov's laws—are first, being of benefit to the humans is the only objective for machines. But the second principle is that the machine does not know what that means. It does not know our preferences for how the future should unfold, and that turns out to be crucial. It knows that it doesn't know the objective. The third principle is essentially what enables it to learn more about the objective, that our choices, our behavior reveals information about our underlying preferences. There are probably other ways you could do it, with an fMRI machine, telepathy, or

something, some way of getting at underlying preferences. But for the foreseeable future, it's based on the choices we make. So the third principle says that preferences produce behavior, and so, by observing behavior, we can infer something about underlying preferences.

This is probably where things get complicated, because the process by which we produce behavior is not perfect. We often do things that we later realize weren't exactly right, including, like Lee Sedol, the famous Go player, playing a losing move in his match against AlphaGo. He realized afterwards that it was a losing move. It wasn't that he was trying to lose. It's just that his cognitive processes did not enable him to play perfectly.

James Manyika: Well, let me ask this. Of your three principles, the one that seems to be the big leap is the second one, which is the presumption that, in fact, our preferences are never fully knowable. Because if you don't have that, then the whole premise falls apart. And you count on the fact that we are inherently unknowable?

David DeLallo: For Stuart, it's not that we are inherently unknowable. Certainly, AI systems can and do learn our preferences at any given time, but that's the key: they learn our preference at *that moment*. We're human, after all, and our preferences can and often do change.

Stuart Russell: They'd never be stable long enough for the machine to learn what they are. Obviously, there are billions of us. We all have different preferences. So what we're actually doing is we're saying, OK, instead of writing algorithms that find optimal solutions for a fixed objective, we can write algorithms that solve this problem of functioning as sort of one half of a combined system with humans. So this actually makes it a game-theory problem, because now there are two entities. So when you solve this kind of problem, where the machine's half of the game is to try to be beneficial to the human, it will do things to learn more, so asking permission allows it to learn more about your preferences.

We simply want the machine to learn what each of the eight billion people on Earth would like the future to be like. Now that's quite feasible, in the sense that, as you know, Facebook already has personal profiles for about a couple of billion individuals. So it's not sci-fi that we could have models for every human.

David DeLallo: This idea of having AI systems ask our permission is critical, according to Stuart. In essence, it's a built-in shut-off switch that ensures we can turn these systems off at any time. To illustrate this, he uses the example of simply finding and buying a cup of coffee in Paris.

Stuart Russell: Let's say it [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. So the only plan it can come up with, because we're in the George V in Paris, is to go and ask for a cup of coffee. And it's €13. It should come back and say, "Would you still like the coffee at €13? Or if you wait another ten minutes, I can go around the corner and find a cafe or a Starbucks and get something cheaper." So if there was any reason why the human wants to switch it off, then it's happy to be switched off, because it doesn't want to do whatever it is that the human is trying to prevent it from doing. That's the exact opposite of a 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.

David DeLallo: So how do we make that shift, James goes on to ask. Stuart believes the answer lies in educating those who develop AI systems to rethink the use of these fixed objectives.

Stuart Russell: One way is we write a new edition of the textbook, which is what I'm doing right now. We have some examples in some of the chapters of how to do things this other way.

David DeLallo: Another way, he suggests, is that we start creating demonstration systems so data scientists can see exactly how these concepts play out in the real world. Basically, these demos

would provide an alternative system to what we're doing now. Content selection systems, which have been under the microscope for helping to fuel distribution of controversial content and perpetuating negative stereotypes, are a prime candidate.

Stuart Russell: An alternative kind of social media content selection that is sensitive to possible negative consequences or, should we say, consequences on parts of the world whose values you're not sure about, at least don't do it without asking first.

David DeLallo: He says self-driving cars also top the list for the demos to start with.

Stuart Russell: Self-driving cars, as they come out, have a relatively narrow range of things they can do. But you can still have preferences for, for example, how fast you want to go. Maybe a little over the speed limit, a lot over the speed limit? Do you want to keep changing lanes? Do you want a nice, steady ride? How far away from the terminal can I drop you off if there's a big queue of traffic waiting to get there? You know, all these kinds of questions. You want a difference between the standard model and the new model. The standard model is like the genie in the lamp. You get exactly the objective you put in, and you always regret it. The new model would be more like the perfect butler, who understands what you want, what you might not want, and knows when to ask and when to defer to what your preferences might be. I think that's, in a nutshell, where we want to go.

David DeLallo: It's certainly a fascinating way to think about AI and how we can get the most and best results out of this truly amazing capability. And with that, we conclude this edition of our podcast series. Many thanks to James and Stuart for letting us listen in on their conversation. And thank you, listeners, for joining. Please do check out some of our additional podcasts on this and other McKinsey channels. Bye for now.

David DeLallo is an executive editor in McKinsey Publishing, based in McKinsey's Stamford office. **James Manyika** is co-chairman of the McKinsey Global Institute and a senior partner in the San Francisco office. **Stuart Russell** is a professor of computer science at the University of California, Berkeley, author of the book *Human Compatible*, and co-author of *Artificial Intelligence: A Modern Approach*, now in its fourth edition.