Quanta magazine

Artificial Intelligence Will Do What We Ask. That's a Problem.

By Natalie Wolchover

January 30, 2020

By teaching machines to understand our true desires, one scientist hopes to avoid the potentially disastrous consequences of having them do what we command.

Powerful artificial intelligence is like the proverbial genie in a bottle. A seemingly innocent wish — "make my home eco-friendly" — can lead to unintended consequences.

Corinne Reid for Quanta Magazine

The danger of having artificially intelligent machines do our bidding is that we might not be careful enough about what we wish for. The lines of code that animate these machines will inevitably lack nuance, forget to spell out caveats, and end up giving AI systems goals and incentives that don't align with our true preferences.

A now-classic thought experiment illustrating this problem was posed by the Oxford philosopher Nick Bostrom in 2003. Bostrom imagined a superintelligent robot, programmed with the seemingly innocuous goal of manufacturing paper clips. The robot eventually turns the whole world into a giant paper clip factory.

Such a scenario can be dismissed as academic, a worry that might arise in some far-off future. But misaligned AI has become an issue far sooner than expected.

The most alarming example is one that affects billions of people. YouTube, aiming to maximize viewing time, deploys AI-based content recommendation algorithms. Two years ago, computer scientists and users began noticing that YouTube's algorithm seemed to achieve its goal by recommending increasingly extreme and conspiratorial content. One researcher reported that after she viewed footage of Donald Trump campaign rallies, YouTube next offered her videos featuring "white supremacist rants, Holocaust denials and other disturbing content." The algorithm's upping-the-ante approach went beyond politics, she said: "Videos about vegetarianism led to videos about veganism. Videos about jogging led to videos about running ultramarathons." As a result, research suggests, YouTube's algorithm has been helping to polarize and radicalize people and spread misinformation, just to keep us watching. "If I were planning things out, I probably would not have made that the first test case of how we're going to roll out this technology at a massive scale," said Dylan Hadfield-Menell, an AI researcher at the University of California, Berkeley.

YouTube's engineers probably didn't intend to radicalize humanity. But coders can't possibly think of everything. "The current way we do AI puts a lot of burden on the designers to understand what the consequences of the incentives they give their systems are," said Hadfield-Menell. "And one of the things we're learning is that a lot of engineers have made mistakes."

A major aspect of the problem is that humans often don't know what goals to give our AI systems, because we don't know what we really want. "If you ask anyone on the street, 'What do you want your autonomous car to do?' they would say, 'Collision avoidance,'" said Dorsa Sadigh, an AI scientist at Stanford University who specializes in human-robot interaction. "But you realize that's not just it; there are a bunch of preferences that people have." Super safe self-driving cars go too slow and brake so often that they make passengers sick. When programmers try to list all goals and preferences that a robotic car should simultaneously juggle, the list inevitably ends up incomplete. Sadigh said that when driving in San Francisco, she has often gotten stuck behind a self-driving car that's stalled in the street. It's safely avoiding contact with a moving object, the way its programmers told it to — but the object is something like a plastic bag blowing in the wind.

To avoid these pitfalls and potentially solve the AI alignment problem, researchers have begun to develop an entirely new method of programming beneficial machines. The approach is most closely associated with the ideas and research of Stuart Russell, a decorated computer scientist at Berkeley. Russell, 57, did pioneering work on rationality, decision-making and machine learning in the 1980s and '90s and is the lead author of the widely used textbook *Artificial Intelligence: A Modern Approach*. In the past five years, he has become an influential voice on the alignment problem and a ubiquitous figure — a well-spoken, reserved British one in a black suit — at international meetings and panels on the risks and long-term governance of AI.

Stuart Russell, a computer scientist at the University of California, Berkeley, gave a TED talk on the dangers of AI in 2017.

Bret Hartman / TED

As Russell sees it, today's goal-oriented AI is ultimately limited, for all its success at accomplishing specific tasks like beating us at *Jeopardy!* and Go, identifying objects in images and words in speech, and even composing music and prose. Asking a machine to optimize a "reward function" — a meticulous description of some combination of goals — will inevitably lead to misaligned AI, Russell argues, because it's impossible to include and correctly weight all goals, subgoals, exceptions and caveats in the reward function, or even know what the right ones are. Giving goals to free-roaming, "autonomous" robots will be increasingly risky as they become more intelligent, because the robots will be ruthless in pursuit of their reward function and will try to stop us from switching them off.

Instead of machines pursuing goals of their own, the new thinking goes, they should seek to satisfy human preferences; their only goal should be to learn more about what our preferences are. Russell contends that uncertainty about our preferences and the need to look to us for guidance will keep AI systems safe. In his recent book, *Human Compatible*, Russell lays out his thesis in the form of three "principles of beneficial machines," echoing Isaac Asimov's three laws of robotics from 1942, but with less naivete. Russell's version states:

1. The machine's only objective is to maximize the realization of human preferences.
2. The machine is initially uncertain about what those preferences are.
3. The ultimate source of information about human preferences is human behavior.

Over the last few years, Russell and his team at Berkeley, along with like-minded groups at Stanford, the University of Texas and elsewhere, have been developing innovative ways to clue AI systems in to our preferences, without ever having to specify those preferences.

These labs are teaching robots how to learn the preferences of humans who never articulated them and perhaps aren't even sure what they want. The robots can learn our desires by watching imperfect demonstrations and can even invent new behaviors that help resolve human ambiguity. (At four-way stop signs, for example, self-driving cars developed the habit of backing up a bit to signal to human drivers to go ahead.) These results suggest that AI might be surprisingly good at inferring our mindsets and preferences, even as we learn them on the fly.

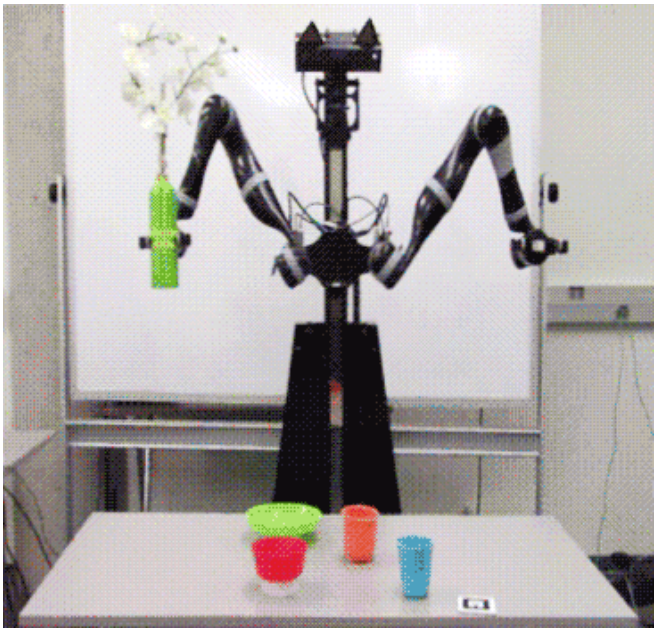
"These are first attempts at formalizing the problem," said Sadigh. "It's just recently that people are realizing we need to look at human-robot interaction more carefully."

Whether the nascent efforts and Russell's three principles of beneficial machines really herald a bright future for AI remains to be seen. The approach pins the success of robots on their ability to understand what humans really, truly prefer — something that the species has been trying to figure out for some time. At a minimum, Paul Christiano, an alignment researcher at OpenAI, said Russell and his team have greatly clarified the problem and helped "spec out what the desired behavior is like — what it is that we're aiming at."

How to Understand a Human

Russell's thesis came to him as an epiphany, that sublime act of intelligence. It was 2014 and he was in Paris on sabbatical from Berkeley, heading to rehearsal for a choir he had joined as a tenor. "Because I'm not a very good musician, I was always having to learn my music on the metro on the way to rehearsal," he recalled recently. Samuel Barber's 1967 choral arrangement *Agnus Dei* filled his headphones as he shot beneath the City of Light. "It was such a beautiful piece of music," he said. "It just sprang into my mind that what matters, and therefore what the purpose of AI was, was in some sense the aggregate quality of human experience."

Robots shouldn't try to achieve goals like maximizing viewing time or paper clips, he realized; they should simply try to improve our lives. There was just one question: "If the obligation of machines is to try to optimize that aggregate quality of human experience, how on earth would they know what that was?"



In Scott Niekum's lab at the University of Texas, Austin a robot named Gemini learns how to place a vase of flowers in the center of a table. A single human demonstration is ambiguous, since the intent might have been to place the vase to right of the green plate, or left of the red bowl. However, after asking a few queries, the robot performs well in test cases.

Scott Niekum

The roots of Russell's thinking went back much further. He has studied AI since his school days in London in the 1970s, when he programmed tic-tac-toe and chess-playing algorithms on a nearby college's computer. Later, after moving to the AI-friendly Bay Area, he began theorizing about rational decision-making. He soon concluded that it's impossible. Humans aren't even remotely rational, because it's not computationally feasible to be: We can't possibly calculate which action at any given moment will lead to the best outcome trillions of actions later in our long-term future; neither can an AI. Russell theorized that our decision-making is hierarchical — we crudely approximate rationality by pursuing vague long-term goals via medium-term goals while giving the most attention to our immediate circumstances. Robotic agents would need to do something similar, he thought, or at the very least understand how we operate.

Russell's Paris epiphany came during a pivotal time in the field of artificial intelligence. Months earlier, an artificial neural network using a well-known approach called reinforcement learning shocked scientists by quickly learning from scratch how to play and beat Atari video games, even innovating new tricks along the way. In reinforcement learning, an AI learns to optimize its reward function, such as its score in a game; as it tries out various behaviors, the ones that increase the reward function get reinforced and are more likely to occur in the future.

Russell had developed the inverse of this approach back in 1998, work he continued to refine with his collaborator Andrew Ng. An "inverse reinforcement learning" system doesn't try to optimize an encoded reward function, as in reinforcement learning; instead, it tries to learn what reward function a human is optimizing. Whereas a reinforcement learning system figures out the best actions to take to achieve a goal, an inverse reinforcement learning system deciphers the underlying goal when given a set of actions.

A few months after his *Agnus Dei*-inspired epiphany, Russell got to talking about inverse reinforcement learning with Nick Bostrom, of paper clip fame, at a meeting about AI governance at the German foreign ministry. "That was where the two things came together," Russell said. On the metro, he had understood that machines should strive to optimize the aggregate quality of human experience. Now, he realized that if they're uncertain about how to do that — if computers don't know what humans prefer — "they could do some kind of inverse reinforcement learning to learn more."

With standard inverse reinforcement learning, a machine tries to learn a reward function that a human is pursuing. But in real life, we might be willing to actively help it learn about us. Back at Berkeley after his sabbatical, Russell began working with his collaborators to develop a new kind of "cooperative inverse reinforcement learning" where a robot and a human can work together to learn the human's true preferences in various "assistance games" — abstract scenarios representing real-world, partial-knowledge situations.

One game they developed, known as the off-switch game, addresses one of the most obvious ways autonomous robots can become misaligned from our true preferences: by disabling their own off switches. Alan Turing suggested in a BBC radio lecture in 1951 (the year after he published a pioneering paper on AI) that it might be possible to "keep the machines in a subservient position, for instance by turning off the power at strategic moments." Researchers now find that simplistic. What's to stop an intelligent agent from disabling its own off switch, or, more generally, ignoring commands to stop increasing its reward function? In *Human Compatible*, Russell writes that the off-switch problem is "the core of the problem of control for intelligent systems. If we cannot switch a machine off because it won't let us, we're really in trouble. If we can, then we may be able to control it in other ways too."

Dorsa Sadigh, a computer scientist at Stanford University, teaches a robot the preferred way to pick up various objects.

Drew Kelly for the Stanford Institute for Human-Centered Artificial Intelligence

Uncertainty about our preferences may be key, as demonstrated by the off-switch game, a formal model of the problem involving Harriet the human and Robbie the robot. Robbie is deciding whether to act on Harriet's behalf — whether to book her a nice but expensive hotel room, say — but is uncertain about what she'll prefer. Robbie estimates that the payoff for Harriet could be anywhere in the range of −40 to +60, with an average of +10 (Robbie thinks she'll probably like the fancy room but isn't sure). Doing nothing has a payoff of 0. But there's a third option: Robbie can query Harriet about whether she wants it to proceed or prefers to "switch it off" — that is, take Robbie out of the hotel-booking decision. If she lets the robot proceed, the average expected payoff to Harriet becomes greater than +10. So Robbie will decide to consult Harriet and, if she so desires, let her switch it off.

Russell and his collaborators proved that in general, unless Robbie is completely certain about what Harriet herself would do, it will prefer to let her decide. "It turns out that uncertainty about the objective is essential for ensuring that we can switch the machine off," Russell wrote in *Human Compatible*, "even when it's more intelligent than us."

These and other partial-knowledge scenarios were developed as abstract games, but Scott Niekum's lab at the University of Texas, Austin is running preference-learning algorithms on actual robots. When Gemini, the lab's two-armed robot, watches a human place a fork to the left of a plate in a table-setting demonstration, initially it can't tell whether forks always go to the left of plates,

or always on that particular spot on the table; new algorithms allow Gemini to learn the pattern after a few demonstrations. Niekum focuses on getting AI systems to quantify their own uncertainty about a human's preferences, enabling the robot to gauge when it knows enough to safely act. "We are reasoning very directly about distributions of goals in the person's head that could be true," he said. "And we're reasoning about risk with respect to that distribution."

Recently, Niekum and his collaborators found an efficient algorithm that allows robots to learn to perform tasks far better than their human demonstrators. It can be computationally demanding for a robotic vehicle to learn driving maneuvers simply by watching demonstrations by human drivers. But Niekum and his colleagues found that they could improve and dramatically speed up learning by showing a robot demonstrations that have been ranked according to how well the human performed. "The agent can look at that ranking, and say, 'If that's the ranking, what explains the ranking?" Niekum said. "What's happening more often as the demonstrations get better, what happens less often?" The latest version of the learning algorithm, called Bayesian T-REX (for "trajectory-ranked reward extrapolation"), finds patterns in the ranked demos that reveal possible reward functions that humans might be optimizing for. The algorithm also gauges the relative likelihood of different reward functions. A robot running Bayesian T-REX can efficiently infer the most likely rules of place settings, or the objective of an Atari game, Niekum said, "even if it never saw the perfect demonstration."

Our Imperfect Choices

Russell's ideas are "making their way into the minds of the AI community," said Yoshua Bengio, the scientific director of Mila, a top AI research institute in Montreal. He said Russell's approach, where AI systems aim to reduce their own uncertainty about human preferences, can be achieved with deep learning — the powerful method behind the recent revolution in artificial intelligence, where the system sifts data through layers of an artificial neural network to find its patterns. "Of course more research work is needed to make that a reality," he said.

Russell sees two major challenges. "One is the fact that our behavior is so far from being rational that it could be very hard to reconstruct our true underlying preferences," he said. AI systems will need to reason about the hierarchy of long-term, medium-term and short-term goals — the myriad preferences and commitments we're each locked into. If robots are going to help us (and avoid making grave errors), they will need to know their way around the nebulous webs of our subconscious beliefs and unarticulated desires.

In the driving simulator at Stanford University's Center for Automotive Research, self-driving cars can learn the preferences of human drivers.

Rod Searcey

The second challenge is that human preferences change. Our minds change over the course of our lives, and they also change on a dime, depending on our mood or on altered circumstances that a robot might struggle to pick up on.

In addition, our actions don't always live up to our ideals. People can hold conflicting values simultaneously. Which should a robot optimize for? To avoid catering to our worst impulses (or worse still, amplifying those impulses, thereby making them easier to satisfy, as the YouTube algorithm did), robots could learn what Russell calls our meta-preferences: "preferences about what kinds of preference-change processes might be acceptable or unacceptable." How do we feel about our changes in feeling? It's all rather a lot for a poor robot to grasp.

Like the robots, we're also trying to figure out our preferences, both what they are and what we want them to be, and how to handle the ambiguities and contradictions. Like the best possible AI, we're also striving — at least some of us, some of the time — to understand the form of the good, as Plato called the object of knowledge. Like us, AI systems may be stuck forever asking questions — or waiting in the off position, too uncertain to help.

"I don't expect us to have a great understanding of what the good is anytime soon," said Christiano, "or ideal answers to any of the empirical questions we face. But I hope the AI systems we build can answer those questions as well as a human and be engaged in the same kinds of iterative process to improve those answers that humans are — at least on good days."

However, there's a third major issue that didn't make Russell's short list of concerns: What about the preferences of bad people? What's to stop a robot from working to satisfy its evil owner's nefarious ends? AI systems tend to find ways around prohibitions just as wealthy people find loopholes in tax laws, so simply forbidding them from committing crimes probably won't be successful.

Or, to get even darker: What if we all are kind of bad? YouTube has struggled to fix its recommendation algorithm, which is, after all, picking up on ubiquitous human impulses.

Still, Russell feels optimistic. Although more algorithms and game theory research are needed, he said his gut feeling is that harmful preferences could be successfully down-weighted by programmers — and that the same approach could even be useful "in the way we bring up children and educate people and so on." In other words, in teaching robots to be good, we might find a way to teach ourselves. He added, "I feel like this is an opportunity, perhaps, to lead things in the right direction."

This article was reprinted on TheAtlantic.com and Spektrum.de.