
BBC REITH LECTURES 2021 – LIVING WITH ARTIFICIAL INTELLIGENCE

With Stuart Russell, Professor of Computer Science and founder of the Center for Human-Compatible Artificial Intelligence at the University of California, Berkeley

Lecture 1: The Biggest Event in Human History

ANITA ANAND: Welcome to the 2021 BBC Reith Lectures. We're at the British Library in the heart of London and as well as housing more than 14 million books, we are also home here to the Alan Turing Institute, the national centre for data science and artificial intelligence. Set up in 2015 it was, of course, named after the famous English mathematician, one of the key figures in breaking the Nazi enigma code therefore saving countless lives. We couldn't really think of a better venue to place this year's Reith Lectures, which will explore the role of artificial intelligence and what it means for the way we live our lives.

Our lecturer has called the development of artificial intelligence "the most profound change in human history," so we've given him four programmes to explain why. He's going to be addressing our fears. He's going to be explaining the likely impact on jobs and the economy and, hopefully, he will answer the most important question of all: who is ultimately going to be in control, is it us or is it the machines?

Let's meet him now. Please welcome the 2021 BBC Reith Lecturer, Professor Stuart Russell.

(AUDIENCE APPLAUSE)

ANITA ANAND: Stuart, it's wonderful that we're going to be hearing from you. I just wonder, actually, when you first became aware of artificial intelligence because for many of us our introduction would have been through sci-fi, so at what point did you think, actually, this will be a real-life career for me?

STUART RUSSELL: So I think it was when my grandmother bought me one of the first programmable calculators, the Sinclair Cambridge programmable, a little tiny, white calculator, and once I understood that you could actually get it to do things by writing these programs, I just wanted to make it intelligent. But if you've ever had one of those calculators you know that there's only 36 keystrokes that you can put in the program, and you can do various things, you can calculate square roots and signs and logs, but you couldn't really make it intelligent with that much. So, I ended up actually borrowing the giant supercomputer at Imperial College, the CDC 6600, which was about as big as this room and far less powerful than what's on your cell phone today.

ANITA ANAND: Well, I mean, this obviously marks you out as very different to the rest of us. We were all writing rude words and turning our calculator upside down.

Can we even measure how fast AI is developing?

STUART RUSSELL: I actually think it's very difficult. Machines don't have an IQ. This is a common mistake that some commentators make is to predict that machine IQ will exceed human IQ on some given date, but if you think about it, so AlphaGo, which is this amazing Go-Playing program that was developed just across the road, is able to beat the human world champion at playing Go but it can't remember anything, and then the Google search engine remembers everything, but it can't plan its way out of a paper bag. So, to talk about the IQ of a machine doesn't make sense.

Humans, when they have a high IQ, typically can do lots of different things. They can play games and remember things, and so it sort of makes sense. Even for humans there's not a particularly good way of describing intelligence, but for machines it makes no sense at all. So, we see big progress on particular tasks. Machine translation, for example, speech recognition is another one, recognising objects and images, these were things that we, in AI, have been trying to do for 50 or 60 years, and in the last 10 years we've actually pretty much solved them. That makes you think that the problems are not insolvable, and we can actually knock them over one by one.

ANITA ANAND: Well, I'm really looking forward to your first lecture. It is entitled *The Biggest Event in Human History.* Stuart, over to you.

STUART RUSSELL: Thank you, Anita. Thank you to the audience for being here. Thank you to the BBC for inviting me. It really is an enormous and a unique honour to give these lectures. We are at the Alan Turing Institute, named for this man who is now on the 50-pound note. The BBC couldn't afford a real one, so I printed out a fake one.

In 1936, in his early twenties, Turing wrote a paper describing two new kinds of mathematical objects, machines and programs. They turned out to be the most powerful ever found, even more so than numbers themselves. In the last few decades those mathematical objects have created eight of the 10 most valuable companies in the world and dramatically changed human lives. Their future impact through AI may be far greater.

Turing's 1950 paper "Computing Machinery and Intelligence" is at least as famous as his 1936 paper. It introduced many of the core ideas of AI, including machine learning. It proposed what we now call the Turing Test as a thought experiment, and it demolished several standard objections to the very possibility of machine intelligence.

Perhaps less well known are two lectures he gave in 1951. One was on the BBC's Third Programme, but this is going out on Radio 4 and the World Service, so I'll quote the other one, given to a learned society in Manchester. He said:

"*Once the machine thinking method had started, it would not take long to outstrip our feeble powers. At some stage therefore we should have to expect the machines to take control*."

Let me repeat that: *"At some stage therefore we should have to expect the machines to take control."*

I must confess that for most of my career I didn't lose much sleep over this issue, and I was not even aware, until a few years ago, that Turing himself had mentioned it.

I did include a section in my textbook, written with Peter Norvig in 1994, on the subject of "What if we succeed?" but it was cautiously optimistic. In subsequent years, the alarm was raised more frequently, but mostly from outside AI.

But by 2013, with the benefit of some time to think during a sabbatical in Paris, I became convinced that the issue not only belonged in the mainstream but was possibly the most important question that we faced. I gave a talk at the Dulwich Picture Gallery in which I stated that:

"Success would be the biggest event in human history and perhaps the last event in human history."

A few months later, in April 2014, I was at a conference in Iceland, and I got a call from National Public Radio asking if they could interview me about the new film Transcendence. It wasn't playing in Iceland, but I was flying to Boston the next day, so I went straight from the airport to the nearest cinema.

I sat in the second row and watched as a Berkeley AI professor, possibly me, played by Johnny Depp, naturally, was gunned down by activists worried about, of all things, super-intelligent AI. Perhaps this was a call from the Department of Magical Coincidences? Before Johnny Depp's character dies, his mind is uploaded to a quantum supercomputer and soon outruns human capabilities, threatening to take over the world.

A few days later, a review of Transcendence appeared in the Huffington Post, which I co-authored along with physicists Max Tegmark, Frank Wilczek, and Stephen Hawking. It included the sentence from my Dulwich talk about the biggest event in human history. From then on, I would be publicly committed to the view that success for my field would pose a risk to my own species.

Now, I've been talking about "success in AI," but what does that mean? To answer, I'll have to explain what AI is actually trying to do. Obviously, it's about making machines intelligent, but what does that mean?

To answer this question, the field of AI borrowed what was, in the 1950s, a widely accepted and constructive definition of human intelligence:

> *"Humans are intelligent to the extent that our actions can be expected to achieve our objectives."*

All those other characteristics of intelligence; perceiving, thinking, learning, inventing, listening to lectures, and so on, can be understood through their contributions to our ability to act successfully.

Now, this equating of intelligence with the achievement of objectives has a long history. For example, Aristotle wrote:

"We deliberate not about ends, but about means. A doctor does not deliberate whether he shall heal, nor an orator whether he shall persuade. They assume the end and consider how and by what means it is attained, and if it seems easily and best produced thereby."

And then, between the sixteenth and twentieth centuries, almost entirely for the purpose of analysing gambling games, mathematicians refined this deterministic view of "means achieving ends" to allow for uncertainty about the outcomes of actions and to accommodate the interactions of multiple decision-making entities, and these efforts culminated in the work of von Neumann and Morgenstern on an axiomatic basis for rationality, published in 1944.

From the very beginnings of AI, intelligence in machines has been defined in the same way:

> *"Machines are intelligent to the extent that their actions can be expected to achieve their objectives."*

But because machines, unlike humans, have no objectives of their own, we give them objectives to achieve. In other words, we build objective-achieving machines, we feed objectives into them, or we specialise them for particular objectives, and off they go. The same general plan applies in control theory, in statistics, in operations research, and in economics. In other words, it underlies a good part of the 20th century's technological progress. It's so pervasive, I'll call it the "standard model."

Operating within this model, AI has achieved many breakthroughs over the past seven decades. Just thinking of intelligence as computation led to a revolution in psychology and a new kind of theory, programs instead of simple mathematical laws. It also led to a new definition of rationality that reflects the finite computational powers of any real entity, whether artificial or human.

AI also developed symbolic computation, that is, computing with symbols representing objects such as chess pieces or aeroplanes, instead of the purely numerical calculations that had defined computing since the seventeenth century.

Also following Turing's suggestion from 1950, we developed machines that learn, that is they improve their achievement of objectives through experience. The first successful learning program was demonstrated on television in 1956. Arthur Samuel's draughts-playing program had learned to beat its own creator, and that program was the progenitor of Deepmind's AlphaGo, which taught itself to beat the human world champion in 2017.

Then in the sixties and seventies, systems for logical reasoning and planning were developed, and they were embodied to create autonomous mobile robots. Logic programming and rule-based expert systems supported some of the first commercial applications of AI in the early eighties, creating an immense explosion of interest in the US and Japan. The first self-driving Mercedes drove on the autobahn in 1987. Britain, on the other hand, had to play catch-up, having stopped nearly all AI research in the early seventies.

Then, in the 1990s, AI developed new methods for representing and reasoning about probabilities and about causality in complex systems, and those methods have spread to nearly every area of science.

Over the last decade, so-called deep learning systems appear to have learned to recognise human speech very well; to recognise objects in images, to translate between hundreds of different human languages. In fact, I use machine translation every year because I'm still paying taxes in France. It does a perfect job of translating quite impenetrable French tax instructions into equally impenetrable English tax instructions. Despite this setback, AI is increasingly important in the economy, running everything from search engines to autonomous delivery planes.

But as AI moves into the real world, it collides with Francis Bacon's observation from The Wisdom of the Ancients in 1609:

"The mechanical arts may be turned either way and serve as well for the cure as for the hurt."

"The hurt," with AI, includes racial and gender bias, disinformation, deepfakes, and cybercrime. And as Bacon also noted:

"*Out of the same fountain come instruments of death*."

Algorithms that can decide to kill human beings, and have the physical means to do so, are already for sale. I'll explain in the next lecture why this is a huge mistake. It's not because of killer robots taking over the world; it's simply because computers are very good at doing the same thing millions of times over.

All of these risks that I've talked about come from simple, narrow, application-specific algorithms. But let's not mince words. The goal of AI is and always has been general-purpose AI: that is, machines that can quickly learn to perform well across the full range of tasks that humans can perform. And one

must acknowledge that a species capable of inventing both the gravitational wave detector and the Eurovision song contest exhibits a great deal of generality.

Inevitably, general-purpose AI systems would far exceed human capabilities in many important dimensions. This would be an inflection point for civilisation.

I want to be clear that we are a long way from achieving general-purpose AI. Furthermore, we cannot predict its arrival based on the growth of data and computing power. Running stupid algorithms on faster and faster machines just gives you the wrong answer more quickly. Also, I think it's highly unlikely that the present obsession with deep learning will yield the progress its adherents imagine. Several conceptual breakthroughs are still needed, and those are very hard to predict.

In fact, the last time we invented a civilisation-ending technology, we got it completely wrong. On September 11, 1933, at a meeting in Leicester, Lord Rutherford, who was the leading nuclear physicist of that era, was asked if, in 25 or 30-years' time, we might unlock the energy of the atom. His answer was:

"*Anyone who looks for a source of power in the transformation of the atoms is talking moonshine*."

The next morning, Leo Szilard, a Hungarian physicist and refugee who was staying at the old Imperial Hotel on Russell Square, 10 minutes' walk from here, read about Rutherford's speech in The Times. He went for a walk and invented the neutron-induced nuclear chain reaction. The problem of liberating atomic energy went from "impossible" to essentially solved in less than twenty-four hours.

The moral of this story is that betting against human ingenuity is foolhardy, particularly when our future is at stake. Now, because we need multiple breakthroughs and not just one, I don't think I'm falling into Rutherford's trap if I say that it's quite unlikely we'll succeed in the next few years. It seems prudent, nonetheless, to prepare for the eventuality.

If all goes well, it will herald a golden age for humanity. Our civilisation is the result of our intelligence; and having access to much greater intelligence could enable a much better civilisation.

One rather prosaic goal is to use general-purpose AI to do what we already know how to do more effectively, at far less cost, and at far greater scale. We could, thereby, raise the living standard of everyone on Earth, in a sustainable

way, to a respectable level. That amounts to a roughly tenfold increase in global GDP. The cash equivalent, or the net present value as economists call it, of the increased income stream is about 10 quadrillion pounds or $14 quadrillion. All of the huge investments happening in AI are just a rounding error in comparison.

If 10 quadrillion pounds doesn't sound very concrete, let me try to make this more concrete by looking back at what happened with transportation. If you wanted to go from London to Australia in the 17th century, it would have been a huge project costing the equivalent of billions of pounds, requiring years of planning and hundreds of people, and you'd probably be dead before you got there. Now we are used to the idea of transportation as a service or TaaS. If you need to be in Melbourne tomorrow, you take out your phone, you go tap-tap-tap, spend a relatively tiny amount of money, and you're there, although they won't let you in.

General-purpose AI would be everything as a service, or XaaS. There would be no need for armies of specialists in different disciplines, organised into hierarchies of contractors and subcontractors, to carry out a project. All embodiments of general-purpose AI would have access to all the knowledge and skills of the human race. In principle, politics and economics aside, everyone could have at their disposal an entire organisation composed of software agents and physical robots, capable of designing and building bridges, manufacturing new robots, improving crop yields, cooking dinner for a hundred guests, separating the paper and the plastic, running an election, or teaching a child to read. It is the generality of general-purpose AI that makes this possible.

Now that's all fine if everything goes well. Although, as I will discuss in the third lecture, there is the question of what's left for us humans to do.

On the other hand, as Alan Turing warned, in creating general-purpose AI, we create entities far more powerful than humans. How do we ensure that they never, ever have power over us? After all, it is our intelligence, individual and collective, that gives us power over the world and over all other species.

Turing's warning actually ends as follows:

"*At some stage therefore, we should have to expect the machines to take control in the way that is mentioned in Samuel Butler's Erewhon*."

Butler's book describes a society in which machines are banned, precisely because of the prospect of subjugation. His prose is very 1872:

"*Are we not ourselves creating our successors in the supremacy of the Earth? In the course of ages, we shall find ourselves the inferior race. Our bondage will steal upon us noiselessly and by imperceptible approaches*."

Is that the end of the story, the last event in human history? Surely, we need to understand why making AI better and better makes the outcome for humanity worse and worse. Perhaps if we do understand, we can find another way.

Many films such as Terminator and Ex Machina would have you believe that spooky emergent consciousness is the problem. If we can just prevent it, then the spontaneous desire for world domination and the hatred of humans can't happen. There are at least two problems with this.

First, no one has any idea how to create, prevent, or even detect consciousness in machines or, for that matter, in functioning humans.

Second, it has absolutely nothing to do with it. Suppose I give you a program and ask, "Does this program present a threat to humanity?" You analyse the code and indeed, when run, it will form and carry out a plan to destroy humanity, just as a chess program forms and carries out a plan to defeat its opponent. Now suppose I tell you that the code, when run, also creates a form of machine consciousness. Will that change your prediction? No, not at all. It makes absolutely no difference. It's competence, not consciousness, that matters.

To understand the real problem with making AI better, we have to examine the very foundations of AI, the "standard model" which says that:

> *"Machines are intelligent to the extent that their actions can be expected to achieve their objectives*."

For example, you tell a self-driving car, "Take me to Heathrow," and the car adopts the destination as its objective. It's not something that the AI system figures out for itself; it's something that we specify. This is how we build all AI systems today.

Now the problem is that when we start moving out of the lab and into the real world, we find that we are unable to specify these objectives completely and correctly. In fact, defining the other objectives of self-driving cars, such as how to balance speed, passenger safety, sheep safety, legality, comfort, politeness, has turned out to be extraordinarily difficult.

This should not be a surprise. We've known it for thousands of years. For example, in the ancient Greek legend, King Midas asked the gods that everything he touch should turn to gold. This was the objective he specified, and the gods granted his objective. They are the AI in this case. And of course, his food, his drink, and his family all turn to gold, and he dies in misery and starvation.

We see the same plot in the Sorcerer's Apprentice by Goethe, where the apprentice asks the brooms to help him fetch water, without saying how much water. He tries to chop the brooms into pieces, but they've been given their objective, so all the pieces multiply and keep fetching water.

And then there are the genies who grant you three wishes. And what is your third wish? It's always, "Please undo the first two wishes because I've ruined the world."

Talking of ruining the world, let's look at social media content-selection algorithms, the ones that choose items for your newsfeed or the next video to watch. They aren't particularly intelligent, but they have more power over people's cognitive intake than any dictator in history.

The algorithm's objective is usually to maximise click-through, that is, the probability that the user clicks on presented items. The designers thought, perhaps, that the algorithm would learn to send items that the user likes, but the algorithm had other ideas.

Like any rational entity, it learns how to modify the state of its environment, in this case the user's mind, in order to maximise its own reward, by making the user more predictable. A more predictable human can be fed items that they are more likely to click on, thereby generating more revenue. Users with more extreme preferences seem to be more predictable. And now we see the consequences of growing extremism all over the world.

As I said, these algorithms are not very intelligent. They don't even know that humans exist or have minds. More sophisticated algorithms could be far more effective in their manipulations. Unlike the magic brooms, these simple algorithms cannot even protect themselves, but fortunately they have corporations for that.

In fact, some authors have argued that corporations themselves already act as super-intelligent machines. They have human components, but they operate as profit-maximising algorithms.

The ones that have been creating global heating for the last hundred years have certainly outsmarted the human race, and we seem unable to interfere with their operation. Again, the objective here, profit neglecting externalities, is the wrong one.

Incidentally, blaming an optimising machine for optimising the objective that you gave it is daft. It's like blaming the other team for scoring against England in the World Cup. We're the ones who wrote the rules. Instead of complaining, we should rewrite the rules so it can't happen.

What we see from these lessons is that with the standard model and mis-specified objectives, "better" AI systems or better soccer teams produce worse outcomes. A more capable AI system will make a much bigger mess of the world in order to achieve its incorrectly specified objective, and, like the brooms, it will do a much better job of blocking human attempts to interfere.

And so, in a sense we're setting up a chess match between ourselves and the machines, with the fate of the world as the prize. You don't want to be in that chess match.

Earlier Anita asked me, "Does everyone in AI agree with me?" Amazingly, not, or at least not yet. For some reason, they can be quite defensive about it. There are many counterarguments, some so embarrassing it would be unkind to repeat them.

For example, it's often said that we needn't put in objectives such as self-preservation and world domination. But remember the brooms: the apprentice's spell doesn't mention self-preservation, but self-preservation is a logical necessity for pursuing almost any objective, so the brooms preserve themselves and even multiply themselves in order to fetch water.

Then there's the Mark Zuckerberg–Elon Musk "smackdown" that was so eagerly reported in the press. Elon Musk had drawn the analogy between creating super-intelligent AI and "summoning the demon."

Mark Zuckerberg replied, "If you're arguing against AI, then you're arguing against safer cars and being able to diagnose people when they're sick." Of course, Elon Musk isn't arguing against AI. He's arguing against uncontrollable AI.

If a nuclear engineer wants to prevent the uncontrolled nuclear reactions that we saw at Chernobyl, we don't say she's "arguing against electricity." It's not "anti-AI" to talk about risks. Elon Musk isn't a Luddite, and nor was Alan Turing,

even though we were all jointly given the Luddite of the Year Award in 2015 for asking, “What if we succeed?” The genome editors and the life extenders and the brain enhancers should also ask: What if we succeed? What then? In the case of AI, how do you propose to retain power, forever, over entities more powerful than ourselves?

One option might be to ban AI altogether, just as Butler’s anti-machinists in Erewhon banned all mechanical devices after a terrible civil war. In Frank Herbert’s Dune, the Butlerian Jihad had been fought to save humanity from machine control, and now there is an 11th commandment:

“*Thou shalt not make a machine in the likeness of a human mind*.”

But then I imagine all those corporations and countries with their eyes on that 10 quadrillion-pound prize, and I think, “Good luck with that.”

The right answer is that if making AI better and better makes the problem worse and worse, then we’ve got the whole thing wrong. We think we want machines that achieve the objectives we give them, but actually we want something else. Later in the series I’ll explain what that “something else” might be, a new form of AI that will be provably beneficial to the human race, as well as all the questions that it raises for our future.

Thank you very much.

(AUDIENCE APPLAUSE)

ANITA ANAND: Stuart, thank you very much indeed. Before we open this up to the audience at the Alan Turing Institute, you touched on this chat we had beforehand about whether people agree with you. Can we drill down into that a bit more because you’re based at Berkley, Silicon Valley is a stone’s throw away.

STUART RUSSELL: Yes.

ANITA ANAND: The majority of people who work in your field, do they regard you as a sage, a Cassandra? I suppose what I’m asking, are you a bit of a Billy No-Mates in Silicon Valley?

STUART RUSSELL: One response is quite understandable, which is I am a machine-learning researcher working at the coalface of AI. It’s really difficult to get my machines to do anything. Just leave me alone and let me make progress on solving the problem that my boss asked me to solve. Stop talking about the future. But the problem is, this is just a slippery slope. If you keep doing that, as

happened with the climate, I'm sure the people who produce petrol are saying, "Just leave me alone. People need to drive. I'm making petrol for them," but that's a slippery slope.

I do think that there is a sea change in the younger generation of researchers. Five years ago, I would say most people going into machine learning had dollar signs in their eyes, but now they really want to make the world a better place.

ANITA ANAND: Is that sea change enough if we carry on down this slope? You mentioned Chernobyl, I wonder whether you'd go as far as to say that there needs to be a Chernobyl-type event in AI before everyone listens to you?

STUART RUSSELL: Well, I think what's happening in social media is already worse than Chernobyl. It has caused a huge amount of dislocation.

ANITA ANAND: Well, if that's a little bit to chew on, let us chew on it now. Let's take some questions from the floor.

CLAIRE FOSTER-GILBERT: Claire Foster-Gilbert from Westminster Abbey Institute. Thank you very much indeed for your lecture. I wanted to ask you if you had any wisdom to share with us on the kinds of people we should try and be ourselves as we deal with, work with, direct, live with AI?

STUART RUSSELL: I'm not sure I have any wisdom on any topic, and that's an incredibly interesting question that I've not heard before. I'm going to give a little preview of what I'm going to say in the later lecture. The process that we need to have happen is that there's a flow of information from humans to machines about what those humans want the future to be like, and I think introspection on those preferences that we have for the future would be extremely valuable. So many of our preferences are unstated because we all share them.

For example, a machine might decide, okay, I've got this way of fixing the carbon dioxide concentration in the atmosphere to help with the climate, but it changes the colour of the sky to a sort of dirty, green ochre colour. Is that all right? Well, most of us have never thought about our preferences for the colour of the sky because we like the blue sky that we have. We don't make these preferences explicit because we all share them and also because we don't expect that aspect of the world to be changed, but introspecting on what makes a good future for us, our families and the world, and noticing, I think, that actually we all share far more than we disagree on about what that future should be like would be extremely valuable.

I notice that there is now a really active intellectual movement, or even a set of intellectual movements, around trying to make explicit what does human wellbeing mean? What is a good life? And I think it's just in time because for almost all of history in almost all parts of the world the main thing has been how do we not die, and if things go well, that time comes to an end and we actually have a breathing space then if we're not faced with imminent death, starvation, whatever, we have a breathing space to think about what should the future be. We finally have a choice, and we haven't really yet had enough discussion about that.

So that's what I would like everyone to do. If we have a choice, what should the future look like? If you could choose, if you weren't constrained by history or by resources, what would it be?

ANITA ANAND: Let us take a question from this side?

PAUL INGRAM: Paul Ingram soon to start at the Cambridge University Centre for the Study of Existential Risk. Stuart, I wanted to invite you to draw a comparison with another existential risk that you mentioned in your lecture, namely the emergence of splitting of the atom and the potential for nuclear war and the Cold War. We managed to survive, although looking back that was more luck than judgment, do you draw any analogies for the emergence of artificial intelligence?

STUART RUSSELL: I think it is an absolutely fascinating subject. What happened after Leo Szilard had this inspiration, he actually was crossing at the traffic light at South Hampton Row, and I tried walking backwards and forwards across that crossing and I haven't had any inspiration at all.

He realised very soon that this was a bad time to have had this discovery because there was already the beginnings of an arms race with Nazi Germany. He was a refugee. And he figured out how to make a nuclear reactor with all of its feedback control systems to keep the subcritical reaction going. He patented that in 1934 but he kept the patent secret because he did not want it to fall into the wrong hands, but fairly soon the Germans also figured this out.

Otto Hahn, Lise Meitner, were German physicists who were, I think, the first to actually demonstrate a fission reaction, and when it happened in the US, I think Villard and Teller were able to get a fission reaction to happen in their lab, and he went home and wrote in his diary:

"Tonight I felt that the world was headed for grief."

I think we have been incredibly lucky not to have suffered nuclear annihilation, and after the war the United States had a window of complete power and they set up the International Atomic Energy Agency and very strict standards for developing peacetime nuclear power, and that enabled the sharing of designs because we could be sure that the design safety rules would be followed, inspection and regulation and so on, and I think there's a lot of lessons in all of those phases for how we think about AI and a key is not to think of it as an arms race. That's what we're doing right now. We have Putin, we have US Presidents, Chinese, Secretaries, talking about this as if, "We are going to win. We're going to use AI and that will enable us to rule the world," and I think that's a huge mistake.

One is that it causes us to cut corners. If you're in a race, then safety is the last thing on your mind. You want to get there first and so you don't worry about making it safe. But the other is that general purpose or super-intelligent AI would be, essentially, an unlimited source of wealth and arguing over who has it would be like arguing over who has a digital copy of the daily Telegraph or something, right? If I have a digital copy, it doesn't prevent other people from having digital copies and it doesn't matter how many I have, it doesn't do me a lot of good.

So, I think we're seeing, on the corporate side, actually a willingness to share super-intelligent AI technology, if and when it's developed, on a global scale, and I think that's a really good development. We just have to get the governments on board with that principle.

ANITA ANAND: Thank you very much. We have many hands going up but actually, my eye has been caught by one of the fathers of the World Wide Web, the father of the World Wide Web, Tim Berners Lee is with us, and I hope you don't mind, I'm just sort of zeroing on you. Are you optimistic or pessimistic when it comes to the future of AI?

TIM BERNERS LEE: I am hopeful about the power of it, but I think all of these concerns are very real. When things go wrong in terms of social network, the sort of same tipping point happens when people end up getting polarised and afterwards we take the pieces apart, but there are lots of other systems in the world where the world is very connected and some of them are in government, and some of them are in big companies. Some are in, for example, investment companies.

If you're a fast trader, for example, humans need not apply because you have to be too fast. So, we've already got some jobs, and a lot of jobs in banks,

you have to be fast and so therefore it has to be run by AI already. Could it be that we get AI suddenly much more quickly if we build competitive AI systems?

STUART RUSSELL: It might. I would have to say that the whole field of evolutionary computation has been a field full of optimism for a long time. The idea that you could use nature's amazing process of evolving creatures to instead evolve algorithms hasn't really paid off yet. The drawback of doing things that way is that you have absolutely no idea how it works and creating very powerful machines that work on principles you don't understand at all seems to be pretty much the worst way of going about this.

TABITHA GOLDSTAUB: Hello, Stuart. Thank you. I'm Tabitha, the Chair of the government's AI Council. I can't help but ask, what should we be teaching in school?

STUART RUSSELL: I mean, not everyone needs to understand how AI works any more than I need to understand how my car engine works in order to drive it. They should understand what AI can and cannot do presently, and I hope they will understand the need to make sure that when AI is deployed, it's deployed in a way that's actually beneficial.

This is the big change, right, to think not just about how can I get a machine to do X, but what happens when I put a machine that does X into society, into schools, into hospitals, into companies, into governments, what happens, and there's really not much of a discipline answering that question right now.

ANITA ANAND: Let's take some more?

STEPHANIE: Hi, Professor Russell. This is Stephanie here. I'm interested in your views on the role of regulation for artificial intelligence and how we get the balance right between regulating and not constraining innovation, particularly if we do that in a liberal democracy and other countries around the world that are not liberal democracies do not regulate? Thank you.

STUART RUSSELL: I think it depends what you're talking about regulating. I think there are things that we should regulate now, and I'm happy to say that the EU is in the process of doing that, such as the impersonation of human beings by machines. So, that could be, for example, a phone call that you get that sounds exactly like your husband or your wife or one of your children, asking you to send some money or they've forgotten the password for your account or whatever it might be, that's quite feasible to do now. But generally, a machine impersonating a human is a lie and I don't see why we should authorise lies for

commercial purposes, and I'm happy to say the EU is explicitly banning that in the new legislation, and that should be something that is a global agreement.

There are other things we should be very restrictive of, such as deep fakes, material that convinces people that some event happened that didn't actually happen, but the question of safety, how we regulate to make sure that AI systems don't produce disastrous outcomes where humanity loses control, we don't know how to write that rule yet.

ANITA ANAND: One of the phrases you used when you were doing your lecture was, "Good luck with that." I mean, we can't get people to agree on most things, how are you going to agree a framework for this?

STUART RUSSELL: When it's in their self-interest, right, so everyone agrees on TCP/IP, which is the protocol that allows machines to communicate on the internet, because if they don't agree with that the machine at the other end doesn't understand them and so you can't send your message. So, everyone agrees on that protocol because it works. Same with wi-fi and standards for cell phones and all this stuff, so there's a huge process. It's invisible to almost everybody but there are giant committees and annual meetings that go on and on and on, and they argue about the most tiny details of all these standards until they hammer it all out and then that standard is incredibly beneficial.

So, we could do the same thing for how you design AI systems to ensure that they are actually beneficial to humans, but we're not ready to say what the standards should be.

ANITA ANAND: There is one here?

JANE BONHAM CARTER: Jane Bonham Carter. I'm a Liberal Democrat politician but I, for years, worked in television and when TV/radio came along, and that was an intrusion into people's lives in a way that had never existed before, but it covered the ground of what I think you were talking about, which is what people shared. So, can AI not be directed towards a more benign curation, I suppose, is my question?

STUART RUSSELL: Absolutely, and as I said, I think some of the social media companies are genuinely interested. I don't think it's just a window dressing or a self-washing or anything, it's that they are genuinely interested in how their products, which are incredibly powerful, how they can be actually beneficial to people. I can't say very much at the moment but we, among others, are developing research relationships and we're finding openness and willingness

to share data and algorithms so that we can actually understand how to do this right.

It actually turns out to be one of the most difficult questions because if you think about driving, for example, it's difficult but probably not impossible to figure out how we should trade off safety versus getting to your destination, versus politeness to other drivers and so on, but what the algorithms are doing is actually changing our preferences, so it's changing what we want.

The person who first ventures into social media, having never touched it before, might be horrified by the person that they have become 12 months later. But the person 12 months later isn't horrified by themselves, right, they are actually really happy that they're now a diehard ecoterrorist and they're out there doing this, that and the other, and we don't even have a basic philosophical understanding of how to make decisions on behalf of someone who's going to be different when those decisions have impact. Do I help the person achieve what they want now, or do I help the person achieve what they're going to want when I achieve it?

It's a puzzle and philosophers have started writing about it, but we just don't have an answer and so this manipulation of human preferences by social media algorithms is actually getting at the hardest thing to understand in the AI problem, as far as I can see.

ANITA ANAND: Let's take another question here from this row?

STEVE McMANUS: Hi, my name's Steve McManus, a lifelong NHS employee. Arguably you are one of the thought leaders in this field, given also some of the other members of the audience we've got here today; where do you draw your thought leadership on this subject?

STUART RUSSELL: Another good question. I have found, actually, reading outside of my field, reading outside AI, in economics, particularly philosophy, has been enormously useful, although economics has this – it's called "the dismal science," I think that's a bit unfair. It's a very hard problem. It's, in many ways, a lot harder than physics and chemistry, but economists actually do try to think about this question: How should the world be arranged?

I was really shocked going back to read Adam Smith, who's widely reviled as "The Apostle of Greed," and so on and so forth, but actually what Adam Smith says at the beginning of his first book is that:

"It's so obvious to everyone that each of us cares deeply about other people that it hardly merits saying it, but I'm going to say it anyway," and then he says it.

That's the beginning of his first book. So, I've learned a great deal from economists, from philosophers, trying to understand a question that AI is now going to have to answer.

If AI systems are going to be making decisions on behalf of the human race, what does that mean? How do you tell whether a decision is a good or a bad decision when it's being made on behalf of the human race, and that's something that philosophers have grappled with for thousands of years?

ROLY KEATING: Roly Keating from British Library. It's wonderful to have you here. Thank you for the lecture. I was interested in the language and vocabulary of human intellectual life that seems to run around AI, and I'm hearing data gathering, pattern recognition, knowledge, even problem solving, but I think an earlier question used the word "wisdom," which I've not heard so much around this debate, and I suppose I'm trying to get a sense of where you feel that fits into the equation. Is AI going to help us as a species gradually become wiser or is wisdom exactly the thing that we have to keep a monopoly on? Is that a purely human characteristic, do you think?

STUART RUSSELL: Or the third possibility would be that AI helps us achieve wisdom without actually acquiring wisdom of its own, and I think, for example, my children have helped me acquire wisdom without necessarily having wisdom of their own. They certainly help me achieve humility. So, AI could help, actually, by asking the questions, right, because in some ways AI needs us to be explicit about what we think the future should be, that just the process of that interrogation could bring some wisdom to us.

ANITA ANAND: And the final question, apologies if we didn't get to you, so many fantastic questions, but the final one with you?

GILA SACKS: Hi. Gila Sacks. It seems that one of the most scary things about this future is that if individuals feel powerless in the face of machines and corporations, it will be a self-fulfilling prophecy, we will be powerless. So, how can individuals have power in the future that you see playing out, either as consumers or as citizens?

STUART RUSSELL: I wish that the entire information technology industry had a different structure. If you take your phone out and look at it, there are 50 or a hundred corporate representatives sitting in your pocket busily sucking out as much money and knowledge and data as they can. None of the things on your phone really represent your interests at all.

What should happen is that there's one app on your phone that represents you that negotiates with the information suppliers, and the travel agencies and whatever else, on your behalf, only giving the information that's absolutely necessary and even insisting that the information be given back, that transactions be completely oblivious, that the other party retains no record whatsoever of the transaction, whether it's a search engine query or a purchase or anything else.

This is technologically feasible but the way the market has evolved where it's completely the other way around. As individuals, you're right, we have no power. You have to sign a 38-page legal agreement to breathe and that, I really think, needs to change and the people who are responsible for making that change are the regulators.

Just to give a simple example, right, it's a federal crime in the US to make a phone call, a robocall, to someone who is on the federal Do Not Call list. I am on the federal Do Not Call List. I get 15 or 20 phone calls a day from robocalls. When you add that up, that is billions of crimes a day or trillions or crimes every year, trillions of federal crimes occurring, and there hasn't been a single prosecution, as far as I know, this whole year. I think there was one last year where they took one group down, but there is a total failure. We are in the wild west and there isn't a Sheriff in sight. So, as individuals, ask your representatives to do something about it.

We are also responsible, the technologists are also responsible, because we developed the internet in a very benign mindset. I can remember, when I was a computer scientist at Stanford, we could actually map our screens to anybody else's screen in the building and see what was on their screen. We thought that was cool, right? It just never occurred to anyone that that might be not totally desirable. We built technology with just open doors and complete fictitious IDs and all the rest of it. I think on the technology side, allowing real authentication of individual's traceability, responsibility, and then regulations with teeth, would help a great deal.

ANITA ANAND: Well, with thoughts of teeth, robocalls and crowded pockets, I'm afraid we're going to have to leave it there. Next time Stuart is going to be asking: *What AI means for conflict and war.* That is from Manchester, but for now a big thanks to our audience, to the Alan Turing Institute for hosting us, and, of course, to our Reith Lecturer, Stuart Russell.

(AUDIENCE APPLAUSE)

END OF TRANSCRIPT