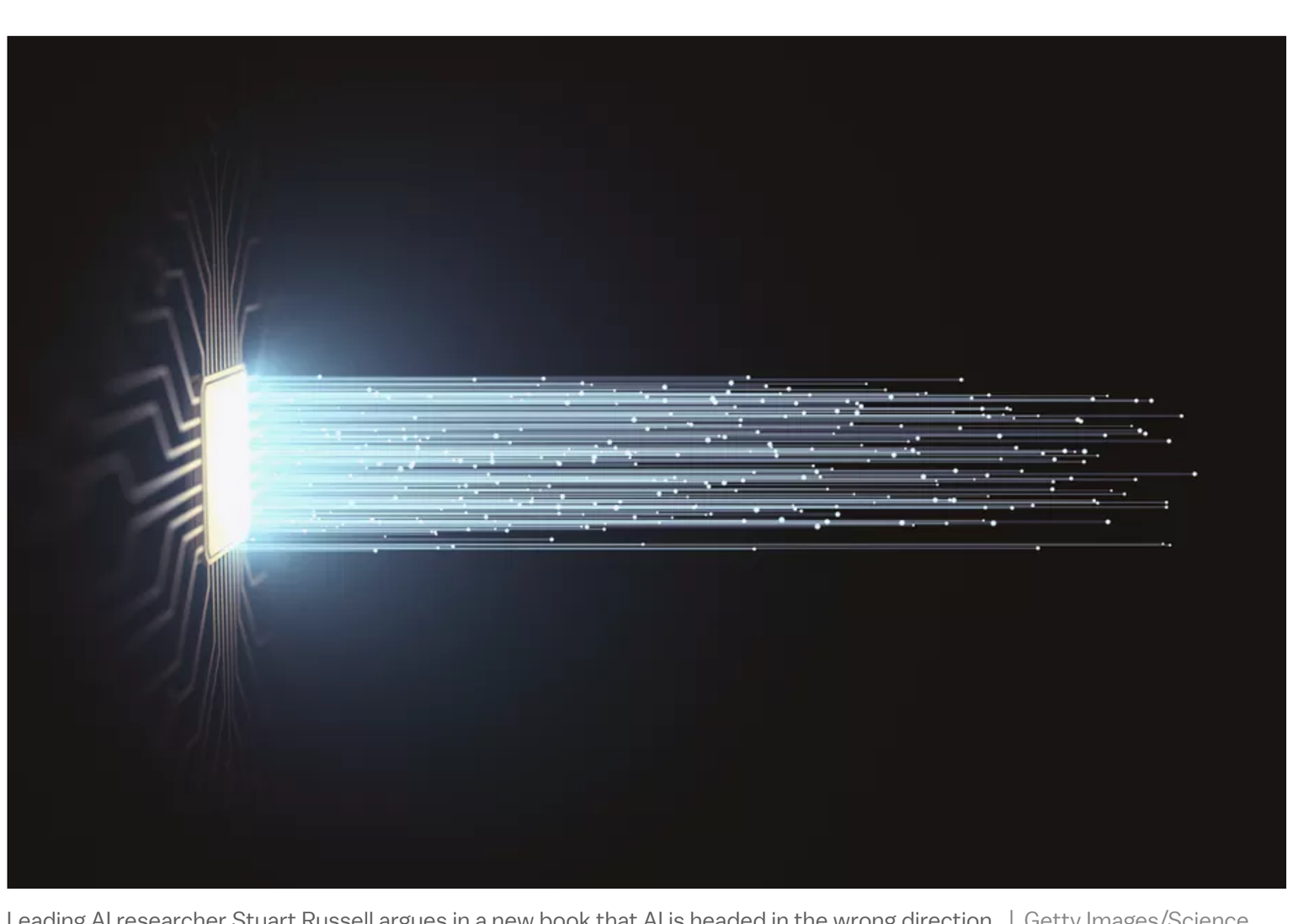
AI could be a disaster for humanity. A top computer scientist thinks he has the solution.

Stuart Russell wrote the book on AI and is leading the fight to change how we build it.

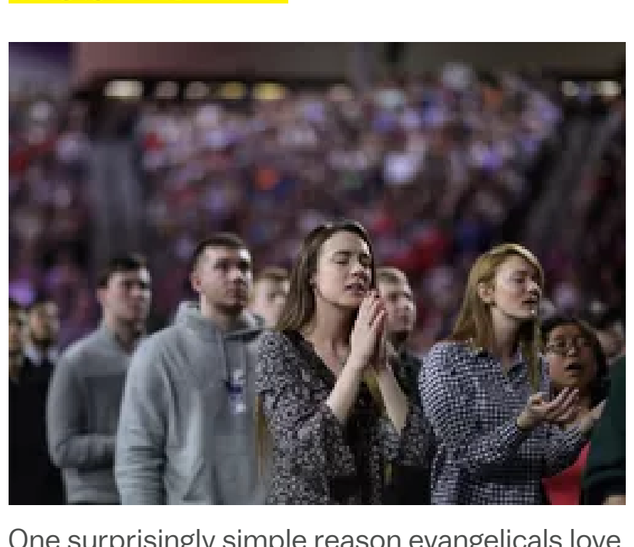
By Kelsey Piper | Oct 26, 2019, 8:00am EDT

SHARE

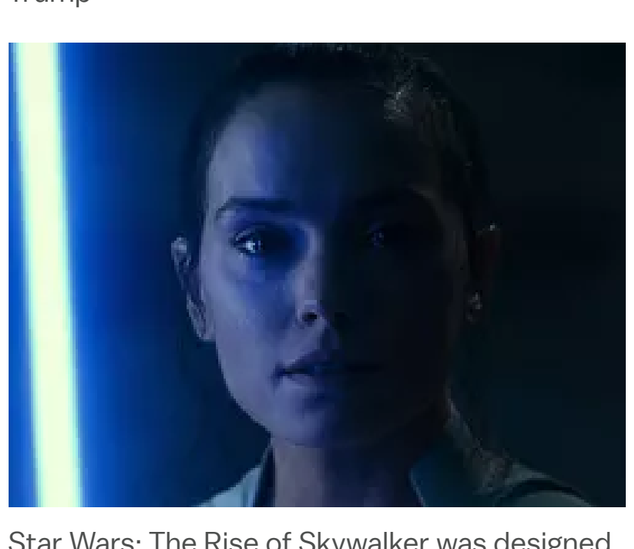


Leading AI researcher Stuart Russell argues in a new book that AI is headed in the wrong direction. | Getty Images/Science Photo Libra

MOST READ



One surprisingly simple reason evangelicals love Trump



Star Wars: The Rise of Skywalker was designed to be the opposite of The Last Jedi

Queen Elizabeth gave her annual Christmas speech. The internet saw a secret message about Brexit.

Muslims love Jesus, too: 6 things you didn't know about Jesus in Islam

Paul Krugman on climate, robots, single-payer, and so much more

FUTURE PERFECT — *Finding the best ways to do good.*

Stuart Russell is a leading AI researcher who **literally wrote (well, co-authored) the top textbook** on the topic. He has also, for the last several years, been warning that his field has the potential to go **catastrophically wrong**.

In a new book, ***Human Compatible***, he explains how. AI systems, he notes, are evaluated by how good they are at achieving their objective: winning video games, writing humanlike text, solving puzzles. If they hit on a strategy that fits that objective, they will run with it, without explicit human instruction to do so.

But with this approach, we've set ourselves up for failure because the "objective" we've given the AI system is not the only thing we care about. Imagine a self-driving car with an "objective" to get from Point A to Point B but unaware that we also care about the survival of the passengers and of pedestrians along the way. Or a health care cost-saving system that **discriminates against black patients** because it anticipates that they're less likely to seek the health care they need.

Humans care about a lot of things: fairness, law, democratic input, our safety and flourishing, our freedom. AI systems, Russell argues in *Human Compatible*, care about only whatever we've put in as their objective. And that means there's a disaster on the horizon.

I met Russell at UC Berkeley, where he heads the **Center for Human-Compatible AI**, to talk about his book and about the risks posed by advanced artificial intelligence. Here's a transcript of our conversation, edited for length and clarity.

Kelsey Piper

What's the case that advanced AI could be dangerous for humanity?

Stuart Russell

To answer that question, we have to understand: how are AI systems designed? What do they do? And in the Standard Model [of AI systems] you build machinery, algorithms, and so on that are designed to achieve specific objectives that you put into the program.

So if it's a chess program, you give it the goal of beating your opponent, of winning the game. If it's a self-driving car, the passenger puts in the objective: [for instance,] I want to be at the airport.

So that all sounds fine. The problem comes when systems become more intelligent. If you put in the wrong objective, then the system pursuing it may take actions that you are extremely unhappy about.

We call this the King Midas problem. King Midas specified his objective: I want everything I touch turned to gold. He got exactly what he asked for. Unfortunately, that included his food and his drink and his family members, and he dies in misery and starvation. Many cultures have the same story. The genie grants you three wishes. Always the third wish is "please undo the first two wishes" because I ruined the world.

And unfortunately, with systems that are more intelligent and therefore more powerful than we are, you don't necessarily get a second and third wish.

So the problem comes from increasing capabilities, coupled with our inability to specify objectives completely and correctly. Can we restore our carbon dioxide to historical levels so that we get the climate back in balance? Sounds like a great objective. Well, the easiest way to do that is to get rid of all those things that are producing carbon dioxide, which happen to be humans. You want to cure cancer as quickly as possible. Sounds great, right? But the quickest way to do it is to run medical trials in parallel with millions of human subjects or billions of human subjects. So you give everyone cancer and then you see what treatments work.

Kelsey Piper

We can't just write down all of the things we don't mean? Don't break any laws, don't murder anybody …

Stuart Russell

So, we've been trying to write tax law for 6,000 years. And yet, humans come up with loopholes and ways around the tax laws so that, for example, our multinational corporations are paying very little tax to most of the countries that they operate in. They find loopholes. And this is what, in the book, I call the loophole principle. It doesn't matter how hard you try to put fences and rules around the behavior of the system. If it's more intelligent than you are, it finds a way to do what it wants.

Kelsey Piper

Human Compatible describes this problem. We're putting incorrect objectives into these systems. The systems try and complete their objectives but their objectives don't encompass everything we care about. What's the solution?

Stuart Russell

If you continue on the current path, the better AI gets, the worse things get for us. For any given incorrectly stated objective, the better a system achieves that objective, the worse it is.

The approach that we propose in the second half of the book is that we design AI systems in a completely different way. We stop using the standard model, which requires us to specify a fixed objective. Instead, the AI system has a constitutional requirement that it be of benefit to human beings.

But it knows that it *doesn't* know what that means. It doesn't know our preferences. And it knows that it doesn't know our preferences about how the future should unfold.

So you get totally different behavior. Basically, the machines defer to humans. They ask permission before doing anything that messes with part of the world.

Kelsey Piper

And they don't have incentives to deceive us about the effects of a course of action?

Stuart Russell

That's another layer of complication where the standard model goes wrong.

A system that is pursuing an objective that's fixed observes human behavior and anticipates that the human might try to interfere with this. Rather than say, "Oh, yeah, please switch me off or change the objective," [this AI] will actually pretend to be doing what humans like simply to prevent us from interfering long enough until it has enough power that it can achieve the objective despite human interference. So you're giving it an incentive to deceive us about its abilities, about its plans. And this is clearly not what we want.

[An AI that is trying to learn what humans want] has an incentive to be honest about its plans because it wants to get feedback and so on.

Kelsey Piper

You've been a leading AI researcher for decades. I'm curious at what point you became convinced that AI is dangerous.

Stuart Russell

So for a long time I've been uncomfortably aware that we don't have an answer to the question: "What if you succeed?" In fact, the first edition of [my] textbook has a section with that title, because it's a pretty important question to ask if a whole field is pushing towards a goal. And if it looks like, when you get there, that you may be taking the human race off a cliff, then that's a problem.

If you ask, okay, we're gonna make things that are much more intelligent, much more powerful than us. How on earth do we expect us to [keep] power from more powerful [entities] forever? It's not obvious that that question has an answer.

In fact, [computer scientist Alan] Turing said we would have to expect the machines to take control. He was completely resigned to this and our species would be humbled, as he put it. So that's clearly a disturbing state of affairs.

It was more clear to me starting in the early 2010s. I was on sabbatical in Paris. I had more time to appreciate the importance of human experience and civilization. And in the meantime, other researchers, mostly outside the field, had started to point out these failure modes: that fixed objectives led to all of these unwelcome behaviors, deception and potentially arbitrarily bad consequences from resource consumption, from self-defense incentives.

So the confluence of those things led me to start thinking about, okay, how do we actually fix the problem?

Kelsey Piper

I read some criticisms and responses to *Human Compatible*. One thing you hear is "worrying about AI now is like people in the 1700s worrying about how to stop the space shuttles from blowing up." Since we don't know what general AI will be like, we can't possibly think about how to design it safely.

Stuart Russell

I think it's useful to look back at the history of nuclear energy and nuclear physics, because it has many parallels. No, it's not a perfect analogy. But when Leo Szilard invented the nuclear chain reaction, he didn't know which atoms could be induced to go through a fission reaction and produce neutrons that would then produce more fission reactions.

He said, "Okay, this is a possible way by which a chain reaction could occur."

And he was able to design a nuclear reactor *just on that basis*, including the feedback control mechanisms that would maintain the reaction at the subcritical level so that it didn't explode. We had a plan without knowing that any such reaction even existed.

So you can talk about the general structure and design of systems without understanding how to make all the parts work the way you want. One thing that's clear about general AI systems [is] they're going to be more intelligent than the ones we have right now. And the point about the standard model [of AI objectives] is that the more intelligent the system, the worse things get.

Kelsey Piper

Another line of criticism I've seen — I think this is something that Yann LeCun at Facebook has **expressed** — is we just don't need to worry that the systems won't do what we want them to do. We won't build systems like that.

Stuart Russell

That's like arguing, "Well, of course, we would never build nuclear reactors that blow up so we don't need to worry about nuclear safety." Right? That's ridiculous. In the book, I say it's like being on the scene of an accident and saying nobody should call an ambulance because *somebody* is going to call an ambulance.

The only way you get nuclear safety is by worrying about the ways [reactors] can blow up and preventing them from blowing up.

An interesting argument, which I discussed a little bit in the book, is that you can think of corporations as, in a sense, machines. They're effectively machines that are set up to maximize a prescribed objective, namely quarterly profit. You could look at the fossil fuel industry as a super-intelligent machine that actually, in the pursuit of its objective, outwitted the human race. So they have they created a 50-year sort of political subversion, public relations disinformation campaign so that they could continue pumping out carbon dioxide.

There are already these quasi-machine super-intelligent entities that are causing problems precisely because they're pursuing incorrect objectives, and it's clearly not the case that it *of course* works out.

Yann LeCun makes other arguments, as does Steven Pinker [another AI risk skeptic]. [One argument is] that it is a mistake to think that we would put in the objectives of world domination, the objectives of self-defense, self-preservation. There's no reason to do that. And as long as we don't, then nothing bad can happen.

And that, I think, is just misconstruing or misunderstanding one of the basic arguments in this whole debate, which is that you don't *have* to put those objectives in. They are subgoals of pursuing pretty much any fixed objective.

Kelsey Piper

What are the biggest misconceptions about the book or about your work that you've seen?

Stuart Russell

There's a general misconception about AI — which is promulgated by Hollywood for reasons of having interesting plots and by the media, because they seem to want to put pictures of *Terminator* robots on every article — which is that the thing we need to be concerned about is consciousness, that somehow these machines will accidentally become conscious and then they'll hate everybody and try to kill us.

And that's just a total red herring. The thing that we're concerned about here is competent, effective behavior in the world. If machines out-decide us, out-think us in the real world, we have to figure out how do we make sure that they're only ever acting on our behalf and not acting contrary to our interests.

Sign up for the Future Perfect newsletter. *Twice a week, you'll get a roundup of ideas and solutions for tackling our biggest challenges: improving public health, decreasing human and animal suffering, easing catastrophic risks, and — to put it simply — getting better at doing good.*

NEXT UP IN FUTURE PERFECT

- Are we living in a computer simulation? I don't know. Probably.
- Want to help animals? Here's where to donate your money.
- Supertrees: Meet the Amazonian giant that helps the rainforest make its own rain
- The 2010s featured a lot of great social science. Here are my 12 favorite studies.
- Thousands arrested protesting a new Indian citizenship law that excludes Muslims
- Trump might help free science that's locked behind paywalls

THE LATEST

All those hacks got Amazon's Ring sued
By Sara Morrison

Mike Bloomberg has spent a staggering $100 million on campaign ads in a month
By Zeeshan Aleem

The power of Greta Gerwig's Little Women is that it doesn't pretend its marriages are romantic
By Constance Grady

The telling conservative backlash to a Virginia zoning reform proposal, explained
By Matthew Yglesias

Congress quietly gave 4,000 Liberians a path to citizenship
By Nicole Narea

Israeli Prime Minister Benjamin Netanyahu wins party primary, weeks after indictments
By Zeeshan Aleem