Stuart Russell: Even though the things it's doing don't seem initially to be disastrous, there's just no telling, and such as this: it might on a whim decide to turn off the earth and go live somewhere else.

Host: That system is artificial intelligence, and although the potential of AI has fascinated researchers for decades, the reality of a system with superhuman intelligence was only a far-off dream. Now, though, a group of leading researchers in physics and computer science including Stephen Hawking have issued a stark warning. They say computers that are as intelligent as humans and can make their own decisions are closer than we think, and we're not nearly as prepared as we should be, so they're ditching the rose-colored glasses and sounding the AI alarm.

Stuart Russell: My name is Stuart Russell. I'm a professor of computer science at the University of California, Berkeley.

Host: Stuart wrote two recent op-eds calling on society to take a hard look at how artificial intelligence is being developed before it's too late. It's not a position many people in the field are taking since most of the development in AI is associated with benefits to humanity, everything from face and voice recognition to the latest in self-driving cars, so I asked Stuart what's got him so concerned.

Stuart Russell: Our concern is that in the future when systems become more and more intelligent and we become more and more dependent on their capabilities, we may find that they get beyond our control – because a system that's, for example, more intelligent than a human being is very hard to predict and even though we might think we're asking it to do something fairly innocuous and helpful like find a cure for cancer, unless you're very, very specific, you get the problem of giving the three wishes to the genie where usually the genie finds a loophole which ends up putting you in the soup, and at best, you get back to where you started. But where the genie that is far more intelligent than the human race put together, it might be very difficult for us to get back to where we started.

What we want to understand is how to develop intelligent technology without losing control over it and put very simply, what we want to do is to say that artificial intelligence is not just about making systems more intelligent, it's making them more intelligent *and beneficial to the human race* – which sounds like apple pie, but is actually something that other fields have taken seriously for a long time. For example, if you look at fusion research, the original goal of fusion research was: can we generate energy by bashing together small atoms to make big ones? And they figured out in the '40s that yes, you can, but you get an enormous explosion and you destroy millions of people. Ever since then, fusion research has had the goal of generating energy *and benefiting the human race* by

controlling the generation of energy so that you don't have a huge explosion. It's just taken for granted that that's what fusion research is about.

Host: Why aren't we already having those kinds of conversations about AI?

Stuart Russell: I think we're starting to. There have been people on the fringes of the field, some philosophers, some futurists, who have been warning for quite some time and you can go back and look at, for example, a paper by I.J. Good. In 1965, he warned that if we did succeed in building super intelligent systems, it would be the last invention that man ever need make. So I think in the last few years, the change has been that we're just seeing this acceleration in the rate of progress and technology and in the level of investment in the commercial sector so that we are starting to feel that it's time to take ourselves more seriously than we have been.

I think five years ago if you asked most AI people what are they doing, they'd say, "Oh, we're trying to build systems that are as intelligent as people," and if you asked them, "When are you going to succeed?" they'd say, "We have no idea. Decades and decades or centuries away." They just didn't worry about it, but now I think it seems as if it's a little bit closer and the fear is moving in from the fringe into the main stream and the feeling that we do need to grow up is growing.

Host: We've seen some startling examples of autonomy in computers and machines. We can think of Google's self-driving car or Watson, the Jeopardy game show winner, but how pervasive is this in our society now?

Stuart Russell: One thing that people might not realize is that search engines are an example of artificial intelligence. In fact, when Google started, the founders thought of it as an AI company and the ability to do search on the web was just one of the things they were going to try and it was the first thing that generated lots of money, but even as Google exists right now, it's doing lots of intelligent things, trying to figure out which of the many, many millions of web pages that contain the same words as your question are actually the web pages that contain the *answer* to your question. And they're using machine learning methods because they now have enormous amounts of data on what questions people typed in and which page did they select from the results that they were given.

All that information goes into improving their algorithms, so we'll see going forward that as systems become more and more capable of reading the web pages and looking at the images and understanding the television broadcasts and radio broadcasts that search will move from what it is now – which is mostly retrieving web pages – to a system that actually has understood the entire web and is going to answer your question based on that. This is one big direction.

We'll also see, as you mentioned, self-driving cars. I think there's still a way to go there because even though they're very impressive, they haven't reached a level of reliability that we would expect from, let's say, a human taxi driver who might drive for 10 years without making a serious mistake. We're still quite a way from that, but progress is being made and one reason progress is being made is that sensing the ability for machines to perceive their environments accurately has really improved.

Host: What potentially could happen if we don't get these systems under control?

Stuart Russell: There have been various scenarios that have been suggested and if you've seen the movie, Transcendence, that movie describes one of those scenarios where a super intelligent machine is combined with the uploaded brain of, as it turns out, a Berkeley AI professor who has been assassinated by terrorists, so that's [crosstalk 00:06:39].

Host: Are you watching your back?

Stuart Russell: [crosstsalk 00:06:40] watching my own assassination. Yeah, it was great.

Once that system comes online, then it has the mind of the human, but it has the computational and cognitive abilities of a superhuman machine, and so very quickly it develops technologies that are far beyond the capabilities of humans even to understand and in the movie, he heals the sick and he raises the dead, and in the space of a day, he develops a technology that puts right all the environmental damage we've done to the whole planet and so on. Unfortunately, or fortunately depending on how you look at it, the human race is terrified by such an entity because even though the things it's doing don't seem initially to be disastrous, there's just no telling, and such as this: it might on a whim decide to turn off the earth and go live somewhere else.

That's the notion of an intelligence explosion, that once a system reaches the same kind of capabilities as a human, then it can redesign itself using better hardware and better algorithms so that very quickly it accelerates and every improvement allows it to accelerate even faster and so, in a very short period of time, it's gone far, far beyond anything the human race can understand.

Host: What should we be doing as a society to safeguard ourselves from this technology getting hideously out of control?

Stuart Russell: I think we have to do, in some sense, the same thing the physicists do, to understand the physics of how we contain this process of intelligent decision making and also, I think, to understand what it is that we want as a human race because if we do succeed in building these kinds of capabilities, then in many

ways the human race can have anything it wants. If it wants to have eternal life, if we want to cure all diseases, if we want to resolve all conflicts between countries or groups of people, all of those things might be available to us, so we really do have a genie. And we have as many wishes as we want, as long as we don't make the wrong wish.

That's really the question. What does it mean to say we made the wrong wish? If you ask the system to cure cancer and nothing else, you don't ask for any more or any less than that, then you've forgotten something very important which is come up with a cure for cancer *and in so doing, don't destroy the planet*, don't turn the whole planet into a giant server farm so that you can improve your cancer discovery algorithms. Whenever you think about giving a goal to a super-intelligent system, it's very easy (as all the stories show) to find a loophole whereby achieving that goal to the greatest possible extent ends up having very undesirable side effects. Various philosophers have tried various ideas and so far, we don't yet have a solution.

To my mind, one of the most promising avenues is not to tell a system what we want it to do because if we don't tell it what we want, then it can't take any rash actions because those actions might be deleterious to what we do actually want, and so its only reasonable course of action at that point is to enter into a conversation with the human race to try to figure out what it is we do want, and the better it can figure that out, the more it's able to take actions to help us. But in the absence of fairly definite knowledge about what it is that humans want, it shouldn't really be taking any serious actions at all.

Host: How worried are you about this?

Stuart Russell: To me, it seems inevitable that the capabilities of intelligent machines will increase not just because of physics but because we're starting to understand better and better how to get those machines to behave intelligently. So it wouldn't surprise me if it happened in my lifetime and I could easily imagine it happening in the lifetime of my children. I think the sooner we start solving this problem, the better because if we don't solve it and the technology starts to become more and more integral to our society, it's going to be very difficult to reverse the process of technology development and improvement.

Host: Stuart, thanks so much for talking to us.

Stuart Russell: Thank you, Nora. It's been a pleasure.