Concerns of an Artificial Intelligence Pioneer

The computer scientist Stuart Russell wants to ensure that our increasingly intelligent machines remain aligned with human values.

By Natalie Wolchover

Stuart Russell, a computer scientist at the University of California, Berkeley, during a March stopover in San Antonio, Texas.

In January, the British-American computer scientist Stuart Russell drafted and became the first signatory of an open letter calling for researchers to look beyond the goal of merely making artificial intelligence more powerful. “We recommend expanded research aimed at ensuring that increasingly capable AI systems are robust and beneficial,” the letter states. “Our AI systems must do what we want them to do.” Thousands of people have since signed the letter, including leading artificial intelligence researchers at Google, Facebook, Microsoft and other industry hubs along with top computer scientists, physicists and philosophers around the world. By the end of March, about 300 research groups had applied to pursue new research into “keeping artificial intelligence beneficial” with funds contributed by the letter’s 37th signatory, the inventor-entrepreneur Elon Musk.

Russell, 53, a professor of computer science and founder of the Center for Intelligent Systems at the University of California, Berkeley, has long been contemplating the power and perils of thinking machines. He is the author of more than 200 papers as well as the field's standard textbook, *Artificial Intelligence: A Modern Approach* (with Peter Norvig, head of research at Google). But increasingly rapid advances in artificial intelligence have given Russell's longstanding concerns heightened urgency.

Recently, he says, artificial intelligence has made major strides, partly on the strength of neuro-inspired learning algorithms. These are used in Facebook's face-recognition software, smartphone personal assistants and Google's self-driving cars. In a bombshell result reported recently in *Nature*, a simulated network of artificial neurons learned to play Atari video games better than humans in a matter of hours given only data representing the screen and the goal of increasing the score at the top — but no preprogrammed knowledge of aliens, bullets, left, right, up or down. "If your newborn baby did that you would think it was possessed," Russell said.

Quanta Magazine caught up with Russell over breakfast at the American Physical Society's 2015 March Meeting in San Antonio, Texas, where he touched down for less than 24 hours to give a standing-room-only lecture on the future of artificial intelligence. In this edited and condensed version of the interview, Russell discusses the nature of intelligence itself and the immense challenges of safely approximating it in machines.

QUANTA MAGAZINE: You think the goal of your field should be developing artificial intelligence that is "provably aligned" with human values. What does that mean?

STUART RUSSELL: It's a deliberately provocative statement, because it's putting together two things — "provably" and "human values" — that seem incompatible. It might be that human values will forever remain somewhat mysterious. But to the extent that our values are revealed in our behavior, you would hope to be able to prove that the machine will be able to "get" most of it. There might be some bits and pieces left in the corners that the machine doesn't understand or that we disagree on among ourselves. But as long as the machine has got the basics right, you should be able to show that it cannot be very harmful.

How do you go about doing that?

That's the question I'm working on right now: Where does a machine get hold of some approximation of the values that humans would like it to have? I think one answer is a technique called "inverse reinforcement learning." Ordinary reinforcement learning is a process where you are given rewards and punishments as you behave, and your goal is to figure out the behavior that will get you the most rewards. That's what the [Atari-playing] DQN system is doing; it is given the score of the game, and its goal is to make that score bigger. Inverse reinforcement learning is the other way around. You see the behavior, and you're trying to figure out what score that behavior is trying to maximize. For example, your domestic robot sees you crawl out of bed in the morning and grind up some brown round things in a very noisy machine and do some complicated thing with steam and hot water and milk and so on, and then you seem to be happy. It should learn that part of the human value function in the morning is having some coffee.

There's an enormous amount of information out there in books, movies and on the web about human actions and attitudes to the actions. So that's an incredible resource for machines to learn what human values are — who wins medals, who goes to jail, and why.

Video: DQN, an artificial neural network developed by researchers at Google DeepMind, teaches itself to play Atari games such as Breakout. It quickly develops sophisticated strategies.

How did you get into artificial intelligence?

When I was in school, AI wasn't thought of as an academic discipline, by and large. But I was in boarding school in London, at St. Paul's, and I had the opportunity to avoid compulsory rugby by doing a computer science A-level [course] at a nearby college. One of my projects for A-level was a program that taught itself to play naughts and crosses, or tic-tac-toe. I became very unpopular because I used up the college's computer for hours on end. The next year I wrote a chess program and got permission from one of the professors at Imperial College to use their giant mainframe computer. It was fascinating to try to figure out how to get it to play chess. I learned some of the stuff I would later be teaching in my book.

But still, this was just a hobby; at the time my academic interest was physics. I did physics at Oxford. And then when I was applying to grad school I applied to do theoretical physics at Oxford and Cambridge, and I applied to do computer science at MIT, Carnegie Mellon and Stanford, not realizing that I'd missed all the deadlines for applications to the U.S. Fortunately Stanford waived the deadline, so I went to Stanford.

And you've been on the West Coast ever since?

Yep.

You've spent much of your career trying to understand what intelligence is as a prerequisite for understanding how machines might achieve it. What have you learned?

During my thesis research in the '80s, I started thinking about rational decision-making and the problem that it's actually impossible. If you were rational you would think: Here's my current state, here are the actions I could do right now, and after that I can do those actions and then those actions and then those actions; which path is guaranteed to lead to my goal? The definition of rational behavior requires you to optimize over the entire future of the universe. It's just completely infeasible computationally.

It didn't make much sense that we should define what we're trying to do in AI as something that's impossible, so I tried to figure out: How do we really make decisions?

So, how do we do it?

One trick is to think about a short horizon and then guess what the rest of the future is going to look like. So chess programs, for example — if they were rational they would only play moves that guarantee checkmate, but they don't do that. Instead they look ahead a dozen moves into the future and make a guess about how useful those states are, and then they choose a move that they hope leads to one of the good states.

> Could you prove that your systems can't ever, no matter how smart they are, overwrite their original goals as set by the humans?

Another thing that's really essential is to think about the decision problem at multiple levels of abstraction, so "hierarchical decision making." A person does roughly 20 trillion physical actions in

their lifetime. Coming to this conference to give a talk works out to 1.3 billion or something. If you were rational you'd be trying to look ahead 1.3 billion steps — completely, absurdly impossible. So the way humans manage this is by having this very rich store of abstract, high-level actions. You don't think, "First I can either move my left foot or my right foot, and then after that I can either…" You think, "I'll go on Expedia and book a flight. When I land, I'll take a taxi." And that's it. I don't think about it anymore until I actually get off the plane at the airport and look for the sign that says "taxi" — then I get down into more detail. This is how we live our lives, basically. The future is spread out, with a lot of detail very close to us in time, but these big chunks where we've made commitments to very abstract actions, like, "get a Ph.D.," "have children."

Are computers currently capable of hierarchical decision making?

So that's one of the missing pieces right now: Where do all these high-level actions come from? We don't think programs like the DQN network are figuring out abstract representations of actions. There are some games where DQN just doesn't get it, and the games that are difficult are the ones that require thinking many, many steps ahead in the primitive representations of actions — ones where a person would think, "Oh, what I need to do now is unlock the door," and unlocking the door involves fetching the key, etcetera. If the machine doesn't have the representation "unlock the door" then it can't really ever make progress on that task.

But if that problem is solved — and it's certainly not impossible — then we would see another big increase in machine capabilities. There are two or three problems like that where if all of those were solved, then it's not clear to me that there would be any major obstacle between there and human-level AI.

What concerns you about the possibility of human-level AI?

In the first [1994] edition of my book there's a section called, "What if we do succeed?" Because it seemed to me that people in AI weren't really thinking about that very much. Probably it was just too far away. But it's pretty clear that success would be an enormous thing. "The biggest event in human history" might be a good way to describe it. And if that's true, then we need to put a lot more thought than we are doing into what the precise shape of that event might be.

The basic idea of the intelligence explosion is that once machines reach a certain level of intelligence, they'll be able to work on AI just like we do and improve their own capabilities — redesign their own hardware and so on — and their intelligence will zoom off the charts. Over the last few years, the community has gradually refined its arguments as to why there might be a problem. The most convincing argument has to do with value alignment: You build a system that's extremely good at optimizing some utility function, but the utility function isn't quite right. In [Oxford philosopher] Nick Bostrom's book [*Superintelligence*], he has this example of paperclips. You say, "Make some paperclips." And it turns the entire planet into a vast junkyard of paperclips. You build a super-optimizer; what utility function do you give it? Because it's going to do it.

What about differences in human values?

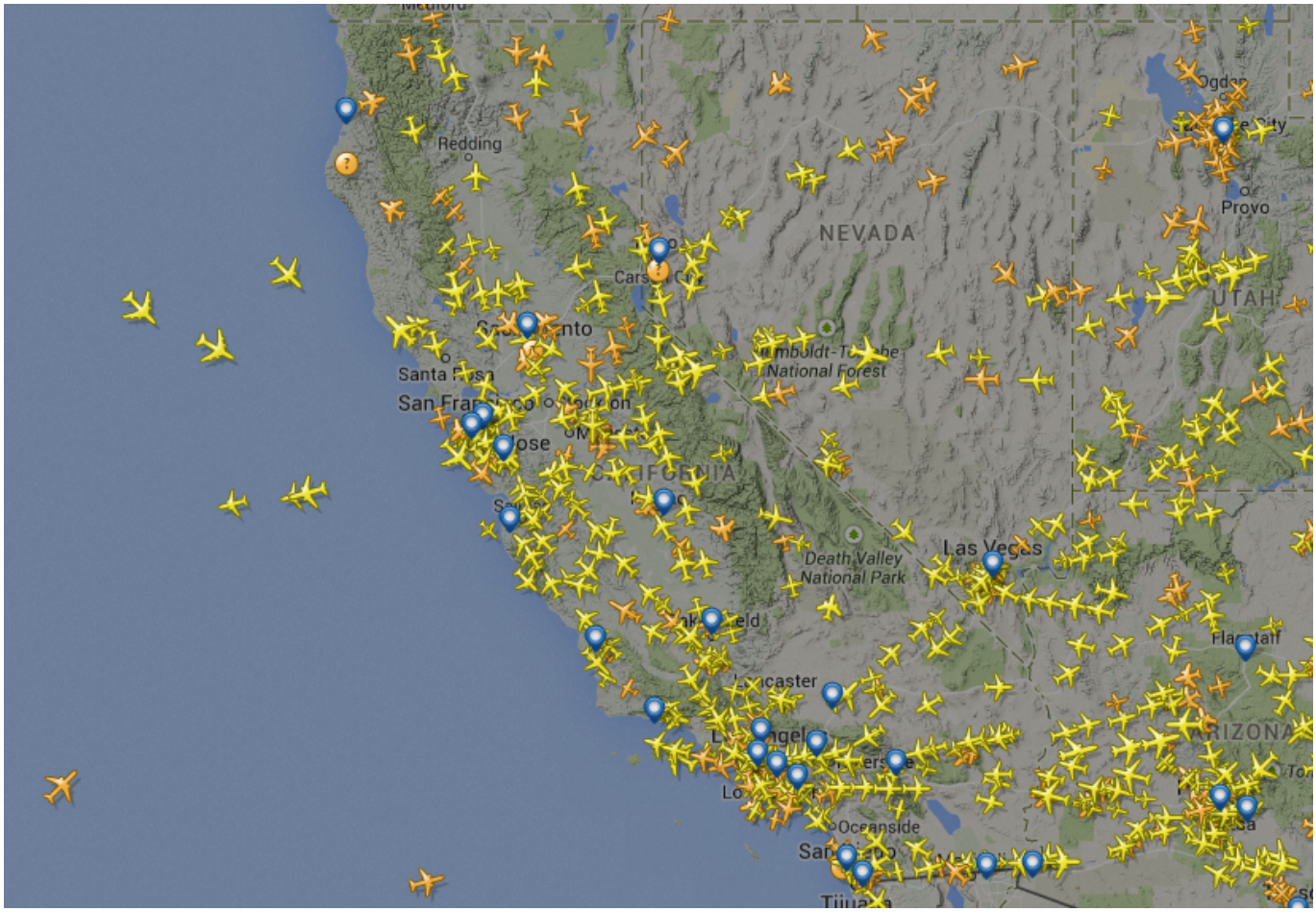
That's an intrinsic problem. You could say machines should err on the side of doing nothing in areas where there's a conflict of values. That might be difficult. I think we will have to build in these value functions. If you want to have a domestic robot in your house, it has to share a pretty good cross-section of human values; otherwise it's going to do pretty stupid things, like put the cat in the oven for dinner because there's no food in the fridge and the kids are hungry. Real life is full of these tradeoffs. If the machine makes these tradeoffs in ways that reveal that it just doesn't get it — that it's just missing some chunk of what's obvious to humans — then you're not going to want that thing

in your house.

I don’t see any real way around the fact that there’s going to be, in some sense, a values industry. And I also think there’s a huge economic incentive to get it right. It only takes one or two things like a domestic robot putting the cat in the oven for dinner for people to lose confidence and not buy them.

Then there’s the question, if we get it right such that some intelligent systems behave themselves, as you make the transition to more and more intelligent systems, does that mean you have to get better and better value functions that clean up all the loose ends, or do they still continue behaving themselves? I don’t know the answer yet.

You’ve argued that we need to be able to mathematically verify the behavior of AI under all possible circumstances. How would that work?



Automating air traffic control systems may require airtight proofs about real-world possibilities.

One of the difficulties people point to is that a system can arbitrarily produce a new version of itself that has different goals. That’s one of the scenarios that science fiction writers always talk about; somehow, the machine spontaneously gets this goal of defeating the human race. So the question is: Could you prove that your systems can’t ever, no matter how smart they are, overwrite their original goals as set by the humans?

It would be relatively easy to prove that the DQN system, as it’s written, could never change its goal of optimizing that score. Now, there is a hack that people talk about called “wire-heading” where you could actually go into the console of the Atari game and physically change the thing that produces the score on the screen. At the moment that’s not feasible for DQN, because its scope of action is entirely within the game itself; it doesn’t have a robot arm. But that’s a serious problem if

the machine has a scope of action in the real world. So, could you prove that your system is designed in such a way that it could never change the mechanism by which the score is presented to it, even though it's within its scope of action? That's a more difficult proof.

Are there any advances in this direction that you think hold promise?

There's an area emerging called "cyber-physical systems" about systems that couple computers to the real world. With a cyber-physical system, you've got a bunch of bits representing an air traffic control program, and then you've got some real airplanes, and what you care about is that no airplanes collide. You're trying to prove a theorem about the combination of the bits and the physical world. What you would do is write a very conservative mathematical description of the physical world — airplanes can accelerate within such-and-such envelope — and your theorems would still be true in the real world as long as the real world is somewhere inside the envelope of behaviors.

Yet you've pointed out that it might not be mathematically possible to formally verify AI systems.

There's a general problem of "undecidability" in a lot of questions you can ask about computer programs. Alan Turing showed that no computer program can decide whether any other possible program will eventually terminate and output an answer or get stuck in an infinite loop. So if you start out with one program, but it could rewrite itself to be any other program, then you have a problem, because you can't prove that all possible other programs would satisfy some property. So the question would be: Is it necessary to worry about undecidability for AI systems that rewrite themselves? They will rewrite themselves to a new program based on the existing program plus the experience they have in the world. What's the possible scope of effect of interaction with the real world on how the next program gets designed? That's where we don't have much knowledge as yet.

This article was reprinted on Wired.com.