3 Takeaways Podcast Transcript

Lynn Thoman
(https://www.3takeaways.com/)

Ep. 187: AI That’s More Powerful Than Humans Is Coming. How Will We Be Able To Control It?

This transcript was auto-generated. Please forgive any errors.

INTRO male voice: Welcome to the 3 Takeaways podcast, which features short, memorable conversations with the world's best thinkers, business leaders, writers, politicians, scientists, and other newsmakers. Each episode ends with the 3 key takeaways that person has learned over their lives and their careers. And now your host and board member of schools at Harvard, Princeton and Columbia, Lynn Thoman.

Lynn Thoman: Hi, everyone. It's Lynn Thoman. Welcome to another 3 Takeaways episode. 42 percent of CEOs surveyed at the Yale CEO Summit say artificial intelligence has the potential to destroy humanity five to 10 years from now. Here to explain the risk is one of the godfathers of artificial intelligence, Stuart Russell. He's a British computer scientist known for his contributions to artificial intelligence. He's a professor at Berkeley who is both the founder and head of the Center for Human Compatible Artificial Intelligence, and he's the co- author of the authoritative AI textbook, which is used in more than 1, 500 universities in 135 countries. He is also the author of the wonderful book, Human Compatible Artificial Intelligence and the Problem of Control. Welcome, Stuart, and thanks so much for joining 3 Takeaways today.

Stuart Russell: Thanks, Lynn. It's nice to be with you.

LT: It is a pleasure. Stuart, you nominated five candidates for the biggest event in the future of humanity. Can you tell us what the five candidates are and which one you believe is the winner?

SR: Sure. So, this was in a talk that I gave at Dulwich Picture Gallery, which is one of the oldest continuously existing public art museums in the UK, and I wanted to explain why I thought we needed to pay attention to what was happening in AI. So, I did it a bit like the Oscars. And here are the, five contenders for the biggest event in the future of humanity.

1. We all die (asteroid impact, climate catastrophe, pandemic, etc.).
2. We all live forever (medical solution to aging).
3. We invent faster-than-light travel and conquer the universe.
4. We are visited by a superior alien civilization.
5. We invent super intelligent AI

SR: And my explanation for why the fifth candidate, super intelligent AI [artificial intelligence] would be the winner, because it would help us avoid physical catastrophes, number 1, and achieve eternal life, number 2, and faster than light travel, number 3.

SR: It would represent a huge leap, a discontinuity in our civilization. The arrival of super intelligent AI is in many ways analogous to the arrival of a superior alien civilization, but much more likely to occur. Perhaps most important, AI, unlike aliens, is something over which we have some say. so this idea of this being essentially the biggest event in human history I still believe that to be correct.

LT: If machines are developing the ability to learn, do we have full control over what those machines will look like when they achieve super intelligent AI? Do we have any idea what the machines will be capable of in 5, 10, 30 years or more?

SR: It depends on how we do it. If we follow the line that we're pursuing right now, which is to basically take very, very large circuits. So we're talking about circuits with a trillion, or perhaps the next generation will be 10 trillion connections, and then train those from tens of trillions of words of text and maybe billions of hours of video, we have very little idea what's going on there. We can't examine the internal operations and understand them. because we're training them to imitate humans, we are probably causing them to acquire internal structures that function as goals. So practical purposes, we could say we are teaching them to acquire human like goals, but we don't even know what those goals are.

SR: So as the system has become more and more capable, we are likely to have less and less control over what they do. So my view is that we ought to redirect our efforts into ways of building AI systems where we can actually understand what's going on, where we can inspect the knowledge that it has, the goals that it's pursuing, the plans that it's proposing and so on.

LT: Where are machines and AI now shaping our lives?

SR: I think the most influence that AI systems have is through social media, because there the way social media platforms work is that there are algorithms constantly choosing what you see. what appears in your feed, what video comes up next when you're on YouTube or TikTok or whatever. Those algorithms in a real sense have more control over human cognitive intake than any dictator in history has ever had. They're completely unregulated and what they have learned to do, in order to optimize the objective that they're pursuing, which is basically to get as many clicks as possible, they have learned to manipulate people so that our consumption is more predictable. And in some sense, planning out a whole sequence of content that will turn us into somebody else, someone who is more easy to predict and manipulate. And that's a serious problem. So, I think this is pretty clear evidence that we had better figure out how to make AI systems safe and beneficial before it's too late.

LT: You talked about the global knowledge that computers will have. They will have in real time, knowledge of everything happening anywhere on the earth.

SR: To a large extent, they'll have access to all the satellite data feeds and satellites can see every object on earth, bigger than a football and, keep track of where it is and what's going on, everywhere. So that's already a really inhuman form of perception.

LT: And they'll have access to essentially all of us, where we are, what our devices are, what our messages are, what we're doing.

SR: The financial transactions, all of the movements of vehicles, and as you say, all the communications that we engage in and increasingly, you know, even inside buildings, right? Buildings are now gradually being equipped with more and more electronic sensors and cameras and so on, which mean that our behaviors can be tracked all the time.

LT: And let's talk about the capacity for action. A human has direct control over only one body, while what can a machine or AI control?

SR: So this is an interesting point because a lot of people seem to be under the impression because it's, quote, just a computer, it can't really do anything in the real world. and so we can always just turn it off. And both of those things are false. AI systems can communicate. There's I think, about between five and six billion people out of seven billion people on the internet, and AI systems in principle can communicate with all of those people individually. So that's enormous. And then you know, if an AI system wants to take physical action, it can pay people, it can persuade people, which is what Hitler did mainly. And that's already been demonstrated because there are websites like TaskRabbit where you can say, I want such and such a thing done in the real world, and, someone signs a contract and gets paid and they do it. And then, you know, we are putting AI systems in control of manufacturing facilities, of scientific research equipment and so on. So they're starting to have direct physical control.

LT: Do we always understand AI and machine decisions? And can you give some examples where unanticipated issues or consequences have arisen?

SR: It depends on the type of AI system, but if we look at the one that's maybe most familiar now, which is the large language model, so Chat GPT being the most prominent example and its descendants, we have very little understanding of how they work. And it's not because they're a secret: the people who built them don't understand how they work. They're just opaque. They don't seem to operate by any of the normal, or anything that corresponds to our understanding of normal reasoning processes and decision making processes, and so on.

SR: And that's the problem, because if we are training systems to imitate humans and the systems are acquiring human-like goals, we need to know, what they're trying to do, and what are their plans for trying to do that? And are they lying to us, and are they trying to manipulate us? And all those things, and if we can't find out, to me, it just seems completely foolish to proceed.

LT: As it does to me. The two examples that stand out to me are the flash crash on the New York Stock Exchange in 2010, where a trillion dollars was wiped out in a few minutes by trading algorithms. And my understanding is that no one still understands what happened, and they had to shut down the algorithms to stop it. And the other one, that to me is interesting, is the $24 million book on Amazon, where an Amazon pricing algorithm caused two booksellers to raise the price of a book about flies to about $24 million each.

SR: That one we did understand because the pricing algorithms that there were, were quite simple. But it was just the fact that those two algorithms, when you put them into a bidding war would ramp up the price basically infinitely. There's, there's no limit for how far it would go. The flash crash actually was a wake-up call, right? It was a sort of mini Chernobyl for the markets. And so they, put in circuit breakers and, you know, it also lost a huge amount of money for some of the traders involved. So now people are much more careful. about supervising the activities of their trading algorithms,

LT: Those are examples of unanticipated issues. But you also believe that there's potential for misuse of artificial intelligence. Can you give some examples, such as perhaps surveillance and influence and persuasion?

SR: Those are just some of the things. Surveillance, we already mentioned that these systems will have access to vast quantities of data about almost everybody on earth. And if you're a government that wants to exert control over its people surveillance is the first thing you need. And then you need a method of coercion. Traditionally coercion meant people showing up at your house and beating you over the head or pointing a gun at you and throwing you in jail or taking your children away, all sorts of methods of coercion that people have devised of taking away your livelihood and so on.

SR: I think borrowing from, actually from AI, this idea of reinforcement learning, which, involves training a system to exhibit certain behaviors by rewarding it when it exhibits the right kind of behavior that you want it to exhibit and punishing it when it fails to do that. You can apply that to humans by for example, rewarding good behavior with access to: good schools for your children, the opportunity to apply for better jobs, the opportunity to travel in first class on the train et cetera, et cetera, et cetera.

SR: So many of the same things that money buys you, you can also provide as rewards in this reinforcement learning process, and the punishments would be the dual of that: your kids can no longer go to a good school, you're not allowed to travel at all, you don't get promoted at work. And the kinds of behaviors that you might be trying to encourage would be supporting the party line in workplace discussions, possibly denouncing your neighbors for unpatriotic opinions and so on. So all the kinds of behaviors that mean that you are a reliable supporter of, the establishment and the power of the government and don't present a threat, those [behaviors] would be rewarded. And I'm afraid that this is probably going to be increasingly feasible and increasingly effective as the systems get built out and enable both better surveillance and better meting out of rewards and punishments.

LT: What are some of the benefits of AI in terms of either solving major problems or increases in wealth? What is the potential?

SR: AI, if it works, and if we can control it, would basically provide us with intelligence on tap whenever we need it. And if we have access to more intelligence, then I think we should be able to have a much better civilization. It's tempting to think of AI as a magic wand, and at the moment it isn't, by any means. But almost by definition, superintelligent AI would be a magic wand because by definition it can do anything that human beings could do, but it can do it at much less cost and much greater scale and speed.

SR: If your rural village needs a hospital or, it needs a road to connect it to a nearby town, then, you take out your phone and book a hospital and. all the trucks and robots would come along and build the hospital. And then you'd have a hospital a week or two. That vision is, I think, what's driving the enormous resources being poured into AI. It really could lift the standard of living of the whole human race. And, when you do the calculations, the value of that step change in the quality of life for everyone is in the quadrillions of dollars. That's just a down payment on how much value this would bring to humanity. And of course, there might be more things besides, but that's more speculative to say that it could, lengthen human lifespans, for example, or cure cancer, or find ways to educate every child to their full potential. But even those things are not unreasonable. It remains to be seen what the other consequences will be.

LT: In the past, we humans used technology as a tool, and now technology is advancing to the point where it is using and even controlling us. How do you see this risk of having our lives shaped by machines and of technology using and controlling us?

SR: Yeah, it's interesting the way you put it. It resembles the language used in Samuel Butler's Erewhon in, I think 1863 where, in that book, it describes a society in the aftermath of a massive conflict between the pro-machinists and the anti-machinists and the anti-machinists basically argue that these machines are becoming more and more and more capable and our bondage will steal upon us unawares. Basically they're saying, we will be to the machines as the beasts in the field are to us. So the anti-machinists win in that story. And so he's looking at a society where they no longer use machines. They've just decided that this is a slippery slope and you can't begin on that slippery slope.

SR: You can't go halfway down and say, okay, that's enough machine. You have to not go down that path at all. So, as AI systems, become more generally intelligent than human beings, they will be more powerful than us, in a literal sense. What does it mean to be more powerful? It means that if you have competing objectives, then the more powerful entity achieves its objective and the less powerful one does not.

SR: And, it's pretty clear from many examples, but mainly from the example of human domination, we get the objectives that we want because we're more intelligent than all of the other species on earth. And the other species have no say in whether they exist or not. So, you have to ask, okay, how do we retain power forever over entities more powerful than ourselves? And that's the basic question. And it's sort of irritating to hear, skeptics sort of poo pooing the idea that there might be a risk to humanity. Say, oh, this is just science fiction, we're just scaremongering because when you ask them that question, they don't have an answer.

SR: The risk comes from having no control over the outcome. Whatever the outcome is, it's dictated by the choices of the machines and no longer by the choices made by humans. There's nothing we can do about it.

LT: Because fundamentally we don't understand what they understand, because they have billions of real time data points all over the world?

SR: They may understand science better than we do. If it comes down to a physical conflict, they could devise weapons that we don't even understand. In some sense, the particular course of events, the particular scenario doesn't really matter. We cannot afford to be in a situation where AI systems could be operating in pursuit of objectives that are in conflict with ours, and doing so more effectively than we can control.

SR: What that means is if we don't figure out how to solve that control problem, how to maintain power forever over entities more powerful than ourselves, then we can't go down that route. We have to do what the anti-machinists did in Erewhon and say, okay, no machines,

LT: Can you summarize again the issues of control - that's the key issue.

SR: The key issue is how do we retain power forever over entities more powerful than ourselves? And, that's a real question, we have to answer it. And I believe it's possible by designing the entities in such a way that their only objective is the furtherance of human interests. But a key point is that the AI system knows that it doesn't know what human interests are. It knows that human interests are what humans want the future to be like, and what ultimately is the cause of the behavior that humans exhibit. And so there is a mechanism for machines to learn more about what human interests are and to become better at serving those interests. But this is the fundamental, as it were, “constitution” of AI systems that will, I think, enable us to retain power forever.

LT: What are the 3 takeaways that you’d like to leave the audience with today?

SR: So I think the 3 takeaways would be that it's very likely that we will achieve super intelligent AI and it's essential that before that happens, we have solved the safety problem. We have to figure out how to retain power over those systems that will become more powerful than us. And the third takeaway is that I believe the question does have an answer. It is possible. But a great deal of work is required to realize this technological approach. But, at the moment, we are investing all of our resources in a direction that doesn't answer the question, but may lead to super intelligent AI systems that we cannot control.

LT: And will there be an opportunity for a do over if something happens?

SR: I think we would have to say we'll be lucky to get a second chance because once you lose control, there may not be any possibility of getting it back.

LT: This is a very clear warning. Thank you very much, Stuart. This has been terrific.

SR: Thank you, Lynn. Very nice chatting.

OUTRO male voice: If you enjoyed today's episode and would like to receive the show notes or get new fresh weekly episodes, be sure to sign up for our newsletter at https://www.3takeaways.com/ or follow us on Instagram, Twitter, LinkedIn and Facebook. Note that 3Takeaways.com is with the number 3, 3 is not spelled out. See you soon at 3Takeaways.com (https://www.3takeaways.com/)

This transcript was auto-generated. Please forgive any errors.